


I'm not robot  reCAPTCHA

[Continue](#)

Propensity score analysis statistical methods and applications pdf

Loading... Top reviews Most recent Top reviews File loading please wait... 2 Advanced Quantitative Techniques in the Social Sciences VOLUMES IN THE SERIES 1. HIERARCHICAL LINEAR MODELS: Applications and Data Analysis Methods, 2nd Edition Stephen W. Raudenbush and Antony S. Bryk 2. MULTIVARIATE ANALYSIS OF CATEGORICAL DATA: Theory John P. van de Geer 3. MULTIVARIATE ANALYSIS OF CATEGORICAL DATA: Applications John P. van de Geer 4. STATISTICAL MODELS FOR ORDINAL VARIABLES Clifford C. Clogg and Edward S. Shihadeh 5. FACET THEORY: Form and Content Ingwer Borg and Samuel Shye 6. LATENT CLASS AND DISCRETE LATENT TRAIT MODELS: Similarities and Differences Ton Heinen 7. REGRESSION MODELS FOR CATEGORICAL AND LIMITED DEPENDENT VARIABLES J. Scott Long 8. LOG-LINEAR MODELS FOR EVENT HISTORIES Jeroen K. Vermunt 9. MULTIVARIATE TAXOMETRIC PROCEDURES: Distinguishing Types From Continua Niels G. Waller and Paul E. Meehl 10. STRUCTURAL EQUATION MODELING: Foundations and Extensions, 2nd Edition David Kaplan 11. PROPENSITY SCORE ANALYSIS: Statistical Methods and Applications, 2nd Edition Shenyang Guo and Mark W. Fraser 3 4 Copyright © 2015 by SAGE Publications, Inc. All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher. Printed in the United States of America Library of Congress Cataloging-in-Publication Data Guo, Shenyang, author. Propensity score analysis : statistical methods and applications / Shenyang Guo & Mark W. Fraser. — 2nd edition. pages cm. — (Advanced quantitative techniques in the social sciences) Includes bibliographical references and index. ISBN 978-1-4522-3500-4 (hardcover : acid-free paper) 1. Social sciences—Statistical methods. 2. Analysis of variance. I. Fraser, Mark W., 1946author. II. Title. HA29.G91775 2014 519.5'3—dc23 2014008147 This book is printed on acid-free paper. 14 15 16 17 18 10 9 8 7 6 5 4 3 2 1 5 FOR INFORMATION: SAGE Publications, Inc. 2455 Teller Road Thousand Oaks, California 91320 E-mail: SAGE Publications Ltd. 1 Oliver's Yard 55 City Road London EC1Y 1SP United Kingdom SAGE Publications India Pvt. Ltd. B 1/1 Mohan Cooperative Industrial Area Mathura Road, New Delhi 110 044 India SAGE Publications Asia-Pacific Pte. Ltd. 3 Church Street #10-04 Samsung Hub Singapore 049483 Acquisitions Editor: Vicki Knight Assistant Editor: Katie Guarino Editorial Assistant: Yvonne McDuffee Production Editor: Stephanie Palermi Copy Editor: Gillian Dickens Typesetter: C&M Digital (P) Ltd. Proofreader: Wendy Jo Dymond Indexer: Sheila Bodell Cover Designer: Candice Harman Marketing Manager: Nicole Elliott 6 Brief Contents List of Tables List of Figures Preface About the Authors 1. Introduction 2. Counterfactual Framework and Assumptions 3. Conventional Methods for Data Balancing 4. Sample Selection and Related Models 5. Propensity Score Matching and Related Models 6. Propensity Score Subclassification 7. Propensity Score Weighting 8. Matching Estimators 9. Propensity Score Analysis With Nonparametric Regression 10. Propensity Score Analysis of Categorical or Continuous Treatments: Dosage Analyses 11. Selection Bias and Sensitivity Analysis 12. Concluding Remarks References Index 7 Detailed Contents List of Tables List of Figures Preface 1. What the Book Is About 2. New in the Second Edition 3. Acknowledgments About the Authors 1 Introduction 1.1 Observational Studies 1.2 History and Development 1.3 Randomized Experiments 1.3.1 Fisher's Randomized Experiment 1.3.2 Types of Randomized Experiments and Statistical Tests 1.3.3 Critiques of Social Experimentation 1.4 Why and When a Propensity Score Analysis Is Needed 1.5 Computing Software Packages 1.6 Plan of the Book 2 Counterfactual Framework and Assumptions Causality, Internal Validity, and Threats Counterfactuals and the Neyman-Rubin Counterfactual Framework The Ignorable Treatment Assignment Assumption The Stable Unit Treatment Value Assumption Methods for Estimating Treatment Effects 2.5.1 Design of Observational Study 2.5.2 The Seven Models 2.5.3 Other Balancing Methods 2.5.4 Instrumental Variables Estimator 2.5.5 Regression Discontinuity Designs 2.6 The Underlying Logic of Statistical Inference 2.7 Types of Treatment Effects 2.1 2.2 2.3 2.4 2.5 8 2.8 Treatment Effect Heterogeneity 2.8.1 The Importance of Studying Treatment Effect Heterogeneity 2.8.2 Checking the Plausibility of the Unconfoundedness Assumption 2.8.3 A Methodological Note About the Hausman Test of Endogeneity 2.8.4 Tests of Treatment Effect Heterogeneity 2.8.5 Example 2.9 Heckman's Econometric Model of Causality 2.10 Conclusion 3 Conventional Methods for Data Balancing 3.1 Why Is Data Balancing Necessary? A Heuristic Example 3.2 Three Methods for Data Balancing 3.2.1 The Ordinary Least Squares Regression 3.2.2 Matching 3.2.3 Stratification 3.3 Design of the Data Simulation 3.4 Results of the Data Simulation 3.5 Implications of the Data Simulation 3.6 Key Issues Regarding the Application of OLS Regression 3.7 Conclusion 4 Sample Selection and Related Models 4.1 The Sample Selection Model 4.1.1 Truncation, Censoring, and Incidental Truncation 4.1.2 Why Is It Important to Model Sample Selection? 4.1.3 Moments of an Incidentally Truncated Bivariate Normal Distribution 4.1.4 The Heckman Model and Its Two-Step Estimator 4.2 Treatment Effect Model 4.3 Overview of the Stata Programs and Main Features of treatreg 4.4 Examples 4.4.1 Application of the Treatment Effect Model to Analysis of Observational Data 4.4.2 Evaluation of Treatment Effects From a Program With a Group Randomization Design 4.4.3 Running the Treatment Effect Model After Multiple Imputations of Missing Data 9 4.5 Conclusion 5 Propensity Score Matching and Related Models 5.1 Overview 5.2 The Problem of Dimensionality and the Properties of Propensity Scores 5.3 Estimating Propensity Scores 5.3.1 Binary Logistic Regression 5.3.2 Strategies to Specify a Correct Model—Predicting Propensity Scores 5.3.3 Hirano and Imbens's Method for Specifying Predictors Relying on Predetermined Critical Values 5.3.4 Generalized Boosted Modeling 5.4 Matching 5.4.1 Greedy Matching 5.4.2 Optimal Matching 5.4.3 Fine Balance 5.5 Postmatching Analysis 5.5.1 Multivariate Analysis After Greedy Matching 5.5.2 Computing Indices of Covariate Imbalance 5.5.3 Outcome Analysis Using the Hodges-Lehmann Aligned Rank Test After Optimal Matching 5.5.4 Regression Adjustment Based on Sample Created by Optimal Pair Matching 5.5.5 Regression Adjustment Using Hodges-Lehmann Aligned Rank Scores After Optimal Matching 5.6 Propensity Score Matching With Multilevel Data 5.6.1 Overview of Statistical Approaches to Multilevel Data 5.6.2 Perspectives Extending the Propensity Score Analysis to the Multilevel Modeling 5.6.3 Estimation of the Propensity Scores Under the Context of Multilevel Modeling 5.6.4 Multilevel Outcome Analysis 5.7 Overview of the Stata and R Programs 5.8 Examples 5.8.1 Greedy Matching and Subsequent Analysis of Hazard Rates 5.8.2 Optimal Matching 5.8.3 Post-Full Matching Analysis Using the Hodges-Lehmann Aligned Rank Test 5.8.4 Post-Pair Matching Analysis Using Regression of Difference Scores 10 5.8.5 Multilevel Propensity Score Analysis 5.8.6 Comparison of Rand-gbm and Stata's boost Algorithms 5.9 Conclusion 6 Propensity Score Subclassification 6.1 Overview 6.2 The Overlap Assumption and Methods to Address Its Violation 6.3 Structural Equation Modeling With Propensity Score Subclassification 6.3.1 The Need for Integrating SEM and Propensity Score Modeling Into One Analysis 6.3.2 Kaplan's (1999) Work to Integrate Propensity Score Subclassification With SEM 6.3.3 Conduct SEM With Propensity Score Subclassification 6.4 The Stratification-Multilevel Method 6.5 Examples 6.5.1 Stratification After Greedy Matching 6.5.2 Subclassification Followed by a Cox Proportional Hazards Model 6.5.3 Propensity Score Subclassification in Conjunction With SEM 6.6 Conclusion 7 Propensity Score Weighting 7.1 Overview 7.2 Weighting Estimators 7.2.1 Formulas for Creating Weights to Estimate ATE and ATT 7.2.2 A Corrected Version of Weights Estimating ATE 7.2.3 Steps in Propensity Score Weighting 7.3 Examples 7.3.1 Propensity Score Weighting With a Multiple Regression Outcome Analysis 7.3.2 Propensity Score Weighting With a Cox Proportional Hazards Model 7.3.3 Propensity Score Weighting With an SEM 7.3.4 Comparison of Models and Conclusions of the Study of the Impact of Poverty on Child Academic Achievement 7.4 Conclusion 8 Matching Estimators 11 8.1 Overview 8.2 Methods of Matching Estimators 8.2.1 Simple Matching Estimator 8.2.2 Bias-Corrected Matching Estimator 8.2.3 Variance Estimator Assuming Homoscedasticity 8.2.4 Variance Estimator Allowing for Heteroscedasticity 8.2.5 Large Sample Properties and Correction 8.3 Overview of the Stata Program nmatch 8.4 Examples 8.4.1 Matching With Bias-Corrected and Robust Variance Estimators 8.4.2 Efficacy Subset Analysis With Matching Estimators 8.5 Conclusion 9 Propensity Score Analysis With Nonparametric Regression 9.1 Overview 9.2 Methods of Propensity Score Analysis With Nonparametric Regression 9.2.1 The Kernel-Based Matching Estimators 9.2.2 Review of the Basic Concepts of Local Linear Regression (lowses) 9.2.3 Asymptotic and Finite-Sample Properties of Kernel and Local Linear Matching 9.3 Overview of the Stata Programs psmatch2 and bootstrap 9.4 Examples 9.4.1 Analysis of Difference-in-Differences 9.4.2 Application of Kernel-Based Matching to One-Point Data 9.5 Conclusion 10 Propensity Score Analysis of Categorical or Continuous Treatments: Dosage Analyses 10.1 Overview 10.2 Modeling Doses With a Single Scalar Balancing Score Estimated by an Ordered Logistic Regression 10.3 Modeling Doses With Multiple Balancing Scores Estimated by a Multinomial Logit Model 10.4 The Generalized Propensity Score Estimator 10.5 Overview of the Stata gpcscore Program 10.6 Examples 12 10.6.1 Modeling Doses of Treatment With Multiple Balancing Scores Estimated by a Multinomial Logit Model 10.6.2 Modeling Doses of Treatment With the Generalized Propensity Score Estimator 10.7 Conclusion 11 Selection Bias and Sensitivity Analysis 11.1 Selection Bias: An Overview 11.1.1 Sources of Selection Bias 11.1.2 Overt Bias Versus Hidden Bias 11.1.3 Consequences of Selection Bias 11.1.4 Strategies to Correct for Selection Bias 11.2 A Monte Carlo Study Comparing Corrective Models 11.2.1 Design of the Monte Carlo Study 11.2.2 Results of the Monte Carlo Study 11.2.3 Implications 11.3 Rosenbaum's Sensitivity Analysis 11.3.1 The Basic Idea 11.3.2 Illustration of Wilcoxon's Signed Rank Test for Sensitivity Analysis of a Matched Pair Study 11.4 Overview of the Stata Program rboundts 11.5 Examples 11.5.1 Sensitivity Analysis of the Effects of Lead Exposure 11.5.2 Sensitivity Analysis for the Study Using Pair Matching 11.6 Conclusion 12 Concluding Remarks 12.1 Common Pitfalls in Observational Studies: A Checklist for Critical Review 12.2 Approximating Experiments With Propensity Score Approaches 12.2.1 Criticism of Propensity Score Methods 12.2.2 Regression and Propensity Score Approaches: Do They Provide Similar Results? 12.2.3 Criticism of Sensitivity Analysis (T) 12.2.4 Group Randomized Trials 12.3 Other Advances in Modeling Causality 12.4 Directions for Future Development References Index 13 List of Tables Table 1.1 Table 2.1 Table 2.2 Table 2.3 Table 2.4 Table 3.1 Table 3.2 Table 3.3 Table 3.4 Table 3.5 Table 3.6 Table 3.7 Table 3.8 Table 4.1 Table 4.2 Table 4.3 Table 4.4 Table 4.5 Table 4.6 Table 4.7 Table 5.1 Stata and R Procedures by Analytic Methods An Artificial Example of Noncompliance With Encouragement (Wi) to Exercise (Di) Descriptive Statistics of the Study Sample Tests for Treatment Effect Heterogeneity Econometric Versus Statistical Causal Models Comparison of Mortality Rates for Three Smoking Groups in Three Databases Artificial Data of Mortality Rates for Three Smoking Groups Adjusted Mortality Rates Using the Age Standardization Method (i.e., Adjustment Based on the Cigarette Smokers' Age Distribution) Data Description and Estimated Effects by Three Methods: Scenario 1 Data Description and Estimated Effects by Three Methods: Scenario 2 Data Description and Estimated Effects by Three Methods: Scenario 3 Data Description and Estimated Effects by Three Methods: Scenario 4 Data Description and Estimated Effects by Three Methods: Scenario 5 Exhibit of Stata treatreg Output for the NSCAW Study Exhibit of Stata treatreg Output: Syntax to Save Nonselection Hazard Exhibit of Stata treatreg Output: Syntax to Check Saved Statistics Sample Description for the Study Evaluating the Impacts of Caregivers' Receipt of Substance Abuse Services on Child Developmental Well-Being Differences in Psychological Outcomes Before and After Adjustments of Sample Selection Estimated Treatment Effect Models of Fifth Graders' Change on ICST Social Competence Score and on CCC Prosocial Behavior Score Exhibit of Combined Analysis of Treatment Effect Models Based on Multiple Imputed Data Files Exhibit of Stata psmatch2 Syntax and Output Running Greedy 14 Table 5.2 Table 5.3 Table 5.4 Table 5.5 Table 5.6 Table 5.7 Table 5.8 Table 5.9 Table 5.10 Table 5.11 Table 5.12 Table 5.13 Table 5.14 Table 5.15 Table 5.16 Table 5.17 Table 5.18 Table 6.1 Table 6.2 Table 6.3 Matching and Mahalanobis Metric Distance Exhibit of Stata boost Syntax and Output Running Propensity Score Model Using GBM Exhibit of R Syntax and Output Running Logistic Regression and Full Matching Sample Description and Logistic Regression Models Predicting Propensity Scores (Example 5.8.1) Description of Matching Schemes and Resample Sizes (Example 5.8.1) Results of Sensitivity Analyses (Example 5.8.1) Status of Child's Use of AFDC by Status of Caregiver's Use of AFDC in Childhood (Example 5.8.2) Sample Description and Results of Regression Analysis (Example 5.8.2) Results of Optimal Matching (Example 5.8.2) Covariate Imbalance Before and After Matching by Matching Scheme (Example 5.8.2) Estimated Average Treatment Effect on Letter-Word Identification Score in 1997 With Hodges-Lehmann Aligned Rank Test (Matching Scheme: Full Matching) (Example 5.8.3) Regressing Difference Score of Letter-Word Identification on Difference Scores of Covariates After Pair Matching (Example 5.8.4) Propensity Score Models and Imbalance Check for the Change of "CCCPROS" in the Fourth Grade (Example 5.8.5) Estimated Coefficients by Multilevel Model for the Change of "CCCPROS" in the Fourth Grade (Example 5.8.5) Propensity Score Models and Imbalance Check for the Change of "CCCRAGG" in the Third Grade (Example 5.8.5) Estimated Coefficients by Multilevel Model for the Change of "CCCRAGG" in the Third Grade (Example 5.8.5) Comparison of Covariate Imbalance Before and After Matching Between Rand-gbm and Stata's boost (Example 5.8.6) Regressing Difference Score of Outcome (i.e., Change of Academic Competence in Third Grade) on Difference Scores of Covariates After Pair Matching: Comparison of Results Between Rand-gbm and Stata's boost (Example 5.8.6) Estimating Overall Treatment Effect After Stratification (Example 6.5.1) Sample Description and Logistic Regression Predicting Propensity Scores (Example 6.5.2) Balance Check After Trimming and Subclassification Using Quintiles (Example 6.5.2) 15 Table 6.4 Table 6.5 Table 6.6 Table 7.1 Table 7.2 Table 7.3 Table 7.4 Table 8.1 Table 8.2 Table 8.3 Table 8.4 Table 8.5 Table 8.6 Table 8.7 Table 8.8 Table 9.1 Table 9.2 Table 9.3 Estimated Cox Regression Models (Estimated Coefficients) by Stratum (Example 6.5.2) Sample Description and Imbalance Check for the Study of the Poverty Impact (Example 6.5.3) Group Comparison in SEM With Propensity Score Subclassification (Example 6.5.3) Covariate Imbalance After Propensity Score Weighting (Example 7.3.1) Regression Analysis of Letter-Word Identification Score in 1997 With Propensity Score Weighting (Example 7.3.1) Estimated Cox Proportional Hazard Models (Example 7.3.2) Comparison of Findings Across Models Estimating the Impact of Poverty on Children's Academic Achievement (Example 7.3.4) An Example of Simple Matching With One Observed Covariate for Seven Observations An Example of Simple Matching With Three Observed Covariates for Seven Observations With Minimum Distance Determined by Vector Norm Using the Inverse of a Sample Variance Matrix An Example of Simple Matching With Three Observed Covariates for Seven Observations With Minimum Distance Determined by Vector Norm Using the Inverse of a Sample Variance-Covariance Matrix Exhibit of Stata nmatch Syntax and Output Running BiasCorrected Matching Estimators With Robust Standard Errors Estimated Treatment Effects (Effects of Child's Use of AFDC) on Passage Comprehension Standard Score in 1997 Using BiasCorrected Matching With Robust Variance Estimators (Example 8.4.1) Estimated Treatment Effects Measured as Change Scores in the Fourth and Fifth Grades by Three Estimators (Example 8.4.2) Sample Size and Distribution of Exposure Time to Program Intervention ("Dosage") by Grade (Example 8.4.2) Efficacy Subset Analysis Using Matching Estimators: Estimated Average Treatment Effects for the Treated (SATT) by Dosage (Example 8.4.2) Exhibit of Stata psmatch2 and bs Syntax and Output Running Matching With Nonparametric Regression Estimated Average Treatment Effects for the Treated on CBCL Change: Difference in Differences Estimation by Local Linear Regression (Example 9.4.1) Estimated Treatment Effect for the Treated (Child's Use of AFDC) on Passage Comprehension Standard Score in 1997: Comparing the Propensity Score Analysis With Nonparametric Regression With Bias-Corrected Matching and Robust Variance Estimator (Example 9.4.2) Table 10.1 Hirano and Imbens's Example: Balance Given the Generalized Propensity Score—t Statistics for Equality of Means Table 10.2 "Order of Magnitude" Interpretations of the Test Statistics Table 10.3 Exhibit of Stata doseresponse Syntax and Output Running GPS Estimator Table 10.4 Distribution of Dose Categories (Example 10.6.1) Table 10.5 Multinomial Logit Model Predicting Generalized Propensity (Example 10.6.1) Table 10.6 Regression Analysis of the Impact of Dosage of Child AFDC Use on the Letter-Word Identification Score in 1997 With and Without Propensity Score Adjustment (Example 10.6.1) Table 10.7 Balance Given the Generalized Propensity Score: t Statistics for Equality of Means and Bayes Factors (Example 10.6.2) Table 11.1 Key Assumptions and Effects by Correction Model Table 11.2 Results of Monte Carlo Study Comparing Models Table 11.3 Results of Monte Carlo Study Comparing Models Not Controlling for Z Under Setting 1 Table 11.4 Example of Sensitivity Analysis: Blood Lead Levels (µg/dl) of Children Whose Parents Are Exposed to Lead at Their Places of Work Versus Children Whose Parents Are Unexposed to Lead at Their Places of Work Table 11.5 Exhibit of Step 1: Take the Absolute Value of Differences, Sort the Data in an Ascending Order of the Absolute Differences, and Create ds That Ranks the Absolute Value of Differences and Adjusts for Ties Table 11.6 Exhibit of Step 2: Calculate Wilcoxon's Signed Rank Statistic for the Differences in the Outcome Variable Between Treated and Control Groups Table 11.7 Exhibit of Step 3: Calculate Statistics Necessary for Obtaining the One-Sided Significance Level for the Standardized Deviate When F = 1 Table 11.8 Exhibit of Step 4: Calculate Needed Statistics for Obtaining the One-Sided Significance Levels for the Standardized Deviates (i.e., the Lower and Upper Bounds of p Value) When F = 2 Table 11.9 Exhibit of Step 4: Calculate Needed Statistics for Obtaining the One-Sided Significance Levels for the Standardized Deviates (i.e., the Lower and Upper Bounds of p Value) When F = 4.25 Table 11.10 Results of the Sensitivity Analysis for Blood Lead Levels of Children: Range of Significance Levels for the Signed Rank Statistic 17 Table 11.11 Exhibit of Stata rboundts Syntax and Output (Example 11.5.1) Table 11.12 Results of the Sensitivity Analysis for the Study of Children's Letter-Word Identification Score: Range of Significance Levels for the Signed Rank Statistic (Example 11.5.2) Table 11.13 Sensitivity to Hidden Bias in Four Observational Studies 18 List of Figures Figure 2.1 Figure 3.1 Figure 3.2 Figure 3.3 Figure 3.4 Figure 3.5 Figure 4.1 Figure 5.1 Figure 5.2 Figure 5.3 Figure 5.4 Figure 5.5 Figure 6.1 Figure 9.1 Figure 9.2 Figure 9.3 Figure 9.4 Figure 9.5 Figure 9.6 Figure 10.1 Figure 10.2 Figure 11.1 Positive Contemporaneous Correlation Scatterplot of Data Under Scenario 1 Scatterplot of Data Under Scenario 2 Scatterplot of Data Under Scenario 3 Scatterplot of Data Under Scenario 5 Decision Tree for Evaluation of Social Experiments General Procedure for Propensity Score Analysis Illustration of Common Support Region Using Hypothetical Data Survivor Functions: Percentage Remaining No Rereport (Example 5.8.1) Distribution of Estimated Propensity Scores (Example 5.8.2) Comparison of Estimated Propensity Scores Generated by Rand-gbm and Those Generated by Stata's boost (Example 5.8.6) A Sample Conceptual Model Depicting the Mediating Role of a Child's Use of a Welfare Program (AFDC) in the Influence of Caregiver's Use of Welfare Program in Caregiver's Childhood on Child Academic Achievement Illustration of the Need for a Better Curve Smoothing Using Nonparametric Regression The Task: Determining the y Value for a Focal Point x120 Weights Within the Span Can Be Determined by the Tricube Kernel Function The y Value at the Focal Point x120 Is a Weighted Mean The Nonparametric Regression Line Connects All 190 Average Values The Local Average Now Is Predicted by a Regression Line. Instead of a Line Parallel to the x-Axis Estimated Dose-Response Function, Estimated Derivative, and 95% Confidence Bands Estimated Dose-response Function, Estimated Derivative, and 95% Confidence Bands Design of the Monte Carlo Study: Two Settings of Selection Bias 19 To our students, the next generation of social and health researchers whose keen interest in learning advanced statistical models inspired us to write this book 20 Preface 1. WHAT THE BOOK IS ABOUT Propensity Score Analysis describes a family of new statistical techniques that are useful in causal modeling when randomized experimentation is infeasible. Causal inference is the core interest of all sciences. Although the randomized clinical trial is deemed the gold standard for research, true experimental designs are not always possible, practical, ethical, or even desirable, leaving quasiexperimental designs predominant in the behavioral, health, and social sciences. Given the continued reliance on quasi-experimental design and its inherent challenges to evaluation and causality studies, researchers have increasingly sought methods to improve estimates of program effects. Over the past 40 years, researchers have recognized the need for more efficient, effective approaches for assessing treatment effects when evaluating programs based on quasiexperimental designs. In response to this need, significant changes have been introduced to evaluation methods used with quasi-experimental designs. This growing interest led to a surge in work focused on estimating average treatment effects under various sets of assumptions. Statisticians such as Paul Rosenbaum and Donald Rubin (1983) and econometricians such as James Heckman (1978, 1979) made substantial contributions to this movement by developing and refining methods of estimating causal effects from observational data. Collectively, these approaches are known as propensity score analysis (PSA). Written in an accessible fashion, this book is intended to introduce robust and efficient causal models using propensity scores. The book is designed to be a clear, comprehensive, and applied guide to social behavioral and health researchers have a quick reference to start their learning of PSA. Chapters pull together and describe eight PSA methods: (1) Heckman's sample selection model (Heckman, 1978, 1979) and Maddala's (1983) treatment effect model, (2) propensity score greedy matching (Rosenbaum & Rubin, 1983) and optimal matching (Rosenbaum, 2002b), (3) propensity score subclassification (Rosenbaum & Rubin, 1983, 1984), (4) propensity score weighting (Hirano & Imbens, 2001; McCaffrey, Ridgeway, & Morral, 2004), (5) matching estimators (Abadie & Imbens, 2002, 2006), (6) propensity score analysis with nonparametric regression (Heckman, Ichimura, & Todd, 1997, 1998), (7) dosage analysis (Hirano & Imbens, 2004; Imbens, 2000; Joffe & Rosenbaum, 1999), and (8) Rosenbaum's (2002a, 2002b) sensitivity analysis to address hidden selection bias. The book employs two conceptual models of observational studies to guide the learning of these new methods—the Neyman21 Rubin counterfactual framework (Neyman, 1923; Rubin, 1974) and Heckman's (2005) econometric model of causality. Content focuses on two critical but often violated assumptions: the strongly ignorable treatment assignment assumption (Rosenbaum & Rubin, 1983) and stable unit treatment value assumption (Rubin, 1986). Statistical software packages such as Stata and R offer a series of programs that allow users to execute analyses of all models in the book. Chapters provide step-by-step illustrations of all eight PSA models, with data and Stata syntax of all examples available in the book's companion webpage (. 2. NEW IN THE SECOND EDITION Since publication, the book has received positive feedback, and readers of various disciplines have sent a variety of helpful comments. The second edition has addressed all these comments and suggestions and incorporated the latest advances of PSA that are not included in the first edition. All errors in the first edition are corrected. Major changes are summarized in the following. The most significant change of the second edition is the relocation of discussion of propensity score subclassification, propensity score weighting, and dosage analysis from Chapter 5 to separate chapters. These methods are closely related to Rosenbaum and Rubin's (1983) seminal study of the development of propensity scores—it is for this reason that Chapter 5 of the first edition pooled these methods together. Because subclassification and weighting methods have been widely applied in recent research and have become recommended models for addressing challenging data issues (Imbens & Wooldridge, 2009), we decided to give each topic a separate treatment. There is an increasing need in social behavioral and health research to model treatment dosage and to extend the propensity score approach from the binary treatment conditions context to categorical and/or continuous treatment conditions contexts. Given these considerations, we treated dosage analysis in the second edition as a separate chapter. As a result, Chapter 5 now focuses on propensity score matching methods alone, including greedy matching and optimal matching. In addition to these major changes, the second edition reviews new developments in the PSA field. By chapter, we highlight below this new information. Chapter 1 is an overview of PSA and retains the original content of the first edition. The overview of computing software packages updates programs newly developed since the publication of the first edition. The chapter also presents a formula for estimating normalized differences (Imbens & Wooldridge, 2009), a procedure that may be used as an omnibus check to discern whether PSA is needed in a specific data setting. Chapter 2 focuses on conceptual frameworks and the assumptions of 22 observational studies. The chapter provides a more detailed description of other corrective methods, particularly the instrumental variable estimator and regression discontinuity designs. The section on treatment effect heterogeneity is new; with illustration, the section shows how to use tests to evaluate effects heterogeneity and the plausibility of the strongly ignorable treatment assignment assumption. Chapter 5 describes propensity score matching methods. The chapter adds a new section that reviews methods, strategies, and principles of propensity score matching with multilevel data—it describes five types of logistic regression models used in estimating propensity scores when the data are multilevel and clustered, as well as methods of multilevel outcome analysis, including a crossclassified random effect model, that may be conducted following greedy matching. Chapter 6 is new and describes propensity score subclassification. The chapter discusses the principles of and steps in running multivariate outcome analysis (e.g., a Cox proportional hazards model or a structural equation modeling) for each propensity score-generated subclass. It then shows how to aggregate the estimated treatment effects from all subclasses to calculate an overall treatment effect for the entire sample. Because the overlap assumption is likely to be violated in subclassification, the chapter describes a trimming approach developed by Crump, Hotz, Imbens, and Mitnik (2009). The section on the stratification-multilevel method is new and describes the approach that Xie, Brand, and Jann (2012) developed for modeling treatment effect heterogeneity. With illustration, the chapter discusses principles for running structural equation modeling in conjunction with propensity score subclassification. Chapter 7 is new and describes the propensity score weighting estimator; specifically, it shows how to use estimated propensity scores to estimate the sample average treatment effect and the sample average treatment effect for the treated. Compared to the first edition, the chapter provides a more detailed description of the statistical principles of weighting. The examples of using propensity score weighting to conduct a Cox proportional hazards model and a structural equation model are new. Chapter 10 describes dosage analysis based on estimated propensity scores. Developed by Hirano and Imbens (2004), the information on the generalized propensity score estimator is new. The approach is an innovative method to generalize propensity score modeling from binary treatment conditions to categorical and/or continuous treatments settings. It may be used in numerous settings to address challenging data issues and to address questions regarding the differential impact of treatment exposure. Chapter 11 discusses selection bias and Rosenbaum's sensitivity analysis. Extending the first edition, the chapter adds two methods to the Monte Carlo study: propensity score weighting and subclassification. As such, the data 23 simulation compares six models to assess the capacity of each method to correct for selection bias: ordinary least squares (OLS) regression, propensity score greedy matching, Heckman-typed treatment effect model, matching estimator, propensity score weighting to estimate average treatment effect, and propensity score subclassification. The newly added information provides readers with a more comprehensive evaluation of different corrective strategies, and more important, it shows the value of assessing assumptions embedded in specific models when making decisions about an analytic plan. Chapter 12 provides concluding remarks. Given that the PSA field is growing, the chapter adds new information from systematic reviews in epidemiology and medical research using PSA. The information may help readers develop a balanced view about the pros and cons of propensity score modeling, including the importance of following guidelines in running propensity score analyses. One of the findings from these systematic reviews is that users of PSA should always check data balance to make sure that a corrective method has succeeded; without such information, the PSA approach runs into the same risk of suffering from an endogeneity problem as a covariance control model. As a new and rapidly growing class of evaluation methods, PSA is by no means considered the best alternative to randomized experiments, and there is ongoing debate on the advantages and disadvantages of PSA. Chapter 12 presents criticisms of PSA to give readers a balanced perspective. The chapter concludes with a discussion of the "maturity" of PSA and ongoing challenges in causal inference in research. 3. ACKNOWLEDGMENTS We have many people to thank for help in preparing this book. First, we thank the original developers of new methods for observational studies, including James Heckman, Paul Rosenbaum, Donald Rubin, Alberto Abadie, Guido Imbens, Keisuke Hirano, Hidehiko Ichimura, and Petra Todd, whose contributions in developing the eight statistical models for observational studies make this book possible. We thank Richard Berk, the editor in chief of the first edition, for his many innovative ideas and helpful guidance in preparing the first edition. When writing this book, we received invaluable comments, suggestions, and direct help from Paul Rosenbaum, Kenneth Bollen, Yu Xie, Ben Hansen, Guido Imbens, Richard Crump, Petra Todd, John Fox, Michael Foster, and two anonymous reviewers of the first edition. Shenyang Guo thanks his mentor William Mason, whose innovative and rigorous research in statistics greatly shaped his career and led him to choose quantitative methodology as his main research area. We thank our former acquisitions editor Lisa Shaw and the current senior acquisitions editor Vicki Knight at Sage Publications for their help in preparing 24 the book. We thank many of our colleagues at the University of North Carolina at Chapel Hill and in the field of social-work research. Dean Jack Richman's vision of promoting research rigor, which was shown by his full support of this project, motivated us to write this book. Richard Barth engaged in the original discussion in designing the book and contributed excellent suggestions for its completion—we cited several studies led by Barth to illustrate the applications of targeted methods in social work research. Diane Wyant provided excellent editorial help for the entire book. Alan Eells helped with programming in R for some examples. Jung-Sook Lee helped manage the PSD and CDS data that were employed in several examples. Carrie Pettus Davis and Jilan Li helped with the search for computing procedures currently available in the field. We thank Cyrette Cottien-Fleming for assistance with copying and other logistics. Part of the financial support was provided by the John A. Tate Distinguished Professorship held by Mark Fraser and the Wallace H. Kuralt, Sr. Distinguished Professorship held by Shenyang Guo. We thank the following reviewers for their time, effort, and feedback: Jason Barabas, Stony Brook University; Priyalatha Govindasamy, University of Denver; Joyce P. Jacobse, Wesleyan University; Antonio Olmos, University of Denver; and James Prieger, Pepperdine University. Finally, we thank our families for their support, understanding, and patience. Specifically, Shenyang Guo thanks his wife Shenyan Li and his children Han and Hanzhe, and Mark Fraser thanks his wife Mary Fraser and his children Alex and Katy. This book is dedicated to you all! 25 About the Authors Shenyang Guo, PhD, holds the Frank J. Bruno Distinguished Professorship of Social Work Research at the George Warren Brown School of Social Work, Washington University in St. Louis. He was the Wallace H. Kuralt, Sr. Distinguished Professor at the School of Social Work at the University of North Carolina at Chapel Hill. He has a MA in economics from Fudan University and a PhD in Sociology from the University of Michigan. He is the author of numerous research articles in child welfare, child mental health services, welfare, and health care. He has expertise in applying advanced statistical models to solving social welfare problems and has taught graduate courses that address event history analysis, hierarchical linear modeling, growth curve modeling, structural equation modeling, and program evaluation. He has given many invited workshops on statistical methods—including event history analysis and propensity score analysis—to the NIH Summer Institute, Children's Bureau, the Society of Social Work and Research conferences, and Statistical Horizons. He served as the Director of Applied Statistical Working Group at UNC. He led the data analysis planning for the National Survey of Child and Adolescent Well-Being (NSCAW) longitudinal analysis and has developed analytic strategies that address issues of weighting, clustering, growth modeling, and propensity score analysis. He also directed the analysis of data from the Making Choices Project, a prevention trial funded by the National Institute on Drug Abuse (NIDA). He has published many articles that include methodological work on the analysis of longitudinal data, multivariate failure time data, program evaluation, and multilevel modeling. He is on the editorial board of Social Service Review and a frequent guest reviewer for journals seeking a critique of advanced methodological analyses. Mark W. Fraser, PhD, holds the John A. Tate Distinguished Professorship for Children in Need at the School of Social Work, University of North Carolina at Chapel Hill. He has written numerous chapters and articles on risk and resilience, child behavior, child and family services, and research methods. With colleagues, he is the coauthor or editor of nine books. These include Families in Crisis, a study of intensive family-centered services, and Evaluating Family-Based Services, a text on methods for family research. In Risk and Resilience in Childhood, he and his colleagues describe resilience-based perspectives for child maltreatment, school dropout, substance abuse, violence, unwanted pregnancy, and other social problems. In Making Choices, 26 Dr. Fraser and his coauthors outline a program to help children build enduring social relationships with peers and adults. In The Context of Youth Violence, he explores violence from the perspective of resilience, risk, and protection, and in Intervention With Children and Adolescents, Fraser and his colleagues review advances in intervention knowledge for social and health problems. His text Social Policy for Children and Families reviews the bases for public policy in child welfare, juvenile justice, mental health, developmental disabilities, and health. Finally, in Intervention Research: Developing Social Programs, he and his colleagues describe five steps in the design and development of interventions. 27 CHAPTER 1 Introduction Propensity score analysis is a class of statistical methods developed for estimating treatment effects with nonexperimental or observational data. Specifically, propensity score analysis offers an approach to program evaluation when randomized trials are infeasible or unethical, or when researchers need to assess treatment effects from survey data, census data, administrative data, medical records data, or other types of data "collected through the observation of systems as they operate in normal practice without any interventions implemented by randomized assignment rules" (Rubin, 1997, p. 757). In the social and health sciences, researchers often face a fundamental task of drawing conditioned causal inferences from quasi-experimental studies. Analytical challenges in making causal inferences can be addressed by a variety of statistical methods, including a range of new approaches emerging in the field of propensity score analysis. This book focuses on seven closely related but technically distinct models for estimating treatment effects: (1) Heckman's sample selection model (Heckman, 1976, 1978, 1979) and its revised version (Maddala, 1983), (2) propensity score matching (Rosenbaum, 2002b; Rosenbaum & Rubin, 1983) and related models, (3) propensity score subclassification (Rosenbaum & Rubin, 1983, 1984), (4) propensity score weighting (Hirano & Imbens, 2001; Hirano, Imbens, & Ridder, 2003; McCaffrey, Ridgeway, & Morral, 2004), (5) matching estimators (Abadie & Imbens, 2002, 2006), (6) propensity score analysis with nonparametric regression (Heckman, Ichimura, & Todd, 1997, 1998), and (7) propensity score analysis of categorical or continuous treatments (Hirano & Imbens, 2004; Imbens, 2000; Joffe & Rosenbaum, 1999). Although statisticians and econometricians have not reached consensus on the scope and content of propensity score analysis, the statistical models described in this book share several similar characteristics: Each has the objective of assessing treatment effects and controlling for covariates, each represents state-of-the-art analysis in program evaluation, and each can be employed to overcome various kinds of challenges encountered in research. Although the randomized controlled trial is deemed to be the gold standard in research design, true experimental designs are not always possible, practical, or even desirable in the social and health sciences. Given a continuing reliance on 28 quasi-experimental design, researchers have increasingly sought methods to improve estimates of program effects. Over the past 35 years, methods of program evaluation have undergone a significant change as researchers have recognized the need to develop more efficient approaches for assessing treatment effects from studies based on observational data and for evaluations based on quasi-experimental designs. This growing interest in seeking consistent and efficient estimators of program effectiveness led to a surge in work focused on estimating average treatment effects under various sets of assumptions. Statisticians

(e.g., Rosenbaum & Rubin, 1983) and econometricians (e.g., Heckman, 1978, 1979) have made substantial contributions by developing and refining new approaches for the estimation of causal effects from observational data. Collectively, these approaches are known as propensity score analysis. Econometricians have integrated propensity score models into other econometric models (i.e., instrumental variable, control function, difference-in-differences estimators) to perform less expensive and less intrusive nonexperimental evaluations of social, educational, and health programs. Furthermore, recent criticism and reformulations of the classical experimental approach in econometrics symbolize an important shift in evaluation methods. The significance of this movement was evidenced by the selection of James Heckman as one of the 2000 Nobel Prize award winners in the field of economics. The prize recognized his development of theory and methods for data analysis in selective samples. As a new and rapidly growing class of evaluation methods, propensity score analysis is by no means conceived as the best alternative to randomized experiments. In empirical research, it is still unknown under what circumstances the approach appears to reduce selection bias and under what circumstances the conventional regression approach (i.e., use of statistical controls) remains adequate. There are certainly debates about the advantages and disadvantages of propensity score modeling. These focus, primarily, on the extent to which propensity score methods offer effective and efficient estimates of treatment effects and on the degree to which they help address many challenging issues embedded in program evaluation, policy evaluation, and causal inference. The call for developing and using strong research designs to provide a comprehensive understanding of causal processes in program evaluation remains a paramount challenge in all fields of practice. However, it is also a consensus among prominent researchers that the propensity score approach has reached a mature level. For instance, Imbens and Wooldridge (2009) evaluated recent developments in the econometrics of program evaluation, primarily the methods described by this book, and concluded that at this stage, the literature has matured to the extent that it has much to offer the empirical researchers. Although the evaluation problem is one where 29 identification problems are important, there is currently a much better understanding of which assumptions are most useful, as well as a better set of methods for inference given different sets of assumptions. (p. 8)

Representing the interest in—and indeed perceived utility of—these new methods, the propensity score approach has been employed in a variety of disciplines and professions such as education (Morgan, 2001), epidemiology (Normand et al., 2001), medicine (e.g., Earle et al., 2001; Gum, Thamilarasan, Watanabe, Blackstone, & Lauer, 2001), psychology (Jones, D’Agostino, Gondolf, & Heckert, 2004), social work (Barth, Greeson, Guo, & Green, 2007; Barth, Lee, Wildfire, & Guo, 2006; Guo, Barth, & Gibbons, 2006; Weigensberg, Barth, & Guo, 2009), and sociology (H. L. Smith, 1997). In social welfare studies, economists and others used propensity score methods in evaluations of the National Job Training Partnership Act program (Heckman, Ichimura, & Todd, 1997), the National Supported Work Demonstration (LaLonde, 1986), and the National Evaluation of Welfare-to-Work Strategies Study (Michalopoulos, Bloom, & Hill, 2004). In describing these new methods, the preparation and writing of this book was guided by two primary objectives. The first objective was to introduce readers to the origins, main features, and debates centering on the seven models of propensity score analysis. We hope this introduction will help accomplish our second objective of illuminating new ideas, concepts, and approaches that social and health sciences researchers can apply to their own fields to solve problems they might encounter in their research efforts. In addition, this book has two overarching goals. Our primary goal is to make the past three decades of theoretical and technological advances in analytic methods accessible and available in a less technical and more practical fashion. The second goal is to promote discussions among social and health sciences researchers regarding emerging strategies and methods for estimating causal effects using nonexperimental methods. The aim of this chapter is to provide an overview of the propensity score approach. Section 1.1 presents a definition of observational study. Section 1.2 reviews the history and development of the methods. Section 1.3 is an overview of the randomized experimental approach, which is the gold standard developed by statisticians and the model that should serve as a foundation for the nonexperimental approach. Section 1.4 offers examples drawn from literature beyond the fields of econometrics and statistics. These examples are intended to help readers determine the situations in which the propensity score approach may be appropriate. Section 1.5 reviews the computing software packages that are currently available for propensity score analysis and the main features of the package used in the models presented throughout this book. Section 1.6 outlines the organization of the book.

1.1 OBSERVATIONAL STUDIES

The statistical methods we discuss may be generally categorized as methods for observational studies. According to Cochran (1965), an observational study is an empirical investigation whose objective is to elucidate causal relationships (i.e., cause and effect) when it is infeasible to use controlled experimentation and to assign participants at random to different procedures. In the general literature related to program evaluation (i.e., nonstatistically oriented literature), researchers use the term quasi-experimental more frequently than observational studies, with the term defined as studies that compare groups but lack the critical element of random assignment. Indeed, quasi-experiments can be used interchangeably with observational studies, as described in the following quote from Shadish, Cook, and Campbell (2002): Quasi-experiments share with all other experiments a similar purpose—to test descriptive causal hypotheses about manipulable causes—as well as many structural details, such as the frequent presence of control groups and pretest measures, to support a counterfactual inference about what would have happened in the absence of treatment. But, by definition, quasiexperiments lack random assignment. Assignment to conditions is by means of self-selection, by which units choose treatment for themselves, or means of administrator selection, by which teachers, bureaucrats, legislators, therapists, physicians, or others decide which persons should get which treatment. (pp. 13–14)

Two features of observational studies merit particular emphasis. First, an observational study concerns treatment effects. A study without a treatment—often called an intervention or a program—is neither an experiment nor an observational study. Most public opinion polls, forecasting efforts, investigations of fairness and discrimination, and many other important empirical studies are neither experiments nor observational studies (Rosenbaum, 2002b). Second, observational studies can employ data from nonexperimental, nonobservational studies as long as the focus is on assessing treatment or the effects of receiving a particular service. By this definition, observational data refer to data that were generated by something other than a randomized experiment and typically include surveys, censuses, or administrative records (Winship & Morgan, 1999).

1.2 HISTORY AND DEVELOPMENT

The term propensity score first appeared in a 1983 article by Rosenbaum and Rubin, who described the estimation of causal effects from observational data. Heckman’s (1978, 1979) work on dummy endogenous variables using 31 simultaneous equation modeling addressed the same issue of estimating treatment effects when assignment was nonrandom; however, Heckman approached this issue from a perspective of sample selection. Although Heckman’s work on the dummy endogenous variable problem employed different terminology, he used the same approach toward estimating a participant’s probability of receiving one of two conditions. Both schools of thought (i.e., the econometric tradition of Heckman and the statistical tradition of Rosenbaum and Rubin) have had a significant influence on the direction of the field, although the term propensity score analysis, coined by Rosenbaum and Rubin, is used more frequently as a general term for the set of related techniques designed to correct for selection bias in observational studies. The development of the propensity score approach signified a convergence of two traditions in studying causal inferences: the econometric tradition that primarily relies on structural equation modeling and the statistical tradition that primarily relies on randomized experiments (Angrist, Imbens, & Rubin, 1996; Heckman, 2005). The econometric tradition dates back to Trygve Haavelmo (1943, 1944), whose pioneering work developed a system of linear simultaneous equations that allowed analysts to capture interdependence among outcomes, to distinguish between fixing and conditioning on inputs, and to parse out true causal effects and spurious causal effects. The task of estimating counterfactuals, a term generally developed and used by statisticians, is explored by econometricians in the form of a switching regression model (Maddala, 1983; Quandt, 1958, 1972). Heckman’s (1978, 1979) development of a two-step estimator is credited as the field’s pioneering work in explicitly modeling the causes of selection in the form of a dummy endogenous variable. As previously mentioned, Heckman’s work followed econometric conventions and solved the problem through structural equation modeling. Historically quite distinct from the econometric tradition, the statistical tradition can be traced back to Fisher (1935/1971), Neyman (1923), and Rubin (1974, 1978). Unlike conventions based on linear simultaneous equations or structural equation models, the statistical tradition is fundamentally based on the randomized experiment. The principal notion in this formulation is the study of potential outcomes, known as the Neyman-Rubin counterfactual framework. Under this framework, the causal effects of treatment on sample participants (already exposed to treatments) are explored by observing outcomes of participants in samples not exposed to the treatments. Rubin extended the counterfactual framework to more complicated situations, such as observational studies without randomization. For a detailed discussion of these two traditions, readers are referred to a special issue of the *Journal of the American Statistical Association* (1996, Vol. 91, No. 434), which presents an interesting dialogue between statisticians and econometricians. Significant scholars in the field—including Greenland, Heckman, Moffitt, Robins, and Rosenbaum—participated in a discussion of a 32 study that used instrumental variables to identify causal effects, particularly the local average treatment effect (Angrist et al., 1996). It is worth noting that the development of propensity score models did not occur in isolation from other important developments. At the same time that propensity score methods were emerging, the social, behavioral, and health sciences witnessed progress in the development of other statistical methods, such as methods for the control of clustering in multilevel data—for instance, the linear mixed model (Laird & Ware, 1982), hierarchical linear modeling (Raudenbush & Bryk, 2002), and robust standard error estimator (Huber, 1967; White, 1980); methods to analyze latent variables and to model complex structural relationships among latent variables (e.g., analyzing moderating as well as mediating effects, or models to depict nonrecursive relationship between latent variables)—that is, the structural equation modeling (Bollen, 1989; Jöreskog, 1971); methods for analyzing categorical and limited dependent variables—that is, the generalized linear models (Nelder & Wedderburn, 1972); methods for analyzing time-to-event data—for instance, the proportional hazards model (Cox, 1972) and marginal approaches to clustered event data (Lee, Wei, & Amato, 1992; Wei, Lin, & Weissfeld, 1989); and more. When researchers are engaged in observational studies, many of these newly developed models need to be applied in conjunction with propensity score methods, and by the same token, a successful propensity score analysis always requires a careful examination of other issues of data analysis, including addressing potential violations of statistical assumptions by employing these newly developed methods. In this book, whenever possible, we describe the application of propensity score models in settings where other data issues are present, and we show how to employ propensity score models in conjunction with the application of other statistical approaches.

1.3 RANDOMIZED EXPERIMENTS

The statistical premise of program evaluation is grounded in the tradition of the randomized experiment. Therefore, a natural starting point in a discussion of causal attribution in observational studies is to review key features of the randomized experiment. According to Rosenbaum (2002b), a theory of observational studies must have a clear conceptual linkage to randomization, so that the consequences of the absence of randomization can be understood. For example, sensitivity analysis is among Rosenbaum’s approaches to handling data with hidden selection bias; this approach includes the use of test statistics that were developed primarily for randomized experiments, such as Wilcoxon’s signed rank statistical test and Hodges-Lehmann estimates. However, the critiques of social experiments by econometricians (e.g., Heckman & Smith, 1995) frequently include description of the conditions under which randomization is infeasible, particularly under the setting of social behavioral research. Thus, it is important to review principles and types of randomized experiments, randomization tests, and the challenges to this tradition. Each of these topics is addressed in the following sections.

1.3.1 Fisher’s Randomized Experiment

The invention of the randomized experiment is generally credited to Sir Ronald Fisher, one of the foremost statisticians of the 20th century. Fisher’s book, *The Design of Experiments* (1935/1971), introduced the principles of randomization, demonstrating them with the now-famous example of testing a British woman’s tea-tasting ability. This example has been cited repeatedly to illustrate the power of randomization and the logic of hypothesis testing (see, e.g., Maxwell & Delaney, 1990; Rosenbaum, 2002b). In a somewhat less technical fashion, we include this example as an illustration of important concepts in randomized experimentation. In Fisher’s (1935/1971) words, the problem is as follows: A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested. (p. 11)

During Fisher’s time, the dominant practice in experimentation was to control covariates or confounding factors that might contaminate treatment effects. Therefore, to test a person’s tasting ability (i.e., the true ability to discriminate two methods of tea preparation), a researcher would control factors that could influence the results, such as the temperature of the tea, the strength of the tea, the use of sugar, and the amount of milk added, in addition to the myriad potential differences that might occur among the cups of tea used in an experiment. As Maxwell and Delaney (1990) pointed out, The logic of experimentation up until the time of Fisher dictated that to have a valid experiment here all the cups to be used “must be exactly alike,” except for the independent variable being manipulated. Fisher rejected this dictum on two grounds. First, he argued that it was logically impossible to achieve, both in the example and in experimentation in general. . . . Second, Fisher argued that, even if it were conceivable to achieve “exact likeness,” or more realistically, “imperceptible difference” on various dimensions of the stimuli, it would in practice be too expensive to attempt. (p. 40)

Instead of controlling for every potential confounding factor, Fisher proposed to control for nothing, namely, to employ a method of randomization. Fisher (1935/1971) described his design as follows: 34 Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of what the test will consist, namely that she will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in a random order, that is in an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance, cards, dice, roulettes, etc., or more expeditiously, from a published collection of random sampling numbers purporting to give the actual results of such a manipulation. Her task is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received. (p. 12)

Before going further, it is crucial to note several important points regarding Fisher’s design. First, in this example, the unit of analysis is not individual ($N \neq 1$), but rather the presentation of the tea cups to the tea taster (i.e., $N = 8$, in which a total of 8 cases comprise the sample). Second, there is a treatment assignment process in this example, namely, the order of presentation of tea cups. Using Rosenbaum’s (2002b) notation, this is a random variable Z , and any specific presentation of the tea cups to the taster is a realization of Z , or $Z = z$. For instance, if a specific presentation of the tea cups consists of four cups with milk added first followed by four cups with tea added first, then we may write $z = (11110000)$, where z is just one of many possible assignments. In Rosenbaum’s notation, these possible treatment assignments form a set of \cdot and $z \in \cdot$. Determining the total number of elements in (which Rosenbaum denoted as K) is an important task for experimental design and a task that can be accomplished using probability theory. This point will be discussed in more detail elsewhere. Third, there is an actual outcome r , which is the result of tasting the eight cups of tea. If the taster gives exactly the same order of tea cups as in the treatment assignment (i.e., she correctly identifies the first four cups as having the milk added first and the next four cups as having the tea added first), then the outcome would be recorded as $r = (11110000)$. Last, the test essentially aims to determine whether the tea taster had the true ability to discriminate the two kinds of tea or whether she made her correct judgment accidentally by guessing. Thus, the null hypothesis (H_0) under testing would be “She has no ability to discriminate,” and the test involves finding statistical evidence to reject the null hypothesis at a given significance level. Building on these explanations, we continue with the tea-tasting test and describe how Fisher implemented his randomized experiment. One important feature of randomized experiments is that, in advance of implementation, the researcher must calculate probable outcomes for each study unit. Fisher (1935/1971) emphasized “forecasting all possible outcomes,” even at a design stage when outcome data are completely absent: “In considering the 35 appropriateness of any proposed experimental design, it is always needful to forecast all possible results of the experiment, and to have decided without ambiguity what interpretation shall be placed upon each one of them” (p. 12). The key of such calculation is to know the total number of elements in the set of (i.e., the value of K). In the above example, we simply made an arbitrary example of treatment assignment 11110000, although many other treatment assignments can be easily figured out, such as alternating cups of tea with the milk added first with the cups prepared by adding the tea infusion first (i.e., 10101010), or presenting four cups with tea infusion added first and then four cups with milk added first (i.e., 00001111). In statistics, the counting rules (i.e., permutations and combinations) inform us that the number of total possible ways to present the eight cups can be solved by finding out the number of combinations of eight things taken four at a time, or $8C4$, as The solution to our problem is Therefore, there are 70 possible ways to present the tea taster with four cups with milk added first and four cups with tea added first. We can keep writing 11110000, 10101010, 00001111, . . . until we exhaust all 70 ways. Here, 70 is the number of total elements in the set of Ω or all possibilities for a treatment assignment. To perform a statistical test of “ H_0 : No ability,” Fisher turned to the task of looking into the possible outcomes r . Furthermore, if we define the taster’s true ability to taste discriminately as requiring all eight cups she identified to match exactly what we presented to her, we can then calculate the probability of having the true outcome. The significance test performed here involves rejecting the null hypothesis, and the null hypothesis is expressed as “no ability.” Fisher used the logic of “guessing the outcome right”; that is, the taster has no ability to discriminate but makes her outcome correct by guessing. Thus, what is the probability of having the outcome r that is identical to the treatment assignment z ? The outcome r should have one set of values from the 70 possible treatment assignments; that is, the taster could guess any outcome from 70 possible outcomes of 11110000, 10101010, 00001111, Therefore, the probability of guessing the right outcome is $1/70 = .0124$, which is a very low probability. Now, we can reject the null hypothesis under a small probability of making a Type I error (i.e., the tea taster did have the ability, but we erroneously rejected the “no ability” hypothesis), and the chance is indeed very low (.0124). In other words, based on statistical evidence (i.e., an examination of all possible 36 outcomes), we can reject the “no ability” hypothesis at a statistical significance level of .05. Thus, we may conclude that under such a design, the taster may have true tasting ability ($p < .05$). Rosenbaum (2002b) used $t(Z, r)$ to denote the test statistic. In the preceding test scenario, we required a perfect match—a total of eight agreements—between the treatment (i.e., the order of tea cups presented to the tea-tasting expert) and the outcome (i.e., the actual outcome identified by the taster); therefore, the problem is to find out the probability $\{t(Z, r) > 8\}$. This probability can be more formally expressed in Rosenbaum’s notation as follows: However, if the definition of “true ability” is relaxed to allow for six exact agreements rather than eight agreements (i.e., six cups in the order of outcome match to the order of presentation), we can still calculate the probability or significance in testing the null hypothesis of “no ability.” As in the earlier computation, this calculation involves the comparison of actual outcome r to the treatment assignment z , and the tea taster’s outcome could be any one of 70 possible outcomes. Let us assume that the taster gives her outcome as $r = (11110000)$. We now need to examine how many treatment assignments (i.e., number of z) match this outcome under the relaxed definition of “true ability.” The answer to this question is one perfect match (i.e., the match with eight agreements) plus 16 matches with six agreements (Rosenbaum, 2002b, p. 30), for a total of 17 treatment assignments. To illustrate, we provide all 17 treatment assignments that match to the taster’s outcome 11110000: perfect match 11110000, and the following assignments with six exact agreements: 01111000, 01110001, 01110010, 01110100, 10110100, 10110010, 10110001, 10110100, 10100100, 10100001, 11100010, 11100100, 11100001, 11100010, 11100100, where bold numbers indicate agreements. 2 Thus, the probability of having six exact agreements is $17/70 = .243$. In Rosenbaum’s notation, the calculation is That is, if we define “true ability” as correctly identifying six out of eight cups of tea, the probability of having a correct outcome increases to .243. The 37 null hypothesis cannot be rejected at a .05 level. In other words, under this relaxed definition, we should be more conservative, or ought to be more reluctant, to declare that the tea taster has true ability. With a sample of eight cups in total and a relaxed definition of “ability,” the statistical evidence is simply insufficient for us to reject the null hypothesis, and, therefore, the experimental design is less significant in testing true tasting ability. We have described Fisher’s famous example of randomized experiment in great detail. Our purpose of doing so is twofold. The first is to illustrate the importance of understanding two processes in generating intervention data: (1) the treatment assignment process (i.e., there is a random variable Z , and the total number of possible ways K is inevitably large) makes it possible to know in advance the probability of receiving treatment in a uniform randomized experiment and (2) the process of generating outcome data (i.e., there is an outcome variable r). This topic is revisited both in Chapters 2 and 3, in the discussion of the so-called ignorable treatment assignment, and in Chapter 11, in the discussion of selection bias and sensitivity analysis. The second purpose in providing a detailed description of Fisher’s experiment was to call attention to the core elements of randomized experiments. According to Rosenbaum (2002b), First, experiments do not require, indeed cannot reasonably require, that experimental units be homogeneous, without variability in their responses. . . . Second, experiments do not require, indeed, cannot reasonably require, that experimental units be a random sample from a population of units. . . . Third, for valid inference about the effects of a treatment on the units included in an experiment, it is sufficient to require that treatments be allocated at random to experimental units—these units may be both heterogeneous in their responses and not a sample from a population. Fourth, probability enters the experiment only through the random assignment of treatments, a process controlled by the experimenter. (p. 23)

1.3.2 Types of Randomized Experiments and Statistical Tests

Fisher’s framework laid the foundation for randomized experimental design. The method has become a gold standard in program evaluation and continues to be an effective and robust means for assessing treatment effects in nearly every field of interest from agriculture and business, to computer science, to education, to medicine and social welfare. Furthermore, many sophisticated randomized designs have been developed to estimate various kinds of treatment effects under various settings of data generation. For example, within the category of uniform randomized experiment, 3 in addition to the traditional method of completely randomized experiment, where stratification is absent 38 (i.e., $S = 1$ and S stands for number of strata), researchers have developed randomized block experiments where two or more strata are permissible (i.e., $S \geq 2$) and paired randomized experiments in which $nS = 2$ (i.e., the number of study participants within stratum S is fixed at 2), $mS = 1$ (i.e., the number of participants receiving treatment within stratum S is fixed at 1), and S could be reasonably large (Rosenbaum, 2002b). A more important reason for studying randomized experiments is that statistical tests developed through randomized experiments may be performed virtually without assumptions, which is not the case for nonrandomized experiments. The class of randomization tests, as reviewed and summarized by Rosenbaum (2002b), includes 1. Tests for binary outcomes: Fisher’s (1935/1971) exact test, the Mantel-Haenszel (1959) statistic, and McNemar’s (1947) test 2. Tests for an outcome variable that is confined to a small number of values representing a numerical scoring of several ordered categories (i.e., an ordinal variable): Mantel’s (1963) extension of the Mantel-Haenszel test 3. Tests for a single stratum $S = 1$, where the outcome variable may take many numerical values (i.e., an interval or ratio variable): Wilcoxon’s (1945) rank sum test 4. Tests for an outcome variable that is ordinal and the number of strata S is large compared with sample size N : the Hodges and Lehmann (1962) test using the signed rank statistic As opposed to drawing inferences using these tests in randomized designs, drawing inferences using these tests in nonrandomized experiments “requires assumptions that are not at all innocuous” (Rosenbaum, 2002b, p. 27).

1.3.3 Critiques of Social Experimentation

Although the randomized experiment has proven useful in many applications since Fisher’s seminal work, the past three decades have witnessed a chorus of challenges to the fundamental assumptions embedded in the experimental approach. In particular, critics have been quick to note the complexities of applying randomized trials in studies conducted with humans rather than mechanical components, agricultural fields, or cups of tea. The dilemma presented in social and health sciences studies with human participants is that assigning participants to a control condition means potentially denying treatment or services to those participants; in many settings, such denial of services would be unethical or illegal. Although the original rationale for using a randomized experiment was the infeasibility of controlling covariates, our evaluation needs have returned to the point where covariant control or its variants (e.g., 39 matching) becomes attractive. This is particularly true in social behavioral evaluations. In a series of publications, Heckman and his colleagues (e.g., Heckman, 1979; Heckman & Smith, 1995) discussed the importance of directly modeling the process of assigning study participants to treatment conditions by using factors that influence participants’ decisions regarding program participation. Heckman and his associates challenged the assumption that we can depend on randomization to create groups in which the treated and nontreated participants share the same characteristics under the condition of nontreatment. They questioned the fundamental assumption embedded in the classical experiment: that randomization removes selection bias. Heckman and Smith (1995) in particular held that social behavioral evaluations need to explicitly address four questions, none of which can be handled suitably by the randomized experiment: (1) What are the effects of factors such as subsidies, advertising, local labor markets, family income, race, and gender on program application decisions? (2) What are the effects of bureaucratic performance standards, local labor markets, and individual characteristics on administrative decisions to accept applicants and place them in specific programs? (3) What are the effects of family background, subsidies, and local market conditions on decisions to drop out of a program and, alternatively, on the length of time required to complete a program? (4) What are the costs of various alternative treatments? 1.4 WHY AND WHEN A PROPENSITY SCORE ANALYSIS IS NEEDED

Drawing causal inferences in observational studies or studies without randomization is challenging, and it is this task that has motivated statisticians and econometricians to explore new analytic methods. The seven analytic models that we discuss in this book derive from that. Although the models differ on the specific means employed, all seven models aim to accomplish data balancing when treatment assignment is nonignorable, to evaluate treatment effects using nonrandomized or nonexperimental approaches, and/or to reduce multidimensional covariates to a one-dimensional score called a propensity score. To provide a sense of why and when propensity score methods are needed, we use examples drawn from the literature across various disciplines. Propensity score analysis is suitable to data analysis and to causal inferences for a variety of studies. Most of these examples will be revisited throughout this book. Example 1: Assessing the Impact of Catholic Versus Public Schools on Learning. A long-standing debate in education is whether Catholic schools (or 40 private schools in general) are more effective than public schools in promoting learning. Obviously, a variety of selections are involved in the formation of “treatment” (i.e., entrance into Catholic schools). To name a few, self-selection is a process that lets those who choose to study in Catholic schools receive the treatment; school selection is a process that permits schools to select only those students who meet certain requirements, particularly minimum academic standards, to enter into the treatment; financial selection is a process that excludes from the treatment those students whose families cannot afford tuition; and geographic selection is a process that selects out (i.e., excludes) students who live in areas where no Catholic school exists. Ultimately, the debate on Catholic schools centers on whether differences observed in outcome data (i.e., academic achievement or graduation rates) between Catholic and public schools are attributable to the intervention or to the fact that the Catholic schools serve a different population. In other words, if the differences are attributable to the intervention, findings suggest that Catholic schools promote learning more effectively than do public schools, whereas if the differences are attributable to the population served by Catholic schools, findings would show that students currently enrolled in Catholic schools would always demonstrate better academic outcomes regardless of whether they attended private or public schools. It is infeasible to conduct a randomized experiment to answer these questions; however, observational data such as the National Educational Longitudinal Survey (NELS) data are available to researchers interested in this question. Because observational data lack randomized assignment of participants into treatment conditions, researchers must employ statistical procedures to balance the data before assessing treatment effects. Indeed, numerous published studies have used the NELS data to address the question of Catholic school effectiveness; however, the findings have been contradictory. For instance, using propensity score matching and the NELS data, Morgan (2001) found that the Catholic school effect is the strongest only among those Catholic school students who, according to their observed characteristics, are least likely to attend Catholic schools. However, in a study that used the same NELS data but employed a new method that directly assessed selectivity bias, Altonji, Elder, and Taber (2005) found that attending a Catholic high school substantially increased a student’s probability of graduating from high school and, more tentatively, attending college. Example 2: Assessing the Impact of Poverty on Academic Achievement. Prior research has shown that exposure to poverty and participation in welfare programs have strong impacts on child development. In general, growing up in poverty adversely affects a child’s life prospects, and the consequences become more severe with greater exposure to poverty (Duncan, Brooks-Gunn, Yeung, & Smith, 1998; Foster & Furstenberg, 1998, 1999; P. K. Smith & Yeung, 1998). 41 Most prior inquiries in this field have applied a multivariate analysis (e.g., multiple regression or regression-type models) to samples of nationally representative data such as the Panel Study of Income Dynamics (PSID) or administrative data, although a few studies employed a correction method such as propensity score analysis (e.g., Yoshikawa, Magusson, Bos, & Hsueh, 2003). Using a multivariate approach with this type of data poses two fundamental problems. First, the bulk of the literature regarding the impact of poverty on children’s academic achievement assumes a causal perspective (i.e., poverty is the cause of poor academic achievement), whereas the analysis using a regression model is, at best, correlational. In addition, a regression model or covariance control approach is less robust in handling endogeneity bias. Second, PSID is an observational survey without randomization and, therefore, researchers must take selection bias into consideration when employing PSID data to assess causal effects. Guo and Lee (2008) have made several efforts to examine the impacts of poverty. First, using PSID data, propensity score models—including optimal propensity score matching, the treatment effects model, and the matching estimator—were used to estimate the impact of poverty. Second, Guo and Lee conducted a more thorough investigation of poverty. That is, in addition to conventional measures of poverty such as the ratio of income to poverty threshold, they examined 30 years of PSID data to create two new variables: (1) the number of years during a caregiver’s childhood (i.e., ages 6–12 years) that a caregiver used Aid to Families With Dependent Children (AFDC) and (2) the percentage of time a child used AFDC between birth and 1997 (i.e., the time point when academic achievement data were compared). Last, Guo and Lee conducted both efficacy subanalysis and intent-to-treat analysis and compared findings. Results using these approaches were more revealing than previous studies. Example 3: Assessing the Impact of a Waiver Demonstration Program. In 1996, the U.S. Congress approved the Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA). Known as welfare reform, PRWORA ended entitlements to cash assistance that were available under the prior welfare policy, AFDC. As part of this initiative, the federal government launched the Waiver Demonstration program, which allowed participating states and counties to use discretionary funding for county-specific demonstration projects of welfare reform—as long as these demonstrations facilitated “cost neutrality.” A key feature of the Waiver Demonstration program, as well as several other programs implemented under welfare reform, is that the county has the option of whether to participate in the Waiver Demonstration. Therefore, by definition, the intervention counties and comparison counties at the state level cannot be formed randomly. Counties that chose to participate differed from counties choosing not to participate. 42 Evaluating such a nonrandomized program is daunting. Using a Monte Carlo study, Guo and Wildfire (2005) demonstrated that propensity score matching is a useful analytic approach for such data and an approach that provides less biased findings than does an analysis using the state-level population data. Example 4: Assessing the Well-Being of Children Whose Parents Abuse Substances. A strong, positive association between parental substance abuse and involvement with the child welfare system has been established (e.g., English, Marshall, Brummel, & Cogan, 1998; U.S. Department of Health and Human Services, 1999). Substance abuse may lead to child maltreatment through several mechanisms, such as child neglect that occurs when substanceabusing parents give greater priority to their drug use than to caring for their children, or substance abuse can lead to extreme poverty and inability to provide for a child’s basic needs (Magura & Laudet, 1996). Policy makers have long been concerned about the safety of children of substance-abusing parents. Drawing on a nationally representative sample from the National Survey of Child and Adolescent Well-Being (NSCAW), Guo et al. (2006) used a propensity score matching approach to address whether children whose caregivers received substance abuse services were more likely to have rereports of maltreatment than were children whose caregivers did not use substance abuse services. Using the same NSCAW data, Guo et al. employed propensity score analysis with nonparametric regression to examine the relationship between the participation of a caregiver in substance abuse services and subsequent child outcomes; that is, they investigated whether children of caregivers who used substance abuse services exhibited more behavioral problems than did children of caregivers who did not use such services. Example 5: Estimating the Impact of Multisystemic Therapy (MST). MST is a multifaceted, short-term (4–6 months), home- and community-based intervention for families with youths who have severe psychosocial and behavioral problems. Funding for MST in the United States rose from US\$5 million in 1995, to approximately US\$18 million in 2000, and to US\$35 million in 2003. Most evaluations of the program used a randomized experiment approach, and most studies generally supported the efficacy of MST. However, a recent study using a systematic review approach (J. H. Littell, 2005) found different results. Among the problems observed in previous studies, two major concerns arose: (1) the variation in the implementation of MST and (2) the integrity of conducting randomized experiments. From a program evaluation perspective, this latter concern is a common problem in social behavioral evaluations: Randomization is often broken or compromised. Statistical approaches, such as propensity score matching, may be helpful when randomization fails or is impossible (Barth et al., 2007). 43 Example 6: Assessing Program Outcomes in Group-Randomized Trials. The Social and Character Development (SACD) program was jointly sponsored by the U.S. Department of Education (DOE) and the Centers for Disease Control and Prevention. The SACD intervention project was designed to assess the impact of schoolwide social and character development education in elementary schools. Using a scientific peer review process, seven proposals to implement SACD were chosen by the Institute of Education Sciences in the U.S. DOE, and the research groups associated with each of the seven proposals implemented different SACD programs in primary schools across the country. At each of the seven sites, schools were randomly assigned to receive either the intervention program or control curricula, and one cohort of students was followed from third grade (beginning in fall 2004) through fifth grade (ending in spring 2007). A total of 84 elementary schools were randomized to intervention and control at seven sites: Illinois (Chicago), New Jersey (Buffalo, New York City, and Rochester), North Carolina, and Tennessee. Evaluating programs generated by a group randomization design is often challenging, because the unit of analysis is a cluster—such as a school—and sample sizes are so small as to compromise randomization. At one of the seven sites, the investigators of SACD designed the Competency Support Program to use a group randomization design. The total number of schools participating in the study within a school district was determined in advance, and then schools were randomly assigned to treatment conditions within school districts; for each treated school, a school that best matched the treated school on academic yearly progress, percentage of minority students, and percentage of students receiving free or reduced-price lunch was selected as a control school (i.e., data collection only without receiving intervention). In North Carolina, over a 2-year period, this group randomization procedure resulted in a total of 14 schools (Cohort 1, 10 schools; Cohort 2, 4 schools) for the study; 7 received the Competency Support Program intervention, and 7 received routine curriculum.

As it turned out—as is often the case when implementing randomized experiments in social behavioral sciences—the group randomization did not work out as planned. In some school districts, as few as four schools met the study criteria and were eligible for participation. Just by the luck of the draw (i.e., by random assignment), the two intervention schools differed systematically on covariates from the two control schools. Thus, when comparing data from the 10 schools, the investigators found the intervention schools differed from the control schools in significant ways: The intervention schools had lower academic achievement scores on statewide tests (Adequate Yearly Progress [AYP]), a higher percentage of students of color, a higher percentage of students receiving free or reduced-price lunches, and lower mean scores on behavioral composite scales at baseline. These differences were statistically significant at the .05 level using bivariate tests and logistic regression models. The researchers were confronted with the failure of 44 randomization. Were these selection effects ignored, the evaluation findings would be biased. It is just at this intersection of design (i.e., failure of randomization) and data analysis that propensity score approaches become very helpful. The preceding examples illustrate conditions under which researchers might consider propensity score modeling. The need for conducting propensity score analysis can also be determined by an imbalance check. This bivariate analysis of the equivalence of covariates by treatment condition may be viewed as an initial check of group or condition comparability. Balance checks help researchers discern whether data correction approaches more sophisticated than covariance control or regression modeling may be warranted. Because of its centrality in making analytic decisions, we present details of the imbalance check here. As noted earlier, researchers are often concerned with the validity of inferences from observational studies, because, in such a setting, the data are generated by a nonrandom process; thus, to determine whether a study requires a correction other than simple covariance control, an initial test using a normalized difference score ΔX may be undertaken (Imbens & Wooldridge, 2009). The test is basically a bivariate analysis using the treatment indicator variable and each covariate X , and X can be either a continuous or dichotomous variable. ΔX is defined as follows: where and are the sample mean values of X , and and are the sample variances of X , for the treatment group and comparison group, respectively. Examples of applying Equation 1.1 to check normalized differences are shown in Section 6.5.2. Following Imbens and Wooldridge, a ΔX exceeding .25 is an indication that selection bias exists and linear regression methods tend to be sensitive to the model specification. In other words, if an initial check of study data shows ΔX exceeding .25 for numerous covariates, researchers should consider employing corrective approaches other than regression, or at least perform corrective analysis in conjunction with regression analysis. As a family of corrective approaches, propensity score models have promising properties and offer several advantages under the condition of imbalance. 1.5 COMPUTING SOFTWARE PACKAGES At the time of the first edition of this book, few software packages offered comprehensive procedures to handle the statistical analyses described in subsequent chapters. Recently, however, more software programs have been developed for propensity score analysis. Our review of software packages 45 indicates that Stata (StataCorp, 2007) and R (R Foundation for Statistical Computing, 2008) offer the most comprehensive computational facilities. Other packages, such as SAS, offer user-developed macros or procedures targeting specific problems (e.g., SAS Proc Assign may be used to implement optimal matching), but they do not offer the variety of analysis options that is available in Stata and R. Table 1.1 lists the Stata and R procedures available for implementing the analyses described in this book. Just like the rapid growth of propensity score methods per se, computing programs have also developed at a fast pace. Table 1.1 shows some of the programs currently available. See also Stuart (2014), who provides a comprehensive list of software programs for implementing matching methods and propensity scores in R, Stata, SAS, and SPSS. We have chosen to use Stata to illustrate most approaches. We chose Stata based on our experience with the software and our conclusion that it is a convenient software package. Specifically, Stata's program test_condate can be used to test treatment effect heterogeneity described in Chapter 2; heckman and treatreg can be used to solve problems described in Chapter 4; psmatch2, pscore, boost, imbalance, hodgesl, logistic, xtlogit, xtlogit, and xtmixed can be used to solve problems described in Chapter 5; pscore, hte, and Stata programming commands can be used to solve problems described in Chapter 6; pweight function specified in a multivariate model can be used to solve problems described in Chapter 7; rmatch can be used to solve problems described in Chapter 8; psmatch2 can be used to solve problems described in Chapter 9; pweight function specified in a multivariate model and gpscore can be used to solve problems described in Chapter 10; and rbounds and mhbounds can be used to perform Rosenbaum's (2002b) sensitivity analysis described in Chapter 11. In each of these chapters, we will provide examples and an overview of Stata syntax. We provide illustrative examples for one R procedure (i.e., optmatch), because this is the only procedure available for conducting optimal matching within R and Stata. All syntax files and illustrative data can be downloaded from this book's companion website (). Many of the Stata programs described above were macros or ado files developed by users. At the time this second edition was completed, Stata released its version 13 (StataCorp, 2013). This version of Stata for the first time includes a series of programs facilitating statistical analyses using propensity scores and other methods. Under the title of "Treatment effects," this group of programs includes regression adjustment, inverse-probability weights (IPW), doubly robust estimators, matching estimators, overlap plots, and endogenous treatment estimators. Many of these newly released programs offer functions similar to those described in this book, although the userdeveloped programs will continue to be needed for many analyses. 46 1.6 PLAN OF THE BOOK Chapter 2 offers a conceptual framework for the development of scientific approaches to causal analysis, namely, the Neyman-Rubin counterfactual framework. In addition, the chapter reviews a closely related, and recently developed, framework that aims to guide scientific inquiry of causal inferences: the econometric model of causality (Heckman, 2005). The chapter includes a discussion of two fundamental assumptions embedded in nearly all outcome-oriented program evaluations: the ignorable treatment assignment assumption and the stable unit treatment value assumption (SUTVA). Violations of these assumptions pose challenges to the estimation of counterfactuals. The chapter provides a review of corrective methods other than propensity score analysis, particularly two methods that are widely employed in economic research, namely, the instrumental variables estimator and regression discontinuity design. The chapter offers a discussion on the importance of modeling treatment effect heterogeneity, two tests of effect heterogeneity, and an example to show the application of these tests. Chapter 3 focuses on the issue of ignorable treatment assignment from the other side of the coin: strategies for data balancing when treatment effects can only be assessed in a nonexperimental design. This chapter aims to answer the key question of what kind of statistical methods should be considered to remedy the estimation of counterfactuals, when treatment assignment is not ignorable. Moreover, the chapter describes three closely related but methodologically distinctive approaches: ordinary least squares (OLS) regression, matching, and stratification. The discussion includes a comparison of estimated treatment effects of the three methods under five scenarios. These methods involve making simple corrections when assignment is not ignorable, and they serve as a starting point for discussing the data issues and features of more sophisticated approaches, such as the seven advanced models described later in the book. The chapter serves as a review of preliminary concepts that are a necessary foundation for learning more advanced approaches. Chapters 4 through 10 present statistical theories using examples to illustrate each of the seven advanced models covered in this book. Chapter 4 describes and illustrates Heckman's sample selection model in its original version (i.e., the model aims to correct for sample selection) and the revised Heckman model developed to evaluate treatment effects. Chapter 5 describes propensity score matching, specifically the creation of matched samples using caliper (or Mahalanobis metric) matching and recently developed methods of optimal matching, propensity score matching with multilevel modeling, estimation of propensity scores with a generalized boosted regression, and various approaches for postmatching analysis of outcomes. Although Chapter 5 focuses on matching, sections on estimating propensity scores and strategies for developing optimal models serve also as a guide for the methods described in 47 Chapters 6, 7, 9, and 10. In these chapters, virtually the same approaches are used to estimate propensity scores. Chapter 6 focuses on propensity score subclassification, a method that can be applied to outcome variables that are not normally distributed and special types of models such as structural equation modeling. Chapter 7 describes propensity score weighting, a robust approach that can also be applied to various types of outcome variables, such as time-to-event data, and complex outcome analyses using structural equation modeling. Chapter 8 describes a collection of matching estimators developed by Abadie and Imbens (2002), who provide an extension of Mahalanobis metric matching. Among the attractive features of this procedure is its provision of standard errors for various treatment effects. Chapter 9 describes propensity score analysis with nonparametric regression. Specifically, it describes the two-limpeoid difference-in-differences approach developed by Heckman and his colleagues (Heckman, Ichimura, & Todd, 1997, 1998). Chapter 10 describes methods to model doses of treatment. This chapter extends the basic methods for binary treatment conditions (treated and control) to more complex situations in which a treatment variable has more than two conditions and can be either categorical or continuous. Table 1.1 Stata and R Procedures by Analytic Methods 48 49 Chapter 11 reviews selection bias, which is the core problem all statistical methods described in this book aim to resolve. This chapter gives the selection bias problem a more rigorous treatment: We simulate two settings of data generation (i.e., selection on observables and selection on unobservables) and compare the performance of six models under these settings using Monte Carlo studies. Hidden selection bias is a problem that fundamentally distinguishes observational studies from randomized experiments. When key variables are missing, researchers inevitably stand on thin ice when drawing inferences about causal effects in observational studies. However, Rosenbaum's (2002b) sensitivity analysis, which is illustrated in Chapter 11, is a useful tool for testing 50 the sensitivity of study findings to hidden selection. This chapter reviews assumptions for all seven models and demonstrates practical strategies for model comparison. Finally, Chapter 12 focuses on continuing issues and challenges in the field. It reviews debates on whether propensity score analysis can be employed as a replacement for randomized experiments. It comments on recent advances. And it suggests directions for the development of new approaches to observational studies. NOTES 1. Excel can be used to calculate the number of combinations of 8 things taken 4 at a time by typing the following in a cell: =COMBIN(8,4), and Excel returns the number 70. 2. If the tea taster gives an outcome other than 11110000, then the number of assignments having six exact agreements remains 17. However, there will be a different set of 17 assignments than those presented here. 3. Uniform here refers to equal probability for elements in the study population to receive treatment. 51 CHAPTER 2 Counterfactual Framework and Assumptions This chapter examines conceptual frameworks that guide the estimation of treatment effects as well as important assumptions that are embedded in observational studies. Section 2.1 defines causality and describes threats to internal validity. In addition, it reviews concepts that are generally discussed in the evaluation literature, emphasizing their links to statistical analysis. Section 2.2 summarizes the key features of the Neyman-Rubin counterfactual framework. Section 2.3 discusses the ignorable treatment assignment assumption. Section 2.4 describes the stable unit treatment value assumption (SUTVA). Section 2.5 provides an overview of statistical approaches developed to handle selection bias. With the aim of showing the larger context in which new evaluation methods are developed, this focuses on a variety of models, including the seven models covered in this book, and two popular approaches widely employed in economics—the instrumental variables estimator and regression discontinuity designs. Section 2.6 reviews the underlying logic of statistical inference for both randomized experiments and observational studies. Section 2.7 summarizes a range of treatment effects and extends the discussion of the SUTVA. We examine treatment effects by underscoring the maxim that different research questions imply different treatment effects and different analytic models must be matched to the kinds of effects expected. Section 2.8 discusses treatment effect heterogeneity and two recently developed tests of effect heterogeneity. With illustrations, this section shows how to use the tests to evaluate effect heterogeneity and the plausibility of the strongly ignorable treatment assignment assumption. Section 2.9 reviews Heckman's scientific model of causality, which is a comprehensive, causal inference framework developed by econometricians. Section 2.10 concludes the chapter with a summary of key points. 2.1 CAUSALITY, INTERNAL VALIDITY, AND THREATS 52 Program evaluation is essentially the study of cause-and-effect relationships. It aims to answer this key question: To what extent can the net difference observed in outcomes between treated and nontreated groups be attributed to the intervention, given that all other things are held constant (or ceteris paribus)? Causality in this context simply refers to the net gain or loss observed in the outcome of the treatment group that can be attributed to malleable variables in the intervention. Treatment in this setting ranges from receipt of a well-specified program to falling into a general state such as "being a service recipient," as long as such a state can be defined as a result of manipulations of the intervention (e.g., a mother of young children who receives cash assistance under the Temporary Assistance to Needy Families program [TANF]). Rubin (1986) argued that there can be no causation without manipulation. According to Rubin, thinking about actual manipulations forces an initial definition of units and treatments, which is essential in determining whether a program truly produces an observed outcome. Students from any social or health sciences discipline may have learned from their earliest research course that association should not be interpreted as the equivalent of causation. The fact that two variables, such as A and B, are highly correlated does not necessarily mean that one is a cause and the other is an effect. The existence of a high correlation between A and B may be the result of the following conditions: (1) Both A and B are determined by a third variable, C, and by controlling for C, the high correlation between A and B disappears. If that's the case, we say that the correlation is spurious. (2) A causes B. In this case, even though we control for another set of variables, we still observe a high association between A and B. (3) In addition, it is possible that B causes A, in which case the correlation itself does not inform us about the direction of causality. A widely accepted definition of causation was given by Lazarsfeld (1959), who described three criteria for a causal relationship. (1) A causal relationship between two variables must have temporal order, in which the cause must precede the effect in time (i.e., if A is a cause and B an effect, then A must occur before B). (2) The two variables should be empirically correlated with one another. And (3), most important, the observed empirical correlation between two variables cannot be explained away as the result of a third variable that causes both A and B. In other words, the relationship is not spurious and occurs with regularity. According to Pearl (2000), the notion that regularity of succession or correlation is not sufficient for causation dates back to the 18th century, when Hume (1748/1959) argued, "We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by an object similar to the second. Or, in other words, where, if the first object had not been, the 53 second never had existed. (sec. vii) On the basis of the three criteria for causation, Campbell (1957) and his colleagues developed the concept of internal validity, which serves a paramount role in program evaluation. Conceptually, internal validity shares common features with causation. We use the term internal validity to refer to inferences about whether observed covariation between A and B reflects a causal relationship from A to B in the form in which the variables were manipulated or measured. To support such an inference, the researcher must show that A preceded B in time, that A covaries with B . . . and that no other explanations for the relationship are plausible. (Shadish et al., 2002, p. 53) In program evaluation and observational studies in general, researchers are concerned about threats to internal validity. These threats are factors affecting outcomes other than intervention or the focal stimuli. In other words, threats to internal validity are other possible reasons to think that the relationship between A and B is not causal, that the relationship could have occurred in the absence of the treatment, and that the relationship between A and B could have led to the same outcomes that were observed for the treatment. Nine well-known threats to internal validity are ambiguous temporal precedence, selection, history, maturation, regression, attrition, testing, instrumentation, and additive and interactive effects of threats to internal validity (Shadish et al., 2002, pp. 54–55). It is noteworthy that many of these threats have been carefully examined in the statistical literature, although statisticians and econometricians have used different terms to describe them. For instance, Heckman, LaLonde, and Smith (1999) referred to the testing threat as the Hawthorne effect, meaning that an agent's behavior is affected by the act of participating in an experiment. Rosenbaum (2002b) distinguished between two types of bias that are frequently found in observational studies: overt bias and hidden bias. Overt bias can be seen in the data at hand, whereas the hidden bias cannot be seen because the required information was not observed or recorded. Although different in their potential for detection, both types of bias are induced by the fact that "the treated and control groups differ prior to treatment in ways that matter for the outcomes under study" (Rosenbaum, 2002b, p. 71). Suffice it to say that Rosenbaum's "ways that matter for the outcomes under study" encompass one or more of the nine threats to internal validity. This book adopts a convention of the field that defines selection threat broadly. That is, when we refer to selection bias, we mean a process that involves one or more of the nine threats listed earlier and not necessarily the more limited definition of selection threat alone. In this sense, then, selection 54 bias may take one or more of the following forms: self-selection, bureaucratic selection, geographic selection, attrition selection, instrument selection, or measurement selection. 2.2 COUNTERFACTUALS AND THE NEYMANRUBIN COUNTERFACTUAL FRAMEWORK Having defined causality, we now present a key conceptual framework developed to investigate causality: the counterfactual framework. Counterfactuals are at the heart of any scientific inquiry. Galileo was perhaps the first scientist who used the thought experiment and the idealized method of controlled variation to define causal effects (Heckman, 1996). In philosophy, the practice of inquiring about causality through counterfactuals stems from early Greek philosophers such as Aristotle (384–322 BCE; Holland, 1986) and Chinese philosophers such as Zhou Zhuang (369–286 BCE; see Guo, 2012). Hume (1748/1959) also was discontent with the regularity of the factual account and thought that the counterfactual criterion was less problematic and more illuminating. According to Pearl (2000), Hume's idea of basing causality on counterfactuals was adopted by John Stuart Mill (1843), and it was embellished in the works of David Lewis (1973, 1986). Lewis (1986) called for abandoning the regularity account altogether and for interpreting "A has caused B" as "B would not have occurred if it were not for A." In statistics, researchers generally credit the development of the counterfactual framework to Neyman (1923) and Rubin (1974, 1978, 1980b, 1986) and call it the Neyman-Rubin counterfactual framework of causality. The terms Rubin causal model and potential outcomes model are also used interchangeably to refer to the same model. Other scholars who made independent contributions to the development of this framework come from a variety of disciplines, including Fisher (1935/1971) and Cox (1958) from statistics, Thurstone (1930) from psychometrics, and Haavelmo (1943), Roy (1951), and Quandt (1958, 1972) from economics. Holland (1986), Sobel (1996), Winship and Morgan (1999), and Morgan and Winship (2007) have provided detailed reviews of the history and development of the counterfactual framework. So what is a counterfactual? A counterfactual is a potential outcome, or the state of affairs that would have happened in the absence of the cause (Shadish et al., 2002). Thus, for a participant in the treatment condition, a counterfactual is the potential outcome under the condition of control; for a participant in the control condition, a counterfactual is the potential outcome under the condition of treatment. Note that the definition uses the subjunctive mood (i.e., contingent on what "would have happened . . ."), which means that the counterfactual is not observed in real data. Indeed, it is a missing value. Therefore, the fundamental 55 task of any evaluation is to use known information to impute a missing value for a hypothetical and unobserved outcome. Neyman-Rubin's framework emphasizes that individuals selected into either treatment or nontreatment groups have potential outcomes in both states: that is, the one in which they are observed and the one in which they are not observed. More formally, assume that each person i under evaluation would have two potential outcomes (Y_{0i} , Y_{1i}) that correspond, respectively, to the potential outcomes in the untreated and treated states. Let $W_i = 1$ denote the receipt of treatment, $W_i = 0$ denote nonreceipt, and Y_i indicate the measured outcome variable. The Neyman-Rubin counterfactual framework can then be expressed as the following model: In the preceding equation, W_i is a dichotomous variable; therefore, both the terms W_i and $(1 - W_i)$ serve as a switcher. Basically, the equation indicates which of the two outcomes would be observed in the real data, depending on the treatment condition or the "on/off" status of the switch. The key message conveyed in this equation is that to infer a causal relationship between W_i (the cause) and Y_i (the outcome), the analyst cannot directly link Y_{1i} to W_i under the condition $W_i = 1$; instead, the analyst must check the outcome of Y_{0i} under the condition of $W_i = 0$ and compare Y_{0i} with Y_{1i} . For example, we might hypothesize that a child i who comes from a low-income family has low academic achievement. Here, the treatment variable is $W_i = 1$ if the child lives in poverty; the academic achievement $Y_{1i} < p$ if the child has a low academic achievement, where p is a cutoff value defining a low test score and $Y_{1i} > p$ otherwise. To make a causal statement that being poor ($W_i = 1$) causes low academic achievement $Y_{1i} < p$, the researcher must examine the outcome under the status of not being poor. That is, the task is to determine the child's academic outcome Y_{0i} under the condition of $W_i = 0$, and ask, "What would have happened had the child not lived in a poor family?" If the answer to the question is $Y_{0i} > p$, then the researcher can have confidence that $W_i = 1$ causes $Y_{1i} < p$. The above argument gives rise to many issues that we will examine in detail. The most critical issue is that Y_{0i} is not observed. Holland (1986, p. 947) called this issue the fundamental problem of causal inference. How could a researcher possibly know $Y_{0i} > p$? The Neyman-Rubin counterfactual framework holds that a researcher can estimate the counterfactual by examining the average outcome of the treatment participants and the average outcome of the nontreatment participants in the population. That is, the researcher can assess the counterfactual by evaluating the difference in mean outcomes between the 56 two groups or "averaging out" the outcome values of all individuals in the same condition. Specifically, let $E(Y_{0W} = 0)$ denote the mean outcome of the individuals who compose the nontreatment group, and $E(Y_{1W} = 1)$ denote the mean outcome of the individuals who comprise the treatment group. Because both outcomes in the above formulation (i.e., $E(Y_{0W} = 0)$ and $E(Y_{1W} = 1)$) are observable, we can then define the treatment effect as a mean difference: where τ denotes treatment effect. This formula is called the standard estimator for the average treatment effect. It is worth noting that under this framework, the evaluation of $E(Y_{1W} = 1) - E(Y_{0W} = 0)$ can be understood as an effort that uses $E(Y_{0W} = 0)$ to estimate the counterfactual $E(Y_{0W} = 1)$. The central interest of the evaluation is not in $E(Y_{0W} = 0)$ but in $E(Y_{0W} = 1)$. Returning to our example with the hypothetical child, the solution to the dilemma of not observing the academic achievement for child i in the condition of not being poor is resolved by examining the average academic achievement for all poor children in addition to the average academic achievement of all nonpoor children in a well-defined population. If the comparison of two mean outcomes leads to $\tau = E(Y_{1W} = 1) - E(Y_{0W} = 0) < 0$, or the mean outcome of all poor children is a low academic achievement, then the researcher can infer that poverty causes low academic achievement and also can provide support for hypotheses advanced under resources theories (e.g., Wolock & Horowitz, 1981). In summary, the Neyman-Rubin framework offers a practical way to evaluate the counterfactuals. Working with data from a sample that represents the population of interest (i.e., using Y_{1i} and Y_{0i} as sample variables denoting, respectively, the population variables Y_1 and Y_0 , and w as a sample variable denoting W), we can further define the standard estimator for the average treatment effect as the difference between two estimated means from sample data: The Neyman-Rubin counterfactual framework provides a useful tool not only for the development of various approaches to estimating potential outcomes but also for a discussion of whether assumptions embedded in randomized experiments are plausible when applied to social and health sciences studies. In this regard, at least eight issues emerge. 1. In the preceding exposition, we expressed the evaluation of causal effects in an overly simplified fashion that did not take into consideration any covariates or threats to internal validity. In our hypothetical example where 57 poor economic condition causes low academic achievement, many confounding factors might influence achievement. For instance, parental education could covary with income status, and it could affect academic achievement. When covariates are entered into an equation, evaluators must impose additional assumptions. These include the ignorable treatment assignment assumption and the SUTVA, which we clarify in the next two sections. Without assumptions, the counterfactual framework leads us nowhere. Indeed, it is violations of these assumptions that have motivated statisticians and econometricians to develop new approaches. 2. In the standard estimator $E(Y_{1W} = 1) - E(Y_{0W} = 0)$, the primary interest of researchers is focused on the average outcome of treatment participants if they had not participated (i.e., $E(Y_{0W} = 1)$). Because this term is unobservable, evaluators use $E(Y_{0W} = 0)$ as a proxy. It is important to understand when the standard estimator consistently estimates the true average treatment effect for the population. Winship and Morgan (1999) decomposed the average treatment effect in the population into a weighted average of the average treatment effect for those in the treatment group and the average treatment effect for those in the control group as2 where π is equal to the proportion of the population that would be assigned to the treatment group, and by the definition of the counterfactual model, let $E(Y_{1W} = 0)$ and $E(Y_{0W} = 1)$ be defined analogously to $E(Y_{1W} = 1)$ and $E(Y_{0W} = 0)$. The quantities $E(Y_{1W} = 0)$ and $E(Y_{0W} = 1)$ that appear in the second and third lines of Equation 2.4 cannot be directly calculated because they are unobservable values of Y . Furthermore, and again on the basis of the definition of the counterfactual model, if we assume that $E(Y_{1W} = 1) = E(Y_{1W} = 0)$ and $E(Y_{0W} = 0) = E(Y_{0W} = 1)$, then through substitution starting in the fourth line of Equation 2.4, we have3 58 Thus, a sufficient condition for the standard estimator to consistently estimate the true average treatment effect in the population is that $E(Y_{1W} = 1) = E(Y_{1W} = 0)$ and $E(Y_{0W} = 0) = E(Y_{0W} = 1)$. This condition, as shown by numerous statisticians such as Fisher (1925), Kempthorne (1952), and Cox (1958), is met in the classical randomized experiment.4 Randomization works in a way that makes the assumption about $E(Y_{1W} = 1) = E(Y_{1W} = 0)$ and $E(Y_{0W} = 0) = E(Y_{0W} = 1)$ plausible. When study participants are randomly assigned either to the treatment condition or to the nontreatment condition, certain physical randomization processes are carried out so that the determination of the condition to which participant i is exposed is regarded as statistically independent of all other variables, including the outcomes Y_1 and Y_0 . 3. The real debate regarding observational studies in statistics centers on the validity of extending the randomization assumption (i.e., that a process yields results independent of all other variables) to analyses in social and health sciences evaluations. Or, to put it differently, whether the researcher engaged in evaluations can continue to assume that $E(Y_{0W} = 0) = E(Y_{0W} = 1)$ and $E(Y_{1W} = 1) = E(Y_{1W} = 0)$. Not surprisingly, supporters of randomization as the central method for evaluating social and health programs answer "yes," whereas proponents of the nonexperimental approach answer "no" to this question. The classical experimental approach assumes no selection bias, and therefore, $E(Y_{0W} = 1) = E(Y_{0W} = 0)$. The assumption of no selection bias is indeed true because of the mechanism and logic behind randomization. However, many authors challenge the plausibility of this assumption in evaluations. Heckman and Smith (1995) showed that the average outcome for the treated group under the condition of nontreatment is not the same as the average outcome of the nontreated group, precisely $E(Y_{0W} = 1) \neq E(Y_{0W} = 0)$, because of selection bias. 4. Rubin extended the counterfactual framework to a more general case—that is, allowing the framework to be applicable to observational studies. Unlike a randomized experiment, an observational study involves complicated situations that require a more rigorous approach to data analysis. Less rigorous approaches are open to criticism; for instance, Sobel (1996) criticized the common practice in sociology that uses a dummy variable (i.e., treatment vs. nontreatment) to evaluate the treatment effect in a regression model (or a regression-type model such as a path analysis or structural equation model) using survey data. As shown in the next section, the primary problem of such an approach is that the dummy treatment variable is specified by these models as exogenous, but in fact it is not. According to Sobel (2005), 59 The incorporation of Neyman's notation into the modern literature on causal inference is due to Rubin (1974, 1977, 1978, 1980b), who, using this notation, saw the applicability of the work from the statistical literature on experimental design to observational studies and gave explicit consideration to the key role of the treatment assignment mechanism in causal inference, thereby extending this work to observational studies. To be sure, previous workers in statistics and economics (and elsewhere) understood well in a less formal way the problems of making causal inferences in observational studies where respondents selected themselves into treatment groups, as evidenced, for example, by Cochran's work on matching and Heckman's work on sample selection bias. But Rubin's work was a critical breakthrough. (p. 100) 5. In the above exposition, we used the most common and convenient statistic (i.e., the mean) to express various counterfactuals and the ways in which counterfactuals are approximated. The average causal effect is an average, and as such, according to Holland (1986, p. 949), "enjoys all of the advantages and disadvantages of averages." One such disadvantage is the insensitivity of an average to the variability of the causal effect. If the variability in individual causal effects ($Y_{1i}W_i = 1) - (Y_{1i}W_i = 0)$ is large over all units, then $\tau = E(Y_{1W} = 1) - E(Y_{0W} = 0)$ may not well represent the causal effect of a specific unit (say, u_0). If u_0 is the unit of interest, then τ may be irrelevant, no matter how carefully we estimate it!" (Holland, 1986, p. 949). This important point is expanded in Sections 2.7 and 2.8, but we want to emphasize that the variability of the treatment effect at the individual level, or violation of an assumption about a constant treatment effect across individuals, can make the estimation of average treatment effects biased; therefore, it is important to distinguish among various types of treatment effects. In short, different statistical approaches employ counterfactuals of different groups to estimate different types of treatment effects. 6. Another limitation of using an average lies in the statistical properties of means. Although means are conventional, distributions of treatment parameters are also of considerable interest (Heckman, 2005, p. 20). In several articles, Heckman and his colleagues (Heckman, 2005; Heckman, Ichimura, & Todd, 1997; Heckman et al., 1999; Heckman, Smith, & Clements, 1997) have discussed the limitation of reliance on means (e.g., disruption bias leading to changed outcomes or the Hawthorne effect) and have suggested using other summary measures of the distribution of counterfactuals such as (a) the proportion of participants in Program A who benefit from the program relative to some alternative B, (b) the proportion of the total population that benefits from Program B compared with Program A, (c) selected quantiles of the impact 60 distribution, and (d) the distribution of gains at selected base state values. 7. The Neyman-Rubin framework expressed in Equation 2.1 is the basic model. However, there are variants that can accommodate more complicated situations. For instance, Rosenbaum (2002b) developed a counterfactual model in which stratification is present and where s stands for the s th stratum: Under this formulation, Equation 2.1 is the simplest case where s equals 1, or stratification is absent 5.8 The Neyman-Rubin counterfactual framework is mainly a useful tool for the statistical exploration of causal effects. However, by no means does this framework exclude the importance of using substantive theories to guide causal inferences. Identifying an appropriate set of covariates and choosing an appropriate model for data analysis are primarily tasks of developing theories based on prior studies in the substantive area. As Cochran (1965) argued, when summarizing the results of a study that shows an association consistent with a causal hypothesis, the investigator should always list and discuss all alternative explanations of his results (including different hypotheses and biases in the results) that occur to him. (sec. 9.5) Dating from Fisher's work, statisticians have long acknowledged the importance of having a good theory of the treatment assignment mechanism (Sobel, 2005). Rosenbaum (2005) emphasized the importance of using theory in observational studies and encouraged evaluators to "be specific" on which variables to match and which variables to control using substantive theories. Thus, similar to all scientific inquiries, the counterfactual framework is reliable only under the guidance of appropriate theories and substantive knowledge. 2.3 THE IGNORABLE TREATMENT ASSIGNMENT ASSUMPTION By thinking of the central challenge of all evaluations as estimating the missing outcomes for participants—each of whom is missing an observed outcome for either the treatment or nontreatment condition—the evaluation problem becomes a missing data issue. Consider the standard estimator of the average treatment effect: $\tau = E(Y_{1W} = 1) - E(Y_{0W} = 0)$. Many sources of error contribute to the bias of τ . It is for this reason that the researcher has to make a few fundamental assumptions to apply the Neyman-Rubin counterfactual model to actual evaluations. One such assumption is the ignorable treatment assignment 61 assumption (Rosenbaum & Rubin, 1983). In the literature, this assumption is sometimes presented as part of the SUTVA (e.g., Rubin, 1986); however, we treat it as a separate assumption because of its importance. The ignorable treatment assignment is fundamental to the evaluation of treatment effects, particularly in the econometric literature. Our discussion follows this tradition. The assumption can be expressed as The assumption says that conditional on covariates X , the assignment of study participants to binary treatment conditions (i.e., treatment vs. nontreatment) is independent of the outcome of nontreatment (Y_0) and the outcome of treatment (Y_1). A variety of terms have emerged to describe this assumption: unconfoundedness (Rosenbaum & Rubin, 1983), selection on observables (Barnow, Cain, & Goldberger, 1980), conditional independence (Lechner, 1999), and exogeneity (Imbens, 2004). These terms can be used interchangeably to denote the key idea that assignment to one condition or another is independent of the potential outcomes if observable covariates are held constant. The researcher conducting a randomized experiment can be reasonably confident that the ignorable treatment assignment assumption holds because randomization typically balances the data between the treated and control groups and makes the treatment assignment independent of the outcomes under the two conditions (Rosenbaum, 2002b; Rosenbaum & Rubin, 1983). However, the ignorable treatment assignment assumption is often violated in quasiexperimental designs and in observational studies because the creation of a comparison group follows a natural process that confounds group assignment with outcomes. Thus, the researcher's first task in any evaluation is to check the tenability of the independence between the treatment assignment and outcomes under different conditions. A widely employed approach to this problem is to conduct bivariate analysis using the dichotomous treatment variable (W) as one and each independent variable available to the analyst (i.e., each variable in the matrix X , one at a time) as another. Chi-square tests may be applied to the case w here X is a categorical variable, and an independent samples t test or Wilcoxon rank sum (Mann-Whitney) test may be applied to the case where X is a continuous variable. Whenever the null hypothesis is rejected as showing the existence of a significant difference between the treated and nontreated groups on the variable under examination, the researcher may conclude that there is a correlation between treatment assignment and outcome that is conditional on an observed covariate; therefore, the treatment assignment is not ignorable, and taking remedial measures to correct the violation is warranted. Although this method is popular, it is worth noting that Rosenbaum (2002b) cautioned that no statistical evidence exists that supports the validity of this convention, because 62 this assumption is basically untestable. To demonstrate that the ignorable treatment assignment is nothing more than the same assumption of ordinary least squares (OLS) regression about the independence of the error term from an independent variable, we present evidence of the associative relation between the two assumptions. In the OLS context, the assumption is also known as contemporaneous independence of the error term from the independent variable or, more generally, exogeneity. To analyze observational data, an OLS regression model using a dichotomous indicator may not be the best choice. To understand this problem, consider the where W_i is a dichotomous following OLS regression model: variable indicating treatment, and X_i is the vector of independent variables for case i . In observational data, because researchers have no control over the assignment of treatment conditions, W is often highly correlated with Y . The use of statistical controls—a common technique in the social and health sciences— involves a modeling process that attempts to extract the independent contribution of explanatory variables (i.e., the vector X) to the outcome Y to determine the net effect of τ . When the ignorable treatment assignment assumption is violated and the correlation between W and e is not equal to 0, the OLS estimator of treatment effect τ is biased and inconsistent. More formally, under this condition, there are three problems associated with the OLS estimator. First, when the treatment assignment is not ignorable, the use of the dummy variable W leads to endogeneity bias. In the above regression equation, the dummy variable W is conceptualized as an exogenous variable. In fact, it is a dummy endogenous variable. The nonignorable treatment assignment implies a mechanism of selection; that is, there are other factors determining W . W is merely an observed variable that is determined by a latent variable W^* in such a way that $W = 1$, if $W^* > C$, and $W = 0$, otherwise, where C is a constant reflecting a cutoff value of utility function. Factors determining W^* should be explicitly taken into consideration in the modeling process. Conceptualizing W as a dummy endogenous variable motivated Heckman

(1978, 1979) to develop the sample selection model and Maddala (1983) to develop the treatment effect model. Both models attempt to correct for the endogeneity bias. See Chapter 4 for a discussion of these models. Second, the presence of the endogeneity problem (i.e., the independent variable is not exogenous and is correlated with the error term of the regression) leads to a biased and inconsistent estimation of the regression coefficient. Our demonstration of the adverse consequence follows Berk (2004). For ease of exposition, assume all variables are mean centered, and there is one predictor in the model: 63 The least squares estimate of its Substituting Equation 2.8 into Equation 2.9 and simplifying, the result is $E(y) = \beta_0 + \beta_1 x + \beta_2 e$, where $\beta_2 = \beta_1 \rho$, and ρ is the correlation between x and e . If x and e are correlated, the expected value for the far right-hand term will be nonzero, and the numerator will not go to zero as the sample size increases without limit. The least squares estimate then will be biased and inconsistent. The presence of a nonzero correlation between x and e may be due to one or more of the following reasons: (a) the result of random measurement error in x , (b) one or more omitted variables correlated with x and y , (c) the incorrect functional form, and (d) a number of other problems (Berk, 2004, used with permission). This problem is also known as asymptotic bias, which is a term that is analogous to inconsistency. Kennedy (2003) explained that when contemporaneous correlation is present, "the OLS procedure, in assigning 'credit' to regressors for explaining variation in the dependent variable, assigns, in error, some of the regressors with which that disturbance is contemporaneously correlated" (p. 158). Suppose that the correlation between the independent variable and the error term is positive. When the error is higher, the dependent variable is also higher, and owing to the correlation between the error and the independent variable, the independent variable is likely to be higher, which implies that too much credit for making the dependent variable higher is likely to be assigned to the independent variable. Figure 2.1 illustrates this scenario. If the error term and the independent variable are positively correlated, negative values of the error will tend to correspond to low values of the independent variable, and positive values of the error will tend to correspond to high values of the independent variable, which will create data patterns similar to that shown in the figure. The OLS estimating line clearly overestimates the slope of the true relationship. Obviously, the estimating line in this hypothetical example provides a much better fit to the sample data than does the true relationship, which causes the variance of the error term to be underestimated. Finally, in observational studies, because researchers have no control over the assignment of treatment conditions, W is often correlated with Y . A statistical control is a modeling process that attempts to extract the independent contribution of explanatory variables to the outcome to determine the net effect of T . Although the researcher aims to control for all important variables by using 64 a well-specified matrix X , the omission of important controls often occurs and results in a specification error. The consequence of omitting relevant variables is a biased estimation of the regression coefficient. We follow Greene (2003, pp. 148–149) to show why this is the case. Suppose that a correctly specified regression model would be where the two parts of X have K_1 and K_2 columns, respectively. If we regress y on X_1 without including X_2 , then the estimator is Figure 2.1 Positive Contemporaneous Correlation Source: Kennedy (2003), p.158. Copyright © 2003 Massachusetts Institute of Technology. Reprinted by permission of The MIT Press. Taking the expectation, we see that unless well-known result is the omitted variable formula is biased. The where Each column of the $K_1 \times K_2$ matrix $P_{1.2}$ is the column of slopes in the least squares regression of the corresponding column of X_2 on the column of X_1 . When the ignorable treatment assignment assumption is violated, remedial 65 action is needed. The popular use of statistical controls with OLS regression is a choice that involves many risks. In Section 2.5, we review alternative approaches that have been developed to correct for biases under the condition of nonignorable assignment (e.g., the Heckman sample selection model directly modeling the endogenous dummy treatment condition) and approaches that relax the fundamental assumption to focus on a special type of treatment effect (e.g., average treatment effect for the treated rather than sample average treatment effect). 2.4 THE STABLE UNIT TREATMENT VALUE ASSUMPTION T he stable unit treatment value assumption (SUTVA) was labeled and formally presented by Rubin in 1980. Rubin (1986) later extended this assumption, arguing that it plays a key role in deciding which questions are adequately formulated to have causal answers. Only under SUTVA is the representation of outcomes by the Neyman-Rubin counterfactual model adequate. Formally, consider the situation with N units indexed by $i = 1, \dots, N$; T treatments indexed by $w = 1, \dots, T$; and outcome variable Y , whose possible values are represented by $Y_{i,w}$ ($i = 1, \dots, N$; $w = 1, \dots, T$).6 SUTVA is simply the a priori assumption that the value of Y for unit i when exposed to treatment w will be the same no matter what mechanism is used to assign treatment w to unit i and no matter what treatments the other units receive, and this holds for all $i = 1, \dots, N$ and all $w = 1, \dots, T$. As it turns out, SUTVA basically imposes exclusive restrictions. Heckman (2005, p. 11) interprets these exclusive restrictions as the following two circumstances: (1) SUTVA rules out social interactions and general equilibrium effects, and (2) SUTVA rules out any effect of the assignment mechanism on potential outcomes. We previously examined the importance of the second restriction (ignorable treatment assignment) in Section 2.3. The following section explains the importance of the first restriction and describes the conditions under which the assumption is violated. According to Rubin (1986), SUTVA is violated when unrepresented versions of treatment exist (i.e., $Y_{i,w}$ depends on which version of treatment w is received) or when there is interference between units (i.e., $Y_{i,w}$ depends on whether i received treatment w or w' , where $i \neq i'$ and $w \neq w'$). The classic example of violation of SUTVA is the analysis of treatment effects in agricultural research, such as rainfall that surreptitiously carries fertilizer from a treated plot to an adjacent untreated plot. In social behavioral evaluations, SUTVA is violated when a treatment alters social or environmental conditions that, in turn, alter 66 potential outcomes. Winship and Morgan (1999) illustrated this idea by describing the impact of a large job training program on local labor markets: Consider the case where a large job training program is offered in a metropolitan area with a competitive labor market. As the supply of graduates from the program increases, the wage that employers will be willing to pay graduates of the program will decrease. When such complex effects are present, the powerful simplicity of the counterfactual framework vanishes. (p. 663) SUTVA is both an assumption that facilitates investigation or estimation of counterfactuals and a conceptual perspective that underscores the importance of analyzing differential treatment effects with appropriate estimators. We return to SUTVA as a conceptual perspective in Section 2.7. It is noteworthy that Heckman and his colleagues (Heckman et al., 1999) treated SUTVA as a strong assumption and presented evidence against the assumption. The limitations imposed by the strong assumption may be overcome by relaxed assumptions (Heckman & Vytlacil, 2005). 2.5 METHODS FOR ESTIMATING TREATMENT EFFECTS As previously discussed in Section 2.3, violating the ignorable treatment assignment assumption has adverse consequences. Indeed, when treatment assignment is not ignorable, the OLS regression estimate of treatment effect is likely to be biased and inefficient. Furthermore, the consequences are worse when important predictors are omitted and in an observational study when hidden selection bias is present (Rosenbaum, 2002b). What can be done? This question served as the original motivation for statisticians and econometricians to develop new methods for program evaluation. As a part of this work, new analytic models have been designed for observational studies and, more generally, for nonexperimental approaches that may be used when treatment assignment is ignorable. The growing consensus among statisticians and econometricians is that OLS regression or simple covariance control is no longer the method of choice, although this statement runs the risk of oversimplification. 2.5.1 Design of Observational Study Under the counterfactual framework, violations of the ignorable treatment assignment and SUTVA assumptions may be viewed as failures in conducting a randomized experiment, although the failures may cover a range of situations, such as failure to conduct an experiment in the first place, broken randomization 67 due to treatment noncompliance or randomized studies that suffer from attrition, or the use of an inadequate number of units in randomization such that randomization cannot fully balance data. To correct for these violations, researchers need to have a sound design. Choosing an appropriate method for data analysis should be guided by the research design. Numerous scholars underscore the importance of having a good design for observational studies. Indeed, Rosenbaum (2010) labeled his book as the Design of Observational Studies to emphasize the relevance and importance of design in the entire business of conducting observational research and evaluation. And, one of Rubin's best-known articles, published in 2008, is titled "For Objective Causal Inference, Design Trumps Analysis." From the perspective of statistical tradition, observational studies aim to accomplish the same goal of causal inference as randomized experiments; therefore, the first design issue is to view observational studies as approximations of randomized experiments. Rubin (2008) argued, "A crucial idea when trying to estimate causal effects from an observational dataset is to conceptualize the observational dataset as having arisen from a complex randomized experiment, where the results used to assign the treatment conditions have been lost and must be reconstructed" (p. 815). In this context, addressing the violation of the unconfoundedness assumption is analogous to an effort to reconstruct and balance the data. Specifically, there are six important tasks involved in the design of observational studies: (1) conceptualize the observational study as having arisen from a complex randomized experiment, (2) understand what was the hypothetical randomized experiment that led to the observed data set, (3) evaluate whether the sample sizes in the data set are adequate, (4) understand who are the decision makers for treatment assignment and what measurements were available to them, (5) examine whether key covariates are measured well, and (6) evaluate whether balance can be achieved on key covariates (Rubin, 2008). 2.5.2 The Seven Models The seven models presented in this book relax the nonignorable treatment assignment assumption: (a) by considering analytic approaches that do not rely on strong assumptions requiring distributional and functional forms, (b) by rebalancing assigned conditions so that they become more akin to data generated by randomization, and (c) by estimating counterfactuals that represent different treatment effects of interest by using a variety of statistics (i.e., means, proportions). In estimating counterfactuals, the seven models have the following core features: 1. Heckman's sample selection model (1978, 1979) and its revision on 68 estimating treatment effects (Maddala, 1983). The crucial features of these models are (a) an explicit modeling of the structure of selection, (b) a switching regression that seeks exogenous factors determining the switch of study participants between two regimes (i.e., the treated and nontreated regimes), and (c) the use of the conditional probability of receiving treatment in the estimation of treatment effects. 2. Propensity score matching model (Rosenbaum, 2002b; Rosenbaum & Rubin, 1983). The fundamental feature of the propensity score matching model is that it balances data through resampling or matching nontreated participants to treated ones on probabilities of receiving treatment (i.e., the propensity scores) and permits follow-up bivariate or multivariate analysis (e.g., stratified analysis of outcomes within quintiles of propensity scores, OLS regression, survival modeling, structural equation modeling, hierarchical linear modeling) as would be performed on a sample generated by a randomized experiment. Reducing the dimensionality of covariates to a one-dimensional score—the propensity—is a substantial contribution that leverages matching. From this perspective, the estimation of propensity scores and use of propensity score matching is the "most basic ingredient of an unconfounded assignment mechanism" (Rubin, 2008, p. 813). Addressing the reduction of sample sizes from greedy (e.g., 1:1) matching, an optimal matching procedure using network flow theory can retain the original sample size where the counterfactuals are based on an optimally full matched sample or optimally matched sample using variable ratios of treated to untreated participants. Estimation of counterfactuals may employ multilevel modeling procedures to account for clustering effects that exist in both the model estimating the propensity scores and the model for outcome analysis. 3. Propensity score subclassification model (Rosenbaum & Rubin, 1983, 1984). Extending the classic work of Cochran (1968), Rosenbaum and Rubin proved that balancing on propensity scores represents all available covariates and yields a one-dimensional score through which one can successfully perform subclassification. The procedure involves estimating the counterfactual for each subclass obtained through propensity score subclassification, aggregating counterfactuals from all subclasses to estimate the average treatment effect for the entire sample and the variance associated with it, and finally testing whether the treatment effect for the sample is statistically significant. Structural equation modeling (SEM) may be performed in conjunction with subclassification, and a test of subclass differences of key SEM parameters is often a built-in procedure in this type of analyses. 4. Propensity score weighting model (Hirano & Imbens, 2001; Hirano et al., 2003; McCaffrey et al. 2004). The key feature of this method is the treatment of estimated propensity scores as sampling weights to perform a weighted outcome 69 analysis. Counterfactuals are estimated through a regression or regression-type model, and the control of selection biases is achieved through weighting, rather than a direct inclusion of covariates as independent variables in a regression model. 5. Matching estimators model (Abadie & Imbens, 2002, 2006). The key feature of this method is the direct imputation of counterfactuals for both treated and nontreated participants by using a vector norm with a positive definite matrix (i.e., the Mahalanobis metric or the inverse of sample variance matrix). Various types of treatment effects may be estimated: (a) the sample average treatment effect (SATE), (b) the sample average treatment effect for the treated (SATT), (c) the sample average treatment effect for the controls (SATC), and (d) the equivalent effects for the population (i.e., population average treatment effect [PATE], population average treatment effect for the treated [PATTE], and population average treatment effect for the controls [PATTC]). Standard errors corresponding to these sample average treatment effects are developed and used in significance tests. 6. Propensity score analysis with nonparametric regression model (Heckman, Ichimura, & Todd, 1997, 1998). The critical feature of this method is the comparison of each treated participant to all nontreated participants based on distances between propensity scores. A nonparametric regression such as local linear matching is used to produce an estimate of the average treatment effect for the treatment group. By applying the method to data at two time points, this approach estimates the average treatment effect for the treated in a dynamic fashion, known as difference-in-differences. 7. Propensity score analysis of categorical or continuous treatments model (Hirano & Imbens, 2004; Imbens, 2000; Joffe & Rosenbaum, 1999). This class of methods is an extension of propensity score analysis of binary treatment conditions to multiple treatment levels, where the researchers are primarily interested in the effects of treatment dosage. Counterfactuals are estimated either through a single scalar of propensity scores (Joffe & Rosenbaum, 1999) or through estimating generalized propensity scores (GPS). The GPS (Hirano & Imbens, 2004) approach involves the following steps: estimating GPS using a maximum likelihood regression, estimating the conditional expectation of the outcome given the treatment and GPS, and estimating the dose-response function to discern treatment effects as well as their 95% confidence bands. It is worth noting that all the models or methods were not originally developed to correct for nonignorable treatment assignment. Quite the contrary, some of these models still assume that treatment assignment is strongly ignorable. According to Rosenbaum and Rubin (1983), showing "strong 70 ignorability" allows analysts to evaluate a nonrandomized experiment as if it had come from a randomized experiment. However, in many evaluations, this assumption cannot be justified. Notwithstanding, in most studies, we wish to conduct analyses under the assumption of ignorability (Abadie, Drukker, Herr, & Imbens, p. 292). Instead of correcting for the violation of the assumption about strongly ignorable treatment assignment, the corrective approaches (i.e., the methods covered in this book) take various measures to control selection bias. These include, for example, (a) relaxation of the assumption (e.g., instead of assuming conditional independence or full independence [Heckman, Ichimura, & Todd, 1997, 1998] or assuming mean independence by only requiring that conditional on covariates, the mean outcome under control condition for the treated cases be equal to the mean outcome under the treated condition for the controls), (b) modeling the treatment assignment process directly by treating the dummy treatment condition as an endogenous variable and using a two-step estimating procedure (i.e., the Heckman sample selection model), (c) developing a onedimensional propensity score so that biases due to observed covariates can be removed by conditioning solely on the propensity score (i.e., Rosenbaum and Rubin's propensity score matching model and Heckman and colleagues' propensity score analysis with nonparametric regression), and (d) employing bias-corrected matching with a robust variance estimator to balance covariates between treatment conditions (i.e., the matching estimators). Because of these features, the methods we describe offer advantages over OLS regression, regression-type models, and other simple corrective methods. Rapidly being developed and refined, propensity score methods are showing usefulness when compared with traditional approaches. Parenthetically, most of these methods correct for overt selection bias only. The sample selection and treatment effect models are exceptions that may partially correct for hidden selections. But, on balance, the models do nothing to directly correct for hidden selection bias. It is for this reason that the randomized experiment remains a gold standard. When properly implemented, it corrects for both overt and hidden selection bias. 2.5.3 Other Balancing Methods We chose to include seven models in this text because they are robust, efficient, and effective in addressing questions that arise commonly in social behavioral and health evaluations. Although the choice of models is based on our own experience, many applications can be found in biostatistics, business, economics, education, epidemiology, medicine, nursing, psychology, public health, social work, and sociology. There are certainly other models that accomplish the same goal of balancing data. To offer a larger perspective, we provide a brief review of additional models. 71 Imbens (2004) summarized five groups of models that serve the common goal of estimating average treatment effects: (1) regression estimators that rely on consistent estimation of key regression functions; (2) matching estimators that compare outcomes across pairs of matched treated and control units, with each unit matched to a fixed number of observations in the opposite treatment; (3) estimators characterized by a central role of the propensity score (i.e., there are four leading approaches in this category: weighting by the reciprocal of the propensity score, blocking on the propensity score, regression on the propensity score, and matching on the propensity score); (4) estimators that rely on a combination of these methods, typically combining regression with one of its alternatives; and (5) Bayesian approaches to inference for average treatment effects. In addition, Winship and Morgan (1999) and Morgan and Winship (2007) reviewed five methods, including research designs that are intended to improve causal interpretation in the context of nonignorable treatment assignment. These include (1) regression discontinuity designs, (2) instrumental variables (IV) approaches, (3) interrupted time-series designs, (4) differential rate of growth models, and (5) analysis of covariance models. Separate from mainstream propensity score models and advances in design, other approaches to causal inference warrant attention. James Robins, for example, developed analytic methods known as marginal structural models that are appropriate for drawing causal inferences from complex observational and randomized studies with time-varying exposure of treatment (Robins, 1999a, 1999b; Robins, Hernn, & Brumback, 2000). Judea Pearl (2000) and others (Glymour & Cooper, 1999; Spirtes, Glymour, & Scheines, 1993) developed a formal framework to determine which of many conditional distributions could be estimated from data using an approach known as directed acyclic graphs. Of these models, the IV approach shares common features with some models discussed in this book, particularly, the switching regression model described in Chapter 4. The IV approach is among the earliest attempts in econometrics to address the endogeneity bias problem, and it has been shown to be useful in estimating treatment effects. Because of its similarities with approaches discussed in this book as well as its popularity in correcting the endogeneity problem when randomized experimentation is not feasible, we give it a detailed review. We also briefly describe the basic ideas of regression discontinuity designs so that readers are aware of how the same analytic issues can be addressed by methods other than propensity score analysis. We do not intend to provide a lengthy treatment of either of these two methods because they are not based on propensity scores. 2.5.4 Instrumental Variables Estimator After OLS regression, the instrumental variable (IV) approach is perhaps the 72 second most widely practiced method in economic research (Wooldridge, 2002). As mentioned earlier, selection bias is a problem of endogeneity in regression analysis. That is, the lack of a randomization mechanism causes the residual term in regression to be correlated with one or more independent variables. To solve the problem, researchers may find an observed variable z_1 that satisfies the following two conditions: z_1 is not correlated with the residual term, but z_1 is highly correlated with the independent variable that causes endogeneity. If z_1 meets these two conditions, then z_1 is called an instrumental variable. The instrument z_1 may not necessarily be a single variable and can be a vector that involves two or more variables. Under this condition, researchers can use a two-stage least squares estimator to estimate the regression coefficients and treatment effects. Together, the method using either a single or a vector of instrumental variables is called the instrumental variables estimator. Following Wooldridge (2002), we describe the basic setup of IV next. Formally, consider a linear population model: Note that in this model, x_k is correlated with ϵ (i.e., $\text{Cov}(x_k, \epsilon) \neq 0$), and x_k is potentially endogenous. To facilitate the discussion, we think of ϵ as containing one omitted variable that is uncorrelated with all explanatory variables except x_k . In the practice of IV, researchers could consider a set of omitted variables. Under such a condition, the model would use multiple instruments. All omitted variables meeting the required conditions are called multiple instruments. To solve the problem of endogeneity bias, the analyst needs to find an observed variable, z_1 , that satisfies the following two conditions: (1) z_1 is uncorrelated with ϵ , or $\text{Cov}(z_1, \epsilon) = 0$, and (2) z_1 is correlated with x_k , meaning that the linear projection of x_k onto all exogenous variables exists. This is otherwise stated as where, by definition, $E(\epsilon) = 0$ and ϵ is uncorrelated with x_1, x_2, \dots , and $x_k - 1, z_1$; the key assumption is that the coefficient on z_1 is nonzero, or $\theta \neq 0$. Next, consider the model (i.e., Equation 2.15) where the constant is absorbed into x_0 so that $x = (1, x_2, \dots, x_k)$. Write the $1 \times k$ vector of all exogenous variables as $z = (1, x_2, \dots, x_k - 1, z_1)$. The 73 preceding two conditions about z_1 imply the K population orthogonality conditions, or Multiplying Equation 2.16 through by z' , taking expectations, and using Equation 2.17, we have where $E(z'x)$ is $K \times K$ and $E(z'y)$ is $K \times 1$. Equation 2.18 represents a system of K linear equations in the K unknowns β_1, \dots, β_k . This system has a unique solution if and only if the $K \times k$ matrix $E(z'x)$ has full rank, or the rank of $E(z'x)$ is K . Under this condition, the solution to β is Thus, given a random sample $\{(x_i, y_i) : i = 1, 2, \dots, N\}$ from the population, the analyst can obtain the instrumental variables estimator of β as The above model (2.19) specifies one instrumental variable, z_1 . In practice, the analyst may have more than one instrumental variable for x_k , such as M instruments of x_k ($i = 1, 2, \dots, z_M$). Define the vector of exogenous variables as $z = (1, x_1, x_2, \dots, x_k - 1, z_1, \dots, z_M)$. Estimated regression coefficients for $x_1, x_2, \dots, x_k - 1, x_k$ can be obtained through the following two stages: (1) obtain the fitted values from the regression x_k on $x_1, x_2, \dots, x_k - 1, z_1, \dots, z_M$, which is called the first-stage regression, and (2) run the regression y on $x_1, x_2, \dots, x_k - 1$, to obtain estimated regression coefficients, which is called the second-stage regression. For each observation i , define the vector $i = 1, 2, \dots, N$. Using from the second-stage regression gives the IV estimator where unity is also the first element of x_i . For details of the IV model with multiple instruments, readers are referred to Wooldridge (2002, pp. 90–92). The two-stage least squares estimator under certain assumptions is the most efficient IV estimator. Wooldridge (2002, pp. 92–97) gives a formal treatment to this estimator and provides proofs for important properties, including consistency, asymptotic normality, and asymptotic efficiency, of the two-stage 74 least squares estimator. In practice, finding instrumental variables can be challenging. It is often difficult to find an instrumental variable z_1 (or M instrumental variables z_1, \dots, z_M) that meets the two conditions required by the procedure; namely, the instrument is not correlated with the residual of the regression model that suffers from endogeneity but it is highly correlated with the independent variable that causes endogeneity. The application of the IV approach requires a thorough understanding of the study phenomenon; processes generating all study variables, including exogenous variables that produce endogeneity problems; independent variables used in the regression model; variables that are not used in the regression; and the outcome variable and its relationships with the independent variables used and not used in the regression. In essence, the IV model requires that researchers have an excellent understanding of the substantive theories as well as the processes generating the data. Although finding good instruments is challenging, innovative studies have employed interesting IVs and applied the approach to address challenging research questions. For instance, in a study on the effects of education on wages, the residual of the regression equation is correlated with education because it contains omitted ability. Angrist and Krueger (1991) used a dichotomous variable indicating whether a study subject was born in the first quarter of the birth year (= 1 if the subject was born in the first quarter and 0 otherwise). They argued that compulsory school attendance laws induce a relationship between education and the quarter of birth: At least some people are forced, by law, to attend school longer than they otherwise would. The birth quarter in this context is obviously random and not correlated with other omitted variables of the regression model. Another well-known example of an IV is the study of the effect of serving in the Vietnam War on the earnings of men (Angrist, 1990). Prior research showed that participation in the military is not necessarily exogenous to unobserved factors that affect earnings even after controlling for education, nonmilitary experience, and so on. Angrist (1990) found that men with a lower draft lottery number were more likely to serve in the military during the Vietnam War, and hence, he used the draft lottery number, initiated in 1969, as an instrument of the binary Vietnam War participation indicator. A similar idea (i.e., of using lottery number as an instrument for serving in the army during the Vietnam War) was employed in a well-known study estimating the effect of veteran status on mortality (Angrist et al., 1996). The study employed an IV to estimate local average treatment effect. Angrist et al. (1996) showed how the IV estimator can be given a precise and straightforward causal interpretation in the potential outcomes framework, despite nonignorability of treatment received. This interpretation avoids drawbacks of the standard 75 structural equation framework, such as constant effects for all units, and delineates critical assumptions needed for a causal interpretation. The IV approach provides an alternative to a more conventional intention-to-treat analysis, which focuses solely on the average causal effect of assignment on the outcome. (p. 444) Other studies that have chosen instrumental variables cleverly and innovatively include the Hoxby (1994) study that used topographic features—natural boundaries created by rivers—as the IV for the concentration of public schools within a school district, where the author was interested in estimating the effects of competition among public schools on student performance; the Evans and Schwab (1995) study examining the effects of attending a Catholic high school on various outcomes, in which the authors used whether a student was Catholic as the IV for attending a Catholic high school; and the Card (1995a) study on the effects of schooling on wages, where the author used a dummy variable that indicated whether a man grew up in the vicinity of a 4-year college as an instrumental variable for years of schooling. Wooldridge (2002, pp. 87–89) provides an excellent review and summary of these studies. It's worth noting that just like the propensity score approach, these IV studies are also controversial and have triggered debates and criticisms. Opponents primarily challenge the problem of a weak correlation between the instruments and the endogenous explanatory variable in these studies (e.g., Bound, Jaeger, & Baker, 1995; Rothstein, 2007). The debate regarding the advantages and disadvantages of the IV approach is ongoing. In addition to empirical challenges in finding good instruments, Wooldridge (2002) finds two potential pitfalls with the two-stage least squares estimator: (1) Unlike OLS under a zero conditional mean assumption, IV methods are never unbiased when at least one explanatory variable is endogenous in the model, and (2) the standard errors estimated by the two-stage least squares or other IV estimators have a tendency to be "large," which may lead to insignificant coefficients or standard errors that are much larger than those estimated by OLS. Heckman (1997) examined the use of the IV approach to estimate the mean effect of treatment on the treated, the mean effect of treatment on randomly selected persons, and the local average treatment effect. He paid special attention to which economic questions were addressed by these parameters and concluded that when responses to treatment vary, the standard argument justifying the use of instrumental variables fails unless person-specific responses to treatment do not influence the decision to participate in the program being evaluated. This condition requires that participant gains from a program—which cannot be predicted from variables in outcome equations—have no influence on the participation decisions of program participants. 2.5.5 Regression Discontinuity Designs 76 Regression discontinuity designs (RDDs) have also drawn the attention of a growing number of researchers. These designs have been increasingly employed in evaluation studies. RDD is a quasi-experimental approach that evaluates the treatment effect by assigning a cutoff or threshold value above or below which a treatment is assigned. By comparing observations lying closely on either side of the threshold, it is possible to estimate the local treatment effect. The method is similar to an interrupted time-series design that compares outcomes before and after an intervention, except that the treatment in RDD is a function of a variable other than time. The RDD method was first proposed by Thistlewaite and Campbell (1960) when they analyzed the effect of student scholarships on career aspirations. In practice, researchers using RDD may distinguish between two general settings: the sharp and the fuzzy regression discontinuity designs. The estimation of treatment effects with both designs can be obtained using a standard nonparametric regression approach such as loess with an appropriately specified kernel function and bandwidth (Imbens & Lemieux, 2008). Discontinuity designs have two assumptions: (1) Treatment assignment is equally as good as random selection at the threshold for treatment, and (2) individuals are sampled independently. Violations of these assumptions lead to biased estimation of treatment effects. The most severe problem with RDD is misspecification of the functional form of the relation between treatment and outcome. Specifically, users run the risk of misinterpreting a nonlinear relationship between treatment and outcome as a discontinuity. Counterfactual values must be extrapolated from observed data below and above the application of the treatment. If the assumptions built into the RDD of extrapolation are unreasonable, then causal estimates are incorrect (Morgan & Winship, 2007). Propensity score methods fall within the broad class of procedures being developed for use when random assignment is not possible or is compromised. These procedures include IV analysis and regression discontinuity designs. They include also directed acyclic graphs and marginal structural models. In the remaining chapters of the book, we describe seven propensity score models that have immediate applications in the social and health sciences and for which software is generally available. We focus on in vivo application more than on theory and proofs. We turn now to basic ideas underlying all seven models. 2.6 THE UNDERLYING LOGIC OF STATISTICAL INFERENCE When a treatment is found to be effective (or not effective), evaluators often want to generalize the finding to the population represented by the sample. They ask whether or not the treatment effect is zero (i.e., perform a nondirectional 77 test) or is greater (less) than some cutoff value (i.e., perform a directional test) in the population. This is commonly known as statistical inference, a process of estimating unknown population parameters from known sample statistics. Typically, such an inference involves the calculation of a standard error to conduct a hypothesis test or to estimate a confidence interval. The statistical inference of treatment effects stems from the tradition of randomized experimentation developed by Sir Ronald Fisher (1935/1971). The procedure is called a permutation test (also known as a randomization test, a re-randomization test, or an exact test) in that it makes a series of assumptions about the sample. When generalizing, researchers often find that one or more of these assumptions are violated, and thus, they have to develop strategies for statistical inference that deal with estimation when assumptions are differentially tenable. In this section, we review the underlying logic of statistical inference for both randomized experiments and observational studies. We argue that much of the statistical inference in observational studies follows the logic of statistical inference for randomized experiments and that checking the tenability of assumptions embedded in permutation tests is crucial in drawing statistical inferences for observational studies. Statistical inference always involves a comparison of sample statistics to statistics from a reference distribution. Although in testing treatment effects from a randomized experiment, researchers often employ a parametric distribution (such as the normal distribution, the t distribution, and the F distribution) to perform a so-called parametric test, such a parametric distribution is not the reference distribution per se; rather, it is an approximation of a randomization distribution. Researchers use parametric distributions in significance testing because these distributions "are approximations to randomization distributions—they are good approximations to the extent that they reproduce randomization inferences with reduced computational effort" (Rosenbaum, 2002a, p. 289). Strictly speaking, all statistical tests performed in randomized experiments are nonparametric tests using randomization distributions as a reference. Permutation tests are based on reference distributions developed by calculating all possible values of a test statistic under rearrangements of the "labels" on the observed data points. In other words, the method by which treatments are allocated to participants in an experimental design is mirrored in the analysis of that design. If the labels are exchangeable under the null hypothesis, then the resulting tests yield exact significance levels. Confidence intervals can then be derived from the tests. Recall the permutation test of a British woman's tea-tasting ability (see Section 1.3.1). To reject the null hypothesis that the taster has no ability in discriminating two kinds of tea (or, equivalently, testing the hypothesis that she makes a correct judgment by accidentally guessing it right), the evaluator lists—of presenting eight all 70 possible ways—that is, 78 cups of tea with four cups adding the milk first and four cups adding the tea infusing first. That is, the evaluator builds a reference distribution of "11110000, 10101010, 00001111, . . ." that contains 70 elements in the series. The inference is drawn on a basis of the following logic: The taster could guess (choose) any one outcome out of the 70 possible ones; the probability of guessing the right outcome is $1/70 = .0124$, which is a low probability; thus, the null hypothesis of "no ability" can be rejected at a statistical significance level of $p < .05$. If the definition of "true ability" is relaxed to allow for six exact agreements rather than eight exact agreements (i.e., six cups are selected in an order that matches the order of actual presentation), then there are a total of 17 possible ways to have six agreements, and the probability of falsely rejecting the null hypothesis increases to $17/70 = .243$. The null hypothesis cannot be rejected at a .05 level. Under this relaxed definition, we should be more conservative, or ought to be more reluctant, in declaring that the tea taster has true ability. All randomization tests listed in Section 1.3.2 (i.e., Fisher's exact test, the Mantel-Haenszel test, McNemar's test, Mantel's extension of the Mantel-Haenszel test, Wilcoxon's rank sum test, and the Hodges and Lehmann signed rank test) are permutation tests that use randomization distributions as references and calculate all possible values of the test statistic to draw an inference. For this reason, this type of test is called nonparametric—it relies on distributions of all possible outcomes. In contrast, parametric tests employ parametric distributions as references. To illustrate, we now follow Lehmann to show the underlying logic of statistical inference employed in Wilcoxon's rank sum test (Lehmann, 2006). Wilcoxon's rank sum test may be used to evaluate an outcome variable that takes many numerical values (i.e., an interval or ratio variable). To evaluate treatment effects, N participants (patients, students, etc.) are divided at random into a group of size n that receives a treatment and a control group of size m that does not receive treatment. At the termination of the study, the participants are ranked according to some response that measures treatment effectiveness. The null hypothesis of no treatment effect is rejected, and the superiority of the treatment is acknowledged, if in this ranking the n treated participants rank sufficiently high. The significance test calculates the statistical significance or probability of falsely rejecting the null hypothesis based on the following where k is the sum of treated participants' equation: ranks under the null hypothesis of no treatment effect, c is a prespecified value at which one wants to evaluate its probability, and w is the frequency (i.e., number of times) of having value k under the null hypothesis. Precisely, if there were no treatment effect, then we could think of each participant's rank as attached before assignments to treatment and control are made. Suppose we have a total of $N = 5$ participants; $n = 3$ are assigned to treatment, and $m = 2$ are 79 assigned to control. Under the null hypothesis of no treatment effect, the five participants may be ranked as 1, 2, 3, 4, and 5. With five participants taken three at a time to form the treatment group, there are a total of 10 possible groupings of outcome ranks under the null hypothesis: The rank sum of treated participants corresponding to each of the previous groups may look like the following: The probabilities of taking various rank sum values under the null hypothesis of no treatment effect are displayed below: For instance, under the null hypothesis of no treatment effect, there are two possible ways to have a rank sum $k = 10$ (i.e., $w = 2$, when the treatment group is composed of treated participants whose ranks are [1, 4, 5] or is composed of treated participants whose ranks are [2, 3, 5]). Because there are a total of 10 possible ways to form the treatment and control groups, the probability of having a rank sum $k = 10$ is $2/10 = .2$. The above probabilities constitute the randomization distribution (i.e., the reference distribution) for this permutation test. From any real sample, one will observe a realized outcome that takes any one of the seven k values (i.e., 6, 7, . . . , 12). Thus, a significance test of no treatment effect is to compare the observed rank sum from the sample data with the preceding distribution and check the probability of having such a rank sum from the reference. If the probability is small, then one can

reject the null hypothesis and conclude that in the population, the treatment effect is not equal to zero. Suppose that the intervention being evaluated is an educational program that aims to promote academic achievement. After implementing the intervention, the 80 program officer observes that the three treated participants have academic test scores of 90, 95, 99, and the two control participants have test scores of 87, 89, respectively. Converting these outcome values to ranks, the three treated participants have ranks of 3, 4, 5, and the two control participants have ranks of 1, 2, respectively. Thus, the rank sum of the treated group observed from the sample is $3 + 4 + 5 = 12$. This observed statistic is then compared with the reference distribution, and the probability of having a rank sum of 12 under the null hypothesis of no treatment effect is $PH(k = 12) = .1$. Because this probability is small, we can reject the null hypothesis of no treatment effect at a significance level of .1 and conclude that the intervention may be effective in the population. Note that in the preceding illustration, we used very small numbers of N , n , and m , and thus, the statistical significance for this example cannot reach the conventional level of .05—the smallest probability in the distribution of this illustrating example is .1. In typical evaluations, N , n , and m tend to be larger, and a significance level of .05 can be attained. Wilcoxon's rank sum test, as described earlier, employs a randomization distribution based on the null hypothesis of no treatment effect. The exact probability of having a rank sum equal to a specific value is calculated, and such a calculation is based on all possible arrangements of N participants into n and m . The probabilities of having all possible values of rank sum based on all possible arrangements of N participants into n and m are then calculated, and it is these probabilities that constitute the reference for significance testing. Comparing the observed rank sum of treated participants from a real sample with the reference, evaluators draw a conclusion about whether they can reject the null hypothesis of no treatment effect at a statistically significant level. The earlier illustrations show a primary feature of statistical inference involving permutation tests: These tests build up a distribution that exhausts all possible arrangements of study participants under a given N , n , and m and calculate all possible probabilities of having a particular outcome (e.g., the specific rank sum of treated participants) under the null hypothesis of no treatment effect. This provides a significance test for treatment in a realized sample. To make the statistical inference valid, we must ensure that the sample being evaluated meets certain assumptions. At the minimum, these assumptions include the following: (a) The sample is a real random sample from a well-defined population, (b) each participant has a known probability of receiving treatment, (c) treatment assignment is strongly ignorable, (d) the individual-level treatment effect (i.e., the difference between observed and potential outcomes $\tau_i = Y_{1i} - Y_{0i}$) is constant, (e) there is a stable unit treatment value, and (f) probabilities of receiving treatment overlap between treated and control groups. When a randomized experiment in the strict form of Fisher's definition is implemented, all the previous assumptions are met, and therefore, statistical inference using permutation tests is valid. Challenges arise when evaluators move from randomized experiments to observational studies, because in the latter case, one or more of the preceding assumptions are not tenable. So what is the underlying logic of statistical inference for observational studies? To answer this question, we draw on perspectives from Rosenbaum (2002a, 2002b) and Imbens (2004). Rosenbaum's framework follows the logic used in the randomized experiments and is an extension of permutation tests to observational studies. To begin with, Rosenbaum examines covariance adjustment in completely randomized experiments. In the earlier examples, for simplicity of exposition, we did not use any covariates. In real evaluations of randomized experiments, evaluators typically would have covariates and want to control them in the analysis. Rosenbaum shows that testing the null hypothesis of no treatment effect in studies with covariates follows the permutation approach, with the added task of fitting a linear or generalized linear model. After fitting a linear model that controls for covariates, the residuals for both conditions (treatment and control groups) are fixed and known; therefore, one can apply Wilcoxon's rank sum test or similar permutation tests (e.g., the Hodges-Lehmann aligned rank test) to model-fitted residuals. A propensity score adjustment can be combined with the permutation approach in observational studies with overt bias. "Overt bias . . . can be seen in the data at hand—for instance, prior to treatment, treated participants are observed to have lower incomes than controls" (Rosenbaum, 2002b, p. 71). In this context, one can balance groups by estimating a propensity score, which is a conditional probability of receiving treatment given observed covariates, and then perform conditional permutation tests using a matched sample. Once again, the statistical inference employs the same logic applied to randomized experiments. We describe in detail three such permutation tests after an optimal matching on propensity scores (see Chapter 5): regression adjustment of difference scores based on a sample created by optimal pair matching, outcome analysis using the Hodges-Lehmann aligned rank test based on a sample created by optimal full or variable matching, and regression adjustment using the Hodges-Lehmann aligned rank test based on a sample created by optimal full or variable matching. Finally, Rosenbaum considers statistical inference in observational studies with hidden bias. Hidden bias is similar to overt bias, but it cannot be seen in the data at hand, because measures that might have revealed a selection effect were omitted from data collection. When bias exists but is not observable, one can still perform propensity score matching and conduct statistical tests by comparing treatment and control participants matched on propensity scores. But caution is warranted, and sensitivity analyses should be undertaken before generalizing findings to a population. Surprisingly and importantly, the core component of Rosenbaum's sensitivity analysis involves permutation tests, which include McNemar's test, Wilcoxon's signed rank test, and the Hodges82 Lehmann point and interval estimates for matched pairs, sign-score methods for matching with multiple controls, sensitivity analysis for matching with multiple controls when responses are continuous variables, and sensitivity analysis for comparing two unmatched groups. We review and illustrate some of these methods in Chapter 11. In 2004, Imbens reviewed inference approaches using nonparametric methods to estimate average treatment effects under the unconfoundedness assumption (i.e., the ignorable treatment assignment assumption). He discusses advances in generating sampling distributions by bootstrapping (a method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample) and observes, There is little formal evidence specific for these estimators, but, given that the estimators are asymptotically linear, it is likely that bootstrapping will lead to valid standard errors and confidence intervals at least for the regression propensity score methods. Bootstrapping may be more complicated for matching estimators, as the process introduces discreteness in the distribution that will lead to ties in the matching algorithm. (p. 21) Furthermore, Imbens, Abadie, and others show that the variance estimation employed in the matching estimators (Abadie & Imbens, 2002, 2006) requires no additional nonparametric estimation and may be a good alternative to estimators using bootstrapping. Finally, in the absence of consensus on the best estimation methods, Imbens challenges the field to provide implementable versions of the various estimators that do not require choosing bandwidths (i.e., a user-specified parameter in implementing kernel-based matching; see Chapter 9) or other smoothing parameters and to improve estimation methods so that they can be applied with a large number of covariates and varying degrees of smoothness in the conditional means of the potential outcomes and the propensity scores. In summary, understanding the logic of statistical inference underscores in turn the importance of checking the tenability of statistical assumptions. In general, current estimation methods rely on permutation tests, which have roots in randomized experimentation. We know too little about estimation when a reference distribution is generated by bootstrapping, but this seems promising. Inference becomes especially challenging when nonparametric estimation requires making subjective decisions, such as specifications of bandwidth size, when data contain a large number of covariates, and when sample sizes are small. Caution seems particularly warranted in observational studies. Omission of important variables and measurement error in the covariates—both of which are difficult to detect—justify use of sensitivity analysis. 83 2.7 TYPES OF TREATMENT EFFECTS Unlike many texts that address treatment effects as the net difference between the mean scores of participants in treatment and control conditions, we introduce and discuss a variety of treatment effects. This may seem pedantic, but there are at least four reasons why distinguishing, both conceptually and methodologically, among types of treatment effects is important. First, distinguishing among types of treatment effects is important because of the limitation in solving the fundamental problem of causal inference (see Section 2.2). Recall that at the individual level, the researcher cannot observe both potential outcomes (i.e., outcomes under the condition of treatment and outcomes under the condition of nontreatment) and thus has to rely on group averages to evaluate counterfactuals. The estimation of treatment effects so derived at the population level uses averages or $\tau = E(Y_{1|W = 1}) - E(Y_{0|W = 0})$. As such, the variability in individuals' causal effects $(Y_{1i}|W_i = 1) - (Y_{0i}|W_i = 0)$ would affect the accuracy of an estimated treatment effect. If the variability is large over all units, then $\tau = E(Y_{1|W = 1}) - E(Y_{0|W = 0})$ may not represent the causal effect of a specific unit very well, and under many evaluation circumstances, treatment effects of certain units (groups) serve a central interest. Therefore, it is critical to ask which effect is represented by the standard estimator. It is clear that the effect represented by the standard estimator may not be the same as those arising from the researcher's interest. Second, there are inevitably different ways to define groups and to use different averages to represent counterfactuals. Treatment effects and their surrogate counterfactuals are then multifaceted. Third, SUTVA is both an assumption and a perspective for the evaluation of treatment effects. As such, when social interaction is absent, SUTVA implies that different versions of treatment (or different dosages of the same treatment) should result in different outcomes. This is the rationale that leads evaluators to distinguish two different effects: program efficacy versus program effectiveness. Last, the same issue of types of treatment effects may be approached from a different perspective—modeling treatment effect heterogeneity, a topic that warrants a separate and more detailed discussion (see Section 2.8). Based on our review of the literature, the following seven treatment effects are most frequently discussed by researchers in the field. Although some are related, the key notion is that researchers should distinguish between different effects. That is, we should recognize that different effects require different estimation methods, and by the same token, different estimation methods estimate different effects. 1. Average treatment effect (ATE) or average causal effect: This is the core effect estimated by the standard estimator 84 Under certain assumptions, one can also write it as 2. In most fields, evaluators are interested in evaluating program effectiveness, which indicates how well an intervention works when implemented under conditions of actual application (Shadish et al., 2002, p. 507). Program effectiveness can be measured by the intent-to-treat (ITT) effect. ITT is generally analogous to ATE: "Statisticians have long known that when data are collected using randomized experiments, the difference between the treatment group mean and the control group mean on the outcome is an unbiased estimate of the ITT" (Sobel, 2005, p. 114). In other words, the standard estimator employs counterfactuals (either estimation of the missing-value outcome at the individual level or mean difference between the treated and nontreated groups) to evaluate the overall effectiveness of an intervention as implemented. 3. Over the past 30 years, evaluators have also become sensitive to the differences between effectiveness and efficacy. The treatment assigned to a study participant may not be implemented in the way it was intended. The term efficacy is used to indicate how well an intervention works when it is implemented under conditions of ideal application (Shadish et al., 2002, p. 507). Measuring the efficacy effect (EE) requires a careful monitoring of program implementation and taking measures to warrant intervention fidelity. EE plays a central role in the so-called efficacy subset analysis (ESA) that deliberately measures impact on the basis of treatment exposure or dose. 4. Average treatment effect for the treated (TT) can be expressed as Heckman (1992, 1996, 1997, 2005) argued that in a variety of policy contexts, it is the TT that is of substantive interest. The essence of this argument is that in deciding whether a policy is beneficial, our interest is not whether on average the program is beneficial for all individuals but whether it is beneficial for those individuals who are assigned or who would assign themselves to the treatment (Winship & Morgan, 1999, p. 666). The key notion here is $TT \neq ATE$. 5. Average treatment effect for the untreated (TUT) is an effect parallel to TT for the untreated: Although estimating TUT is not as important as TT, noting the existence of 85 such an effect is a direct application of the Neyman-Rubin model. In policy research, the estimation of TUT addresses (conditionally and unconditionally) the question of how extension of a program to nonparticipants as a group might affect their outcomes (Heckman, 2005, p. 19). The matching estimators described in Chapter 8 offer a direct estimate of TUT, although the effect is labeled as the sample (or population) average treatment effect for the controls (SATC or PATC). 6. Marginal treatment effect (MTE) or its special case of the treatment effect for people at the margin of indifference: In some policy and practice situations, it is important to distinguish between marginal and average returns (Heckman, 2005). For instance, the average student going to college may do better (i.e., have higher grades) than the marginal student who is indifferent about going to school or not. In some circumstances, we wish to evaluate the impact of a program at the margins. Heckman and Vytlacil (1999, 2005) have shown that MTE plays a central role in organizing and interpreting a wide variety of evaluation estimators. 7. Local average treatment effect (LATE): Angrist et al. (1996) outlined a framework for causal inference where assignment to binary treatment is ignorable, but compliance with the assignment is not perfect so that the receipt of treatment is nonignorable. LATE is defined as the average causal effect for compliers. It is not the average treatment effect either for the entire population or for a subpopulation identifiable from observed values. Using the instrumental variables approach, Angrist et al. demonstrated how to estimate LATE. To illustrate the importance of distinguishing different treatment effects, we invoke an example originally developed by Rosenbaum (2002b, pp. 181–183). Using hypothetical data in which responses under the treatment and control conditions are known, it demonstrates the inequality of four effects: Consider a randomized trial in which patients with chronic obstructive pulmonary disease are encouraged to exercise. Table 2.1 presents an artificial data set of 10 patients (i.e., $N = 10$ and $i = 1, \dots, 10$). The treatment, W_i , is encouragement to exercise: $W_i = 1$, signifying encouragement, and $W_i = 0$, signifying no encouragement. The assignment of treatment conditions to patients is randomized. The pair (d_{1i}, d_{0i}) indicates whether patient i would exercise, with or without encouragement, where 1 signifies exercise and 0 indicates no exercise. For example, $i = 1$ would exercise whether encouraged or not, $(d_{11}, d_{01}) = (1, 1)$, whereas $i = 10$ would not exercise in either case, $(d_{11}, d_{01}) = (0, 0)$, but $i = 3$ exercises only if encouraged, $(d_{11}, d_{01}) = (1, 0)$. 86 The response, (Y_{1i}, Y_{0i}) , is a measure of lung function, or forced expiratory volume on a conventional scale, with higher numbers signifying better lung function. By design, the efficacy effect is known in advance ($EE = 5$); that is, switching from no exercise to exercise raises lung function by 5. Note that counterfactuals in this example are hypothesized to be known. For $i = 3$, $W_i = 1$ or exercise is encouraged, $Y_{1i} = 64$ is the outcome under the condition of exercise, and $Y_{0i} = 59$ is the counterfactual (i.e., if the patient did not exercise, the outcome would have been 59), and for this case, the observed outcome $R_i = 64$. In contrast, for $i = 4$, $W_i = 0$ or exercise is not encouraged, $Y_{1i} = 62$ is the counterfactual, and $Y_{0i} = 57$ is the outcome under the condition of no exercise, and for this case, the observed outcome $R_i = 57$. D_i is a measure of compliance with the treatment; $D_i = 0$, signifying exercise actually was not performed; and $D_i = 1$, signifying exercise was performed. So for $i = 2$, even though $W_i = 0$ (no treatment, or exercise is not encouraged), the patient exercised anyway. Likewise, for $i = 10$, even though exercise is encouraged and $W_i = 1$, the patient did not exercise, $D_i = 0$. Comparing the difference between W_i and D_i for each i gives a sense of intervention fidelity. In addition, on the basis of the existence of discrepancies in fidelity, program evaluators claim that treatment effectiveness is not equal to treatment efficacy. Rosenbaum goes further to examine which patients responded to encouragement. Patients $i = 1$ and $i = 2$ would have the best lung function without encouragement, and they will exercise with or without encouragement. Patients $i = 9$ and $i = 10$ would have the poorest lung function without encouragement, and they will not exercise even when encouraged. Patients $i = 3, 4, \dots, 8$ have intermediate lung function without exercise, and they exercise only when encouraged. The key point noted by Rosenbaum is that although treatment assignment or encouragement, W_i , is randomized, compliance with assigned treatment, (d_{1i}, d_{0i}) , is strongly confounded by the health of the patient. Therefore, in this context, how can we estimate the efficacy? Table 2.1 An Artificial Example of Noncompliance With Encouragement (W_i) to Exercise (D_i) 87 Source: Rosenbaum (2002b, p. 182). Reprinted with kind permission of Springer Science + Business Media. To estimate a naive ATE, we might ignore the treatment state (i.e., ignoring W_i) and (naively) take the difference between the mean response of patients who exercised and those who did not exercise (i.e., using D_i as a grouping variable). In this context and using the standard estimator, we would estimate the naive ATE as which is nearly three times the true effect of 5. The problem with this estimate is that the people who exercised were in better health than the people who did not exercise. Alternatively, a researcher might ignore the level of compliance with the treatment and use the treatment state W_i to obtain ATE (i.e., taking the mean difference between those who were encouraged and those who were not). In this context and using the standard estimator, we find that the estimated ATE is nothing more than the intent-to-treat (ITT) effect: which is much less than the true effect of 5. This calculation demonstrates that ITT is an estimate of program effectiveness but not of program efficacy. Finally, a researcher might ignore the level of compliance and estimate the average treatment effect for the treated (TT) by taking the average differences between Y_{1i} and Y_{0i} for the five treated patients: 88 Although TT is substantially lower than efficacy, this is an effect that serves a central substantive interest in many policy and practice evaluations. In sum, this example illustrates the fundamental differences among four treatment effects, $EE \neq ITT$ ($ATE \neq TT \neq$ Naive ATE, and one similarity, $ITT = ATE$). Our purpose for showing this example is not to argue which estimate is the best but to show the importance of estimating appropriate treatment effects using appropriate methods suitable for research questions. 2.8 TREATMENT EFFECT HETEROGENEITY In the social and health sciences, researchers often need to test heterogeneous treatment effects. This stems from substantive theories and the designs of observational studies in which study participants are hypothesized to respond to treatments, interventions, experiments, or other types of stimuli differentially. The coefficient of an indicator variable measuring treatment condition often does not reflect the whole range of complexity within treatment effects. Treatment effect heterogeneity serves important functions in addressing substantive research and evaluation questions. For this reason, we give the topic separate treatment here. In this section, we discuss the need to model treatment effect heterogeneity and we describe two tests developed by Crump, Hotz, Imbens, and Mitnik (2008). With these tests, not only can researchers test whether a conditional average treatment effect is zero or whether a conditional average treatment effect is constant among subpopulations, but also they can use these tests to gauge whether the strongly ignorable treatment assignment assumption is plausible in a real setting, an assumption that is in general untestable. 2.8.1 The Importance of Studying Treatment Effect Heterogeneity Treatment effects are by no means uniform across subpopulations. Consider the three treatment effects depicted in the previous section: the average treatment effect (ATE), the average treatment effect for the treated (TT), and the average treatment effect for the untreated (TUT). Xie, Brand, and Jann (2012) show that these three quantities should not always be identical, and differences in these quantities reveal treatment effect heterogeneity. Xie et al. show, in addition, that the standard estimator for ATE is valid if and only if treatment effect heterogeneity is absent. By definition and using the counterfactual framework, ATE is the expected difference between two outcomes, or $ATE = E(Y_{1i} - Y_{0i})$. Using the iterative expectation rule, Xie et al. show that the quantity of ATE can be further decomposed as 89 where q is the proportion of untreated participants. Note that the first term in the is the ATE estimated by the standard above equation, estimator. The estimator is valid and unbiased, if and only if the last two terms are equal to zero. Xie et al. underscore that in reality, these two terms often are not equal to zero; therefore, the standard estimator of ATE assumes no treatment effect heterogeneity. When these two terms are not equal to zero or, equivalently, when treatment effect heterogeneity is present, using the standard estimator for ATE is biased. Specifically, these two terms indicate two types of selection biases that are produced by ignorance of the treatment effect heterogeneity. is the average difference between the two First, the term groups in outcomes if neither group receives the treatment. Xie et al. (2012) call this "pretreatment heterogeneity bias." This source of selection bias exists, for instance, when preschool children who attended Head Start programs, which are designed typically for low-income children and their families, are compared unfavorably with other children who did not attend Head Start programs. Comparisons would be affected by the absence of a control for family socioeconomic resources. Second, the term $(TT - TUT)q$ indicates the difference in the average treatment effect between the two groups, TT and TUT, weighted by the proportion untreated, q . Xie et al. (2012) call this "treatment-effect heterogeneity bias." This source of selection bias exists, for instance, when researchers ignore the fact that attending college and earning a degree is selective. An evaluation of the effect of higher education should account for the tendency of colleges to attract people who are likely to gain more from college experiences. Crump et al. (2008) developed two nonparametric tests of treatment effect heterogeneity. The first test is for the null hypothesis that the treatment has a zero average effect for all subpopulations defined by covariates. The second test is for the null hypothesis that the average effect conditional on the covariates is identical for all subpopulations. Section 2.8.4 describes these two tests, and Section 2.8.5 illustrates their applications with an empirical example. The motivation for developing these two tests, according to the authors, was threefold. The first was to address substantive questions. In many projects, researchers are primarily interested in establishing whether the average treatment effect differs from zero; when this is true (i.e., when there is evidence supporting a nonzero ATE), researchers may be further interested in whether there are subpopulations for which the effect is substantively and statistically significant. A concrete example is a test of the effectiveness of a new drug. The evaluators in such a context are interested not only in whether a new drug has a nonzero average effect but whether it has a nonzero (positive or negative) effect 90 for identifiable subgroups in the population. The presence of an effect might permit better targeting who should or should not use the drug: "If one finds that there is compelling evidence that the program has nonzero effect for some subpopulations, one may then further investigate which subpopulations these are, and whether the effects for these subpopulations are substantively important" (Crump et al., 2008, p. 392). In practice, each observed covariate available to the evaluator defines a subpopulation, and therefore, one faces a challenge to test many null hypotheses about a zero treatment effect for these subpopulations. The test Crump et al. (2008) developed offers a single test for zero conditional average treatment effects so that the multiple-testing problem is avoided. The second part of the motivation for developing these tests was concern related to whether there is heterogeneity in the average effect conditional on observed covariates, such as race/ethnicity, education, and age. According to Crump et al. (2008), "If there is strong evidence in favor of heterogeneous effects, one may be more reluctant to recommend extending the program to populations with different distributions of the covariates" (p. 392). The third motivation for developing tests of treatment effect heterogeneity was related to developing an indirect assessment of the plausibility of the strongly ignorable treatment assignment assumption. As described earlier, this crucial assumption is usually not testable. However, there exist indirect approaches, primarily those developed by Heckman and Hotz (1989) and Rosenbaum (1997), from which users can check whether the assumption is plausible or whether additional efforts should be made if ignorability is obviously not the case. Comparing to these two approaches, the tests developed by Crump et al. (2008) are easier to implement. Because of the importance of checking the unconfoundedness assumption in observational studies and the unique advantages offered by the tests from Crump et al., we give this issue a closer examination in the next subsection. Much of the discussion on testing and modeling treatment effect heterogeneity may be illustrated by the inclusion of interactions in an outcome analysis. By definition, the existence of a significant interaction indicates that the impact of an independent variable on the dependent variable varies by the level of another independent variable. Heterogeneous treatment effects, on one hand, are analogous to the existence of significant interactions in the regression model and reflect slope differences of the treatment among subpopulations. When we say that the treatment effect is heterogeneous, we mean that the treatment effect is not uniform and varies by subpopulations defined by covariates. It may be measured by interactions, such as age group by treatment, race group by treatment, income by treatment, and gender by treatment group indicators. The issue of testing and modeling treatment effect heterogeneity, on the other hand, is more complicated than checking and testing significant interactions. Indeed, treatment effect heterogeneity may not be discovered by testing for 91 interactions. Elwert and Winship (2010) argue that the meaning of "main" effects in interaction models is not always clear. Crump et al. (2008) found that treatment effect heterogeneity exists even when the main treatment effect is not statistically significant (see, e.g., reevaluation of the MDRC study of California's Greater Avenues to Independence [GAIN] programs; Crump et al., 2008, pp. 396–398). As discussed regarding the counterfactual framework, the potential outcome can be estimated only at the group level, so the meaning of interactions in an outcome regression using individuals as units is not clearly defined and, therefore, does not truly show treatment heterogeneity. Xie et al. (2012) recommend focusing on the interaction of the treatment effect and the propensity score as one useful way to study effect heterogeneity. Although testing the interaction of treatment by a propensity score is not the only means for assessing effect heterogeneity and the method is controversial, it is often more interpretable because the propensity score summarizes the relevance of the full range of covariates. According to Xie et al., this is the only interaction that warrants attention if selection bias in models of treatment effect heterogeneity is a concern. On the basis of this rationale, Xie and his colleagues developed three methods to model effect heterogeneity: the stratification-multilevel (SM) method, the matching-smoothing (MS) method, and the smoothing-differencing (SD) method. We discuss and illustrate the SM method in Chapter 6. 2.8.2 Checking the Plausibility of the Unconfoundedness Assumption Following Crump et al. (2008), in this subsection, we describe two types of tests that are useful in assessing the plausibility of the unconfoundedness assumption. The first set of tests was developed by Heckman and Hotz (1989). Partitioning the vector of covariates X into two parts, a variable V and the remainder Z , so that $X = (V, Z')$, Heckman and Hotz propose that one can analyze the data (V, W, Z) as if V is the outcome, W is the treatment indicator, and as if unconfoundedness holds conditional on Z . The researcher is certain that the effect of the treatment on V is zero for all units, because V is a pretreatment variable or covariate. Under this context, if the researcher finds statistical evidence suggesting a treatment effect on V , it must be the case that the unconfoundedness conditional on Z is incorrect, or it is suggestive that unconfoundedness is a delicate assumption. The test cannot be viewed as direct evidence against unconfoundedness, because it is not conditional on the full set of covariates $X = (V, Z')$. The tests are effective if the researcher has data on multiple lagged values of the outcome, that is, one may choose V to be the oneperiod lagged value of the outcome. Instead of using multiple lagged values of the outcome, Rosenbaum (1997) considers using two or more control groups. If potential biases would likely be different for both groups, then evidence that all control groups led to similar estimates is suggestive that unconfoundedness may be appropriate. Denote T_i as 92 an indicator for the two control groups, $T_i = 0$ for the first control group and $T_i = 1$ for the second group. The researcher can test whether $Y_i(0) | X_i | T_i$ in the two control groups. If one finds evidence that this pseudo treatment has a systematic effect on the outcome, then it must be the case that unconfoundedness is violated for at least one of the two control groups. The test of a zero conditional average treatment effect developed by Crump et al. (2008) is equivalent to the tests Heckman and Hotz (1989) and Rosenbaum (1997) developed. However, it is much easier to implement. The test does not require the use of lagged values of an outcome variable or multiple control groups, and it can be applied directly to all covariates readily available to the researcher. 2.8.3 A Methodological Note About the Hausman Test of Endogeneity Earlier in the description of the strongly ignorable treatment assignment assumption, we showed that this assumption is equivalent to the OLS assumption regarding the independence of error term from an independent variable. In fact, violation of the unconfoundedness assumption is the same problem of endogeneity one may encounter in a regression analysis. In Subsection 2.8.2, we showed two indirect tests of unconfoundedness, and we mentioned that this assumption in empirical research is virtually not directly testable. To understand the utility of the nonparametric tests that Crump et al. (2008) developed, particularly their usefulness in checking the unconfoundedness assumption, we need to offer a methodological note about a test commonly employed in econometric studies for the endogeneity problem. The test is the Hausman (1978) test of endogeneity, sometimes known as the misspecification test in a regression model. We intend to show that the Hausman test has limitations for accomplishing the goal of checking unconfoundedness. Denoting the dependent variable by y_1 and the potentially endogenous explanatory variable by y_2 , we can express our population regression model as where z is 1×1 (including a constant), δ_1 is 1×1 , and u_1 is the unobserved disturbance. The set of all exogenous variables is denoted by the $1 \times L$ vector z , where z is a strict subset of z . The maintained exogeneity assumption is $E(z'u_1) = 0$. Hausman suggested comparing the OLS and two-stage least squares estimators of α as a formal test of endogeneity: If y_2 is uncorrelated with u_1 , the OLS and two-stage least squares estimators should differ only by sampling error. For more details about the test, we refer to Wooldridge (2002, pp. 118–122). It is important to note that to implement the Hausman test, the analyst should have knowledge about the source of endogeneity, that is, the 93 source of omitted variables in the regression that causes the correlation of the error term with the endogenous explanatory variable. In reality, particularly in the observational studies, this information is often absent, and the analyst does not have a clear sense about unobserved variables that may cause selection biases. Therefore, just like running an IV model where the analyst has difficulty finding an appropriate instrumental variable that is not correlated with the regression error term but is highly correlated with the endogenous explanatory variable, the analyst has the same difficulty in specifying source variables for endogeneity to run the Hausman test. It is for this reason that researchers find appeal in the indirect methods, such as the Heckman and Hotz (1989) test and the Rosenbaum (1997) test, to gauge the level of violation of ignorability. And it is for this reason that the tests developed by Crump et al. (2008) appear to be very useful. 2.8.4 Tests of Treatment Effect Heterogeneity We now return to the tests developed by Crump and colleagues (2008) for treatment effect heterogeneity. With empirical data for treatment indicator W_i ($W_i = 1$, treated; and $W_i = 0$, control), a covariate vector X_i , and outcome variable Y_i for the i th observation, the researcher can test two pairs of hypotheses concerning the conditional average treatment effect $\tau(x)$ when $X = x$. The first pair of hypotheses, called "a test of zero conditional average treatment effect," is Under the null hypothesis H_0 , the average treatment effect is zero for all values of the covariates, whereas under the alternative H_a , there are some values of the covariates for which the treatment effect differs from 0. The second pair of hypotheses, called "a test of constant conditional average treatment effect," is Under the null hypothesis H_0 , all subgroups defined by covariate vector x have treatment effects a constant treatment effect τ , whereas under the alternative of subgroups defined by x do not equal a constant value τ , and therefore, there exists effects heterogeneity. Crump et al. (2008) developed procedures to test the above two pairs of hypotheses. There are two versions of the tests: parametric and nonparametric tests. The Stata programs to implement these tests are available at the following website: omnitk/software.html. Users need to 94 download the program and help files by clicking the add file and help file from the section of "Nonparametric Tests for Treatment Effect Heterogeneity." The Stata ado file is named "test_condate.ado," and the help file is named "test_condate.hlp." Users need to save both files in the folder storing usersupplied ado programs, typically "C:\ado\plus1," in a Windows operating system. Each version of the tests is based on additional assumptions about the data. For the parametric version of the tests, the assumptions are similar to those for most analyses described by this book, such as an independent and identically distributed random sample of (Y_i, W_i, X_i) , unconfoundedness, and overlap of the two groups (treated and nontreated) in the covariate distribution. For the nonparametric version of the tests, Crump et al. (2008) make the following assumptions: the Cartesian product of intervals about the covariate distributions, conditional outcome distributions, and rates for series estimators. The parametric version of the tests is standard. The test statistic T for the first pair of hypotheses H_0 and H_a has a chi-square distribution with K degrees of freedom, where K is the number of covariates being tested, including the treatment indicator variable: To implement the test, the analyst specifies the outcome variable, the set of covariates being tested for effects heterogeneity, and the treatment indicator variable. After running the test_condate program, the analyst obtains the test statistic T labeled "Chi-Sq Test," the degree-of-freedom K labeled "dof Chisq," and the observed p value of the chi-square labeled "p-val Chi-sq" from the output. All three quantities are shown under the column heading of "Zero_Cond_ATE"—that is, they are the results for testing the first pair of hypotheses H_0 and H_a . A p value such as $p < .05$ suggests that the null hypothesis H_0 can be rejected at a statistically significant level. That is, a significant chi-square value indicates that the treatment effect is nonzero for subgroups in the sample, and the unconfoundedness assumption is probably violated. For the parametric model, the test statistic T^* for the second pair of also has a chi-square distribution with $K - 1$ degrees of hypotheses, and freedom, where K is the number of covariates being tested, including the treatment indicator variable: After running the test_condate program, the analyst obtains three statistics: T^* labeled "Chi-Sq Test," degree-of-freedom or $K - 1$ labeled "dof Chi-sq," and the observed p value of the chi-square labeled "p-val Chi-sq" under the 95 column heading of "Const_Cond_ATE." These are test results for the second A p value such as $p < .05$ suggests that the null pair of hypotheses and hypothesis can be rejected at a statistically significant level. When a significant chi-square is observed, the analyst can reject the hypothesis of a constant treatment effect across subgroups defined by covariates and conclude that the treatment effect varies across subgroups. This suggests that treatment effect heterogeneity exists. Perhaps the most important contribution made by Crump et al. (2008) is the extension of the tests from a parametric to nonparametric setting. Crump et al. developed equivalent tests by applying the series estimator of regression function and provided theorems with proofs. The development of this procedure employs sieve methods (Chen, Hong, & Tarozzi, 2008; Imbens, Newey, & Ridder, 2006). It is for this reason that Crump et al. refer to their tests as "nonparametric tests for treatment effect heterogeneity," although the two tests using chi-square are really parametric rather than nonparametric. Crump et al. show that in large samples, the test statistic of the nonparametric version has a standard normal distribution. Both T for the first pair of hypotheses, H_0 and H_a , and T^* for the second pair of hypotheses, and are distributed as The output of test_condate shows two types of quantities: the test statistic T (or T^*) labeled "Norm Test" and the observed p value of T (or T^*) labeled "pval Norm." Like the output for the parametric tests, both quantities are shown in two columns: One is under the column heading of "Zero_Cond_ATE," which shows the results for testing the first pair of hypotheses, H_0 and H_a , and the second is under the column heading of "Const_Cond_ATE," which shows the results for testing the second pair of hypotheses, and If the p value of T or T^* is less than .05 ($p < .05$), the analyst can reject the null hypothesis at a statistically significant level; otherwise, the analyst fails to reject the null hypothesis. With the nonparametric tests, the analyst may conclude that the treatment effect is nonzero for subgroups in the sample and that the unconfoundedness assumption is not plausible, if the "p-val Norm" under "Zero_Cond_ATE" is less than .05 ($p < .05$); the analyst may conclude that the treatment effect varies by subgroup and treatment effect heterogeneity exists if the "p-val Norm" under "Const_Cond_ATE" is less than .05 ($p < .05$). The output of the test_condate program also presents results of a test of the zero average treatment effect under the column heading of "Zero_ATE" for comparison purposes. This is the test commonly used to estimate the average treatment effect and its standard error. This is typically the main effect used in analysis, and it does not explicitly test or model treatment effect heterogeneity. The crucial message conveyed by the comparison is that the test of the zero ATE may show a nonsignificant p value, but the tests of treatment effect heterogeneity 96 could still be statistically significant. If this is observed, effect heterogeneity exists even when the main treatment effect is not statistically significant. 2.8.5 Example We now present a study investigating intergenerational dependence on welfare and its relation to child academic achievement. The data for this study are used in several examples throughout this book. Conceptual issues and substantive interests. As described in Chapter 1, prior research has shown that both childhood poverty and childhood welfare dependency have an impact on child development. In general, growing up in poverty adversely affects life course outcomes, and the consequences become more severe by length of poverty exposure (P. K. Smith & Yeung, 1998). Duncan et al. (1998) found that family economic conditions in early childhood had the greatest impact on achievement, especially among children in families with low incomes. Foster and Furstenberg (1998, 1999) found that the most disadvantaged children tended to live in female-headed households with low incomes, receive public assistance, and/or have unemployed heads of

household. In their study relating patterns of childhood poverty to children's IQs and behavioral problems, Duncan, Brooks-Gunn, and Klebanov (1994) found that the duration of economic deprivation was a significant predictor of both outcomes. Focusing on the effects of the timing, depth, and length of poverty on children, Brooks-Gunn and Duncan's study (1997) reported that family income has selective but significant effects on the well-being of children and adolescents, with greater impacts on ability and achievement than on emotional development. In addition, Brooks-Gunn and Duncan found that poverty had a far greater influence on child development if children experienced poverty during early childhood. The literature clearly indicates a link between intergenerational welfare dependence and child developmental outcomes. From the perspective of a resources model (see, e.g., Wolock & Horowitz, 1981), this link is repetitive and leads to a maladaptive cycle that traps generations in poverty. Children born to families with intergenerational dependence on welfare may lack sufficient resources to achieve academic goals, which will ultimately affects employability and the risk for using public assistance in adulthood. Corcoran and Adams (1997) developed four models to explain poverty persistence across generations: (1) The lack of economic resources hinders human capital development; (2) parents' noneconomic resources, which are correlated with their level of poverty, determine children's poverty as adults; (3) the welfare system itself produces a culture of poverty shared by parents and children; and (4) structural-environmental factors associated with labor market conditions, demographic changes, and racial discrimination shape 97 intergenerational poverty. Corcoran and Adams's findings support all these models to some extent, with the strongest supports established for the economic resources argument. Prior research on poverty and its impact on child development has shed light on the risk mechanisms linking resources and child well-being. Some of these findings have shaped the formation of welfare reform policies, some have fueled the ongoing debate about the impacts of welfare reform, and still other findings remain controversial. There are two major methodological limitations in this literature. First, prior research did not analyze a broad range of child outcomes (i.e., physical health, cognitive and emotional development, and academic achievement). Second, and more central to this example, prior research implicitly assumed a causal effect of poverty on children's academic achievement. However, most such studies used covariance control methods such as regression or regression-type models without explicit control for sample selection and confounding covariates. As we have shown earlier, studies using covariance control may fail to draw valid causal inferences. Throughout the book, we use different propensity score models to analyze the causal inference of poverty on child academic achievement. Data. This study uses the 1997 Child Development Supplement (CDS) to the Panel Study of Income Dynamics (PSID) and the core PSID annual data from 1968 to 1997 (Hofferth et al., 2001). The core PSID comprises a nationally representative sample of families. In 1997, the Survey Research Center at the University of Michigan collected information on 3,586 children between the ages of birth and 12 years who resided in 2,394 PSID families. Information was collected from parents, teachers, and the children themselves. The objective was to provide researchers with comprehensive and nationally representative data about the effects of maternal employment patterns, changes in family structure, and poverty on child health and development. The CDS sample contained data on academic achievement for 2,228 children associated with 1,602 primary caregivers. To address the research question about intergenerational dependence on welfare, we analyze a subset of this sample. Children included in the study were those who had valid data on receipt of welfare programs in childhood (e.g., AFDC [Aid to Families With Dependent Children]) and whose caregivers were 36 years or younger in 1997. The study involved a careful examination of 30 years of data using the 1968 PSID ID number of primary caregivers as a key. Due to limited information, the study could not distinguish between the types of assistance programs. The study criteria defined a child as a recipient of public assistance (e.g., AFDC) in a particular year if his or her caregiver ever received the AFDC program in that year and defined a caregiver as a recipient of AFDC in a particular year if the caregiver's primary caregiver (or the study child's grandparent) ever received the program in that year. The definition of receipt of AFDC in a year cannot 98 disentangle short-term use (e.g., receipt of AFDC for only a single month) from long-term use (e.g., all 12 months). One limitation of the study is posed by the discrete nature of AFDC data and the fact that the AFDC study variable (i.e., "caregiver's number of years using AFDC in childhood") was treated as a continuous variable in the analysis, which may not accurately measure the true influence of AFDC. After screening the data, applying the inclusion criteria, and deleting missing data listwise, the study sample comprised 1,003 children associated with 708 caregivers. Tests for treatment effect heterogeneity . Table 2.2 shows descriptive statistics of the study sample. For this illustration, we report findings that examine one domain of academic achievement: the age-normed "letter-word identification" score of the Woodcock-Johnson Revised Tests of Achievement (Hofferth et al., 2001). A high score on this measure indicates high achievement. The score is defined as the outcome variable for this study. The "treatment" in this study is child AFDC use from birth to current age in 1997. Of 1,003 study children, 729 never used AFDC or "untreated," and 274 used AFDC or "treated." The six covariates are major control variables observed from the PSID and CDS surveys. Table 2.2 Descriptive Statistics of the Study Sample Table 2.3 shows findings for the tests of treatment effect heterogeneity. Results suggest that we can reject the null hypothesis of a zero conditional average treatment effect using the parametric test ($\chi^2(df = 7) = 24.55$, $p < .001$) and the nonparametric test (test statistic following a normal distribution = 4.69, $p < .000$). The results confirm that the unconfoundedness assumption in this data set is not plausible, and corrective approaches to control for selection bias are needed if we want to draw a causal inference that is more rigorous and valid. 99 The results also suggest that there are some values of the covariates for which the treatment effect differs from zero. With regard to the tests regarding a constant conditional average treatment effect, we find that both tests show a nonsignificant p value (i.e., $p = .0844$ from the parametric test and $p = .0692$ from the nonparametric test). With these findings, we fail to reject the null hypothesis, and hence, we cannot confirm that treatment effect heterogeneity exists in this sample. The test of a zero average treatment effect shows that the main treatment variable is statistically significant ($p < .000$), meaning that AFDC has a nonzero impact on child academic achievement. This commonly used test shows the main effect of treatment. It does not tell us whether AFDC use affects child academic achievement differentially or whether treatment effect heterogeneity exists. As such, it does not reflect the whole range of complexity of treatment effects. 2.9 HECKMAN'S ECONOMETRIC MODEL OF CAUSALITY I n Chapter 1, we described two traditions in drawing causal inferences: the econometric tradition that relies on structural equation modeling and the statistical tradition that relies on randomized experiment. The economist James Heckman (2005) developed a conceptual framework for causal inference that he called the scientific model of causality. In this work, Heckman sharply contrasted his model with the statistical approach—primarily the NeymanRubin counterfactual model—and advocated for an econometric approach that directly models the selection process. Heckman argued that the statistical literature on causal inferences was incomplete because it had not attempted to model the structure or process by which participants are selected into treatments. Heckman further argued that the statistical literature confused the task of identifying causal models from population distributions (where the sampling variability of empirical distributions is irrelevant) with the task of identifying causal models from actual data (where sampling variability is an issue). Because this model has stimulated a rich debate, we highlight its main features in this section. The brevity of our presentation is necessitated by the fact that the model is a comprehensive framework and includes forecasting the impact of interventions in new environments, a topic that exceeds the scope of this book. We concentrate on Heckman's critique of the Neyman-Rubin model, which is a focal point of this chapter. Table 2.3 Tests for Treatment Effect Heterogeneity 100 Source: Data from Hofferth et al., 2001. First, Heckman (2005, pp. 9–21) developed a notation system for his scientific model that explicitly encompassed variables and functions that were not defined or treated comprehensively in prior literature. In this system, Heckman defined outcomes for persons in a universe of individuals and corresponding to possible treatments within a set of treatments where assignment is subject to certain rules; the valuation associated with each possible treatment outcome, including both private evaluations based on personal utility and evaluations by others (e.g., the "social planner"); and the selection mechanism appropriate under alternative policy conditions. Using this notation system and assumptions, Heckman further defined both individual-level treatment (causal) effects and population-level treatment effects. Second, Heckman (2005, p. 3) specified three distinct tasks in the analysis of causal models: (1) defining the set of hypotheticals or counterfactuals, which requires a scientific theory; (2) identifying parameters (causal or otherwise) from hypothetical population data, which requires mathematical analysis of point or set identification; and (3) identifying parameters from real data, which requires estimation and testing theory. Third, Heckman (2005, pp. 7–9) distinguished three broad classes of policy evaluation questions: (1) evaluating the impact of previous interventions on outcomes, including their impact in terms of general welfare (i.e., a problem of internal validity); (2) forecasting the impacts (constructing counterfactual states) of interventions implemented in one environment on other environments, including their impacts in terms of general welfare (i.e., a problem of external validity); and (3) forecasting the impacts of interventions (constructing counterfactual states associated with interventions) never historically experienced for other environments, including impacts in terms of general welfare (i.e., using history to forecast the consequences of new policies). Fourth, Heckman (2005, pp. 35–38) contrasted his scientific model (hereafter denoted as H) with the Neyman-Rubin model (hereafter denoted as NR) in terms of six basic assumptions. Specifically, NR assumes (1) a set of counterfactuals defined for ex post outcomes (no evaluations of outcomes or specification of treatment selection rules); (2) no social interactions; (3) invariance of counterfactual to assignment of treatment; (4) evaluating the impact of historical interventions on outcomes, including their impact in terms of welfare is the only problem of interest; (5) mean causal effects are the only objects of interest; and (6) there is no simultaneity in causal effects, that is, outcomes cannot cause each other reciprocally. In contrast, H (1) decomposes outcomes under competing states (policies or treatments) into their determinants; (2) considers valuation of outcomes as an essential ingredient of any study of causal inference; (3) models the choice of treatment and uses choice data to infer subjective valuations of treatment; (4) uses the relationship between outcomes and treatment choice equations to motivate, justify, and interpret alternative identifying strategies; (5) explicitly accounts for the arrival of information through ex ante and ex post analyses; (6) considers distributional causal parameters as well as mean effects; (7) addresses all three policy evaluation problems; and (8) allows for nonrecursive (simultaneous) causal models. The comparison of the NR and H models is summarized and extended in Table 2.4. Finally, Heckman (2005, pp. 50–85) discussed the identification problem and various estimators to evaluate different types of treatment effects. In Section 2.7, we have highlighted the main effects of interest that are commonly found in the literature (i.e., ATE, TT, TUT, MTE, and LATE). Heckman carefully weighed the implicit assumptions underlying four widely used methods of causal inference applied to data in the evaluation of these effects: matching, control functions, the instrumental variable method, and the method of directed acyclic graphs (i.e., Pearl, 2000). The scientific model of causality has clearly influenced the field of program evaluation. Perhaps the most important contribution of the model is its comprehensive investigation of the estimation problem, effects of interest, and estimation methods under a general framework. This is pioneering. Although it is too early to make judgments about the model's strengths and limitations, it is stimulating widespread discussion, debate, and methodological innovation. To conclude, we cite Sobel's (2005) comment that, to a great extent, coincides with our opinion: Table 2.4 Econometric Versus Statistical Causal Models 102 Source: Heckman, J. J. (2005). The scientific model of causality. *Sociological Methodology*, 35, p. 87. Heckman argues for the use of an approach to causal inference in which structural models play a central role. It is worth remembering that these models are often powerful in part because they make strong assumptions. . . . But I do not want to argue that structural modeling is not useful, nor do I want to suggest that methodologists should bear complete responsibilities for the use of the tools they have fashioned. To my mind, both structural modeling and approaches that feature weaker assumptions have their place, and in some circumstances, one will be more appropriate than the other. Which approach is more reasonable in a particular case will often depend on the feasibility of conducting a randomized study, what we can actually say about the reasonableness of invoking various assumptions, as well as the question facing the investigator (which might be dictated by a third party, such as a policy maker). An investigator's tastes and preferences may also come into play. A cautious and risk-averse investigator may care primarily about being right, even if this limits the conclusions he or she draws, whereas another investigator who wants (or is required) to address a bigger question may have (or need to have) a greater tolerance for uncertainty about the validity of his or her conclusions. (pp. 127–128) 2.10 CONCLUSION This chapter examined the Neyman-Rubin counterfactual framework, the ignorability treatment assignment assumption, the SUTVA assumption, the underlying logic of statistical inference, treatment effect heterogeneity and its 103 tests, and the econometric model of causality. We began with an overview of the counterfactual perspective that serves as a conceptual tool for the evaluation of treatment effects, and we ended with a brief review of Heckman's comprehensive and controversial scientific model of causal inference. It is obvious that there are disagreements among research scholars. In particular, debate between the econometric and statistical traditions continues to play a central role in the development of estimation methods. Specifically, we have emphasized the importance of disentangling treatment effects from treatment assignment and evaluating different treatment effects suitable to evaluation objectives under competing assumptions. Although the unconfoundedness assumption is untestable and the classic test of endogeneity is not helpful in the context of observational studies, new nonparametric tests of treatment effect heterogeneity are useful. They offer a convenient test for gauging the heterogeneity of treatment effects and evaluating the plausibility of the unconfoundedness assumption. We will revisit these issues throughout the book. NOTES 1. In the literature, there are notation differences in expressing this and other models. To avoid confusion, we use consistent notation in the text and present the original notation in footnotes. Equation 2.1 was expressed by Heckman and Vytlacil (1999, p. 4730) as 2. In Winship and Morgan's (1999, p. 665) notation, Equation 2.4 is expressed as 3. In Winship and Morgan's (1999) notation, Equation 2.5 is expressed as 4. Holland (1986) provides a thorough review of these statisticians' work under the context of randomized experiment. 5. In Rosenbaum's (2002b, p. 41) notation, Equation 2.6 is expressed as $R_{si} = Z_{si}\tau_i - (1 - Z_{si})\tau_{ci}$. 6. We have changed notation to make the presentation of SUTVA consistent with the notation system adopted in this chapter. In Rubin's original 104 presentation, he used u in place of i and t in place of w . 105 CHAPTER 3 Conventional Methods for Data Balancing As preparation for understanding advances in data balancing, this chapter reviews conventional approaches to the analysis of observational data. In Chapter 2, we examined the Neyman-Rubin counterfactual framework and its associated assumptions. Although the assumption of "no social interactions" is often deemed too strong, researchers generally agree that the ignorability treatment assignment assumption is necessary for program evaluations, including evaluations using observational data. Given the importance of this assumption in all emerging approaches, this chapter scrutinizes the mechanisms that produce violations of ignorability treatment assignment as well as conventional corrective approaches. We do so by creating data under five scenarios, and we show how to balance data by using three common methods, namely, ordinary least squares (OLS) regression, matching, and stratification. All three methods provide for control of covariates and may lead to unbiased estimation of treatment effects. All three methods are seemingly different but essentially aim to accomplish common objectives. All three also have limitations in handling selection bias. Understanding these methods is helpful in developing an understanding of advanced models. The key message this chapter conveys is that covariance control does not necessarily correct for nonignorability treatment assignment, and it is for this reason that other methods for estimating treatment effects should be considered in practice. Section 3.1 presents a heuristic example to address the question of why data balancing is necessary. Section 3.2 presents the three correction models (i.e., OLS regression, matching, and stratification). Section 3.3 describes the procedure for data simulation used in this chapter, particularly the five scenarios under which the ignorability treatment assignment assumption is violated to varying degrees. Section 3.4 shows the biases related to each method for estimating treatment effects under each of the five scenarios (i.e., the results of data simulation). Section 3.5 summarizes the implications of the data simulation. Section 3.6 is a succinct review of important aspects of running OLS regression, including important assumptions embedded in the OLS model, and a review of Berk's (2004) work on the pitfalls in running regression. Section 3.7.106 concludes with a summary of key points. 3.1 WHY IS DATA BALANCING NECESSARY? A HEURISTIC EXAMPLE To illustrate the importance of controlling for covariates in observational studies—which is analogous to balancing data—we repeat the famous example first published by Cochran (1968) and repeatedly cited by others. Rubin (1997) also used this example to show the value of data balancing and the utility of stratification. The importance of controlling for covariates is underscored by Table 3.1, which presents a comparison of mortality rates for two groups of smokers (i.e., cigarette smokers and cigar and pipe smokers) and one nonsmoking group that were measured in three countries: Canada, the United Kingdom, and the United States. Rubin (1997) argued that analyses of these data failed to control for age—a crucial covariate of mortality rate—and that, as a result, the observed data appear to show that cigarette smoking is good for health, especially relative to cigar and pipe smoking. For example, the Canadian data showed that the cigar and pipe smokers had a 35.5% mortality rate, whereas the mortality rates for the cigarette smokers and nonsmokers were not only similar but also much lower than those of the cigar and pipe smokers (20.5% and 20.2%, respectively). The pattern of mortality was consistent in the data from the other two countries. However, this finding is contradictory to our knowledge about the adverse consequences of smoking. Why is this the case? The primary reason is that age is an important confounding variable that affects the outcome of mortality rate. Note that the cigar and pipe smokers in Canada had the oldest average age (i.e., 65.9 years), while the average age of the nonsmokers was in the middle range for the three groups (i.e., 54.9 years), and the cigarette smokers had the youngest average age (i.e., 50.5 years). As presented, the unadjusted mortality rates are known as crude death rates because they do not take into account the age distribution of the three groups. To conduct a meaningful evaluation, we must balance the data to control for covariance. Balancing implies data manipulation that, in this case, would render trivial the confounding effect of age. Balancing raises the question of what the mortality rates will look like if we force all three groups to follow the same age distribution. To address the balancing problem, both Cochran (1968) and Rubin (1997) used a method called stratification, which is synonymous with subclassification. The lower panel of Table 3.1 shows the adjusted mortality rates for all three groups after applying three schemes of stratification. Cochran tested three schemes by stratifying the sample into 2 subclasses, 3 subclasses, and 9 to 11 subclasses. All three subclassifications successfully removed estimation bias, under which the apparent advantages of cigarette smoking disappear, and nonsmokers are shown as the healthiest group. 107 The stratification method used by Cochran is one of the three methods that are the focus of this chapter, and we elaborate on stratification in subsequent sections. The key message conveyed by this example is that estimation bias can be substantial if covariates in data from observational studies are uncontrolled. There are a number of other methods available to accomplish the same objective of controlling for covariates. One popular method in demography is age standardization. The core idea of age standardization is simple: Choose one age distribution from the three groups and, for each age group, multiply the proportion of persons in the group from the standard population to each agespecific death rate, then sum up all products. The resulting number is the adjusted mortality rate or standardized mortality rate for which age is no longer a confounding variable. To illustrate this concept, we created an artificial data set (Table 3.2) that simulates the same problem as found in Table 3.1. Note that in Table 3.2, the unadjusted mortality rate is the highest for cigar and pipe smokers, while the groups of cigarette smokers and nonsmokers appear to have the same rate. In addition, note that the simulated data follow the same pattern of age distribution as in the Table 3.1 data, where the cigar and pipe smokers are the oldest (i.e., 25% of this group were 61 years or older), while the cigarette smokers are the youngest (i.e., 8.33% of this group were 61 years or older). Table 3.1 Comparison of Mortality Rates for Three Smoking Groups in Three Databases Source: Cochran (1968, Tables 1–3) and Rubin (1997, p. 758). Table 3.2 Artificial Data of Mortality Rates for Three Smoking Groups 108 To conduct age standardization, we can choose the age distribution from any one of the three groups as a standard. Even when a different age is chosen—that is, a different standard is selected—the results will be the same in terms of the impact of smoking on mortality rate. In our illustration, we selected the age distribution of the cigarette smokers as the standard. Table 3.3 shows the results of the standardization. Using the age distribution of the cigarette smokers as the standard, we obtained three adjusted mortality rates (Table 3.3). The adjusted mortality rate of the cigarette smokers (i.e., 27.17 per 1,000) is the same as the unadjusted rate because both rates were based on the same age distribution. The adjusted mortality rate of the cigar and pipe smokers (or, more formally, "the adjusted mortality rate of cigar and pipe smokers with the cigarette smokers' age distribution as standard") is 26.06 per 1,000, which is much lower than the unadjusted rate of 38.5 per 1,000, and the adjusted mortality rate of the nonsmokers is 22.59, which is also lower than the unadjusted rate of 27.5 per 1,000. Simple corrective methods such as covariate standardization often work well, if the researcher is confident that the sources of confounding are known and that all confounding variables are observed (i.e., measured). When confounding is overtly understood and measured, it is often called overt bias (Rosenbaum, 2002b). Overt bias may be removed by simple approaches. The following section describes three methods that are equally straightforward and work well when sources of bias are understood and observed in measurement. 3.2 THREE METHODS FOR DATA BALANCING This section formally describes three conventional methods that help balance data—that is, OLS regression, matching, and stratification. Our interest centers on the key questions of how each method operates to balance data and to what extent each method accomplishes this goal. 3.2.1 The Ordinary Least Squares Regression 109 The material presented in this section is not new and can be found in most textbooks on regression. Let $Y = \alpha + \beta X + e$ represent a population regression model, where Y is an $(n \times 1)$ vector of the dependent variable for the n participants, X is an $(n \times p)$ matrix containing a unit column (i.e., all elements in the column take value 1) and $p - 1$ independent variables, e is an $(n \times 1)$ vector of the error term, and β is a $(p \times 1)$ vector of regression parameters containing one intercept and $p - 1$ slopes. Assuming repeated sampling and fixed X , and $e \sim iid, N(0, \sigma^2ln)$, where ln is an $(n \times n)$ identity matrix and σ^2 is a scalar, so that $\sigma^2ln = E(ee')$ is the variance-covariance matrix of the error term. With the observed data of Y and X , we can use the least squares criterion to choose the estimate of the parameter vector β that makes the sum of the squared errors of the error vector e a minimum, that is, we minimize the quadratic form of the error vector: Table 3.3 Adjusted Mortality Rates Using the Age Standardization Method (i.e., Adjustment Based on the Cigarette Smokers' Age Distribution) Taking the partial derivative of l with respect to β , and letting the partial derivative be zero, we obtain the optimizing vector β , $\beta = (X'X)^{-1}X'Y$. If we have sample data and use lowercase letters to represent sample variables and statistics, we have the sample estimated vector of regression coefficients as b , $s2\{b\}$ of an order $(p \times p)$, can be obtained by the following matrix: where $MSE = SSE/(n - p)$, and $SSE = \sum (y_i - \hat{y}_i)^2$, in which i is an identity matrix with the order of $(n \times n)$, and $H(n \times n) = x(x'x)^{-1}x'$. Taking the square root of each estimated variance, the researcher obtains standard errors (se) of the estimated regression coefficients, as $With$ estimated regression coefficients and corresponding standard errors, we can perform statistical significance tests or estimate the confidence intervals of the coefficients as follows: 1. Two-tailed test (when hypothetical direction of β is unknown): $H_0: \beta_1 = 0$ (meaning: X_1 has no impact on Y), $H_a: \beta_1 \neq 0$ (meaning: X_1 has an impact on Y). Calculate $t^* = b_1/se\{b_1\}$. If $|t^*| \leq (t_{1-\alpha/2; n-p})$, conclude H_0 ; if $|t^*| > (t_{1-\alpha/2; n-p})$, conclude H_a . 2. One-tailed test (when one can assume a direction for β based on theory): $H_0: \beta_1 \leq 0$ (meaning: X_1 has either a negative or no impact on Y), $H_a: \beta_1 > 0$ (meaning: X_1 has a positive impact on Y). Calculate $t^* = b_1/se\{b_1\}$. If $|t^*| \leq (t_{1-\alpha; n-p})$, conclude H_0 ; if $|t^*| > (t_{1-\alpha; n-p})$, conclude H_a . 3. Confidence interval of b : $(1 - \alpha) \times 100\%$ confidence interval can be computed as follows: The Gauss-Markov theorem reveals the best linear unbiased estimator or BLUE property of the OLS regression. The theorem states that given the assumptions of the classical linear regression model, the least squares estimators, in the class of unbiased linear estimators, have minimum variance; 111 that is, they are BLUEs. Using this setup for the OLS regression estimator, the researcher can control for the covariates and balance data. We now rewrite the regression model $Y = \alpha + \beta X + e$ to $Y = \alpha + \tau W + X_1\beta + e$ by treating W as a separate variable and taking it out from the original matrix X . Here W is a dichotomous variable indicating treatment condition (i.e., $W = 1$, if treated, and $W = 0$ otherwise). Thus, X_1 now contains one variable less than X and contains no unit column. If the model is appropriately specified—that is, if the X_1 contains all variables affecting outcome Y (i.e., one or more of the threats to internal validity have been taken care of by a correct formulation of the X_1 matrix)—and if all variables take a correct functional form, then τ will be an unbiased and consistent estimate of the average treatment effect of the sample, or However, it is important to note that we can draw this conclusion only when all other assumptions of OLS regression are met, which is questionable under many conditions. For now, we will assume that all other assumptions of OLS regression have been met. Recall the interpretation of a regression coefficient: Other things being equal (or ceteris paribus), a one-unit increase in variable x_1 will decrease (or increase, depending on the sign of the coefficient) b_1 units in the dependent variable y . This is an appealing feature of OLS: By using least squares minimization, a single regression coefficient captures the net impact of an independent variable on the dependent variable. This mechanism is exactly how OLS regression controls for covariates and balances data. If the researcher successfully includes all covariates, and if the regression model meets other assumptions, then is an unbiased and consistent estimate of the average treatment effect. In sum, the key feature of OLS regression, or more precisely the mechanism for balancing data through regression, is to include important covariates in the regression equation and to ensure that the key assumptions embedded in the regression model are plausible. By doing so, the regression coefficient of the dichotomous treatment variable indicates the average treatment effect in the sample. Employing the regression approach requires making strong assumptions that are often violated in the real world. Key assumptions and issues related to regression analysis are examined in Section 3.6. 3.2.2 Matching Before the development of the new estimation methods for observational data, matching was the most frequently used conventional method for handling observational data. For instance, Rossi and Freeman (1989) described how to balance data by conducting ex post matching based on observed covariates. Matching can be performed with or without stratification (Rosenbaum, 2002b, p. 80). The central idea of the method is to match each treated participant ($x_{ij} | w_{ij} = 1$) to n nontreated participants ($x_j | w_j = 0$) on x , where x is a vector of matching variables (i.e., variables that covary with the treatment), and then compare the average of y of the treated participants with the average of y of the matched nontreated participants. The resultant difference is an estimate of the sample average treatment effect. Under this condition, the standard estimator may be rewritten as where the subscript "match" indicates the matched subsample. For $w_{match} = 1$, the group comprises all treated participants whose matches are found (i.e., the group excludes treated participants without matches), and for $w_{match} = 0$, the group is composed of all nontreated participants who were matched to treated participants. Denoting M as the original sample size, M_{match} as the matched sample size, $N_{w=1}$ the number of treated participants before matching, $N_{w=0}$ the number of untreated participants before matching, $N_{match,w=1}$ the number of treated participants after matching, and $N_{match,w=0}$ the number of untreated participants after matching, we have $N_{match,w=1} = N_{w=1} - N_{w=1}$ (because some treated participants cannot find a match in the nontreated group), and $M_{match} < M$ (because of the loss of both treated and untreated participants due to nonmatching). The following equality is also true: $N_{match,w=1} = n(N_{match,w=0})$ depending on how many matches (i.e., n) the researcher chooses to use. If the researcher chooses to match one participant from the untreated pool to each treated participant (i.e., $n = 1$), then $N_{match,w=1} = N_{match,w=0}$; if one chooses to match four participants from the untreated pool to each treated participant (i.e., $n = 4$), then $N_{match,w=1} = 4(N_{match,w=0})$, or $N_{match,w=0}$ is four times as large as $N_{match,w=1}$. The choice of n deserves scrutiny. If we choose $n = 1$ to perform a one-to-one match, then we lose variation in multiple matches who all match to a treated participant. If we choose a large n such as $n = 6$, then we may not find six matches for each treated participant. Abadie et al. (2004) developed a procedure that allows the researcher to specify a maximum number of matches. Abadie and Imbens's (2002) data simulation suggested that $n = 4$ (i.e., matching up to 4 untreated participants for each treated participant) usually worked well in terms of mean squared error. Rosenbaum (2002b) recommended the use of optimal matching to solve the problem. These issues are reviewed and extended in later chapters. In sum, the key feature of matching, or more precisely the mechanism for balancing data through matching, involves identifying untreated participants who are similar on covariates to treated participants and using the mean outcome of the nontreated group as a proxy to estimate the counterfactual of the treated 113 group. 3.2.3 Stratification Stratification is a procedure that groups participants into strata on the basis of a covariate x (Rosenbaum, 2002b). From the M participants, select $N < M$ participants and group them into S nonoverlapping strata with n_s participants in stratum s . In selecting the N units and assigning them to strata, use only the x variable and possibly using a table of random numbers to select units. A stratification formed in this way is called stratification on x . In addition, an exact stratification on x has strata that are homogeneous in x , so two participants are included in the same stratum only when both have the same value of x ; that is, $x_{si} = x_{sj}$ for all s, i, j . Exact stratification on x is practical only when x is of low dimensionality and its coordinates are discrete; otherwise, it will be difficult to locate many participants with the same x . For instance, exact stratification is feasible when we have a discrete age variable with three groups ($x = 1$, if ages 18 to 40 years; $x = 2$, if ages 41 to 60 years; and $x = 3$, if ages 61 years or older). For the purpose of balancing data, we introduce a stratification procedure based on the percentile of x , called quartile or quintile stratification, that is designed for situations where x is a continuous variable. The quintile stratification procedure involves the following five steps: 1. Sort data on x so that all participants are in an ascending order of x . 2. Choose participants whose values on x are equal to or less than the first quintile of variable x to form the first stratum (i.e., Stratum 1 contains all participants whose x values are equal to or less than the value of the 20th percentile of x); then choose participants whose values on x fall into the range bounded by the second quintile to form the second stratum (i.e., Stratum 2 contains all participants whose x is in a range between values of the 21st percentile and the 40th percentile of x); keep working in this fashion until the fifth stratum is formed. 3. For each stratum s ($s = 1, 2, \dots, 5$), perform the standard estimator to calculate the difference of mean outcome y between the treated and where $s = 1, 2, 3, 4, 5$. nontreated groups; that is, 4. Calculate the arithmetic mean of the five means to obtain the average treatment effect of the sample using the equation and calculate the variance of the estimated average treatment effect using the equation 114 5. Use these statistics to perform a statistical significance test to determine whether the sample average treatment effect is statistically significant. Quartile stratification is performed in a similar fashion, except the statistic quartile is used and four strata are created. There are more sophisticated methods involving stratification. For instance, in the example using data from the three smoking groups, Cochran's (1968) method for balancing the mortality rate data (see Table 3.1) was to choose the number of strata so that each stratum contained a reasonable number of participants from each treatment condition. This explains why there were three stratification schemes in Table 3.1 (i.e., 2 subclasses, 3 subclasses, and 9 to 11 subclasses). Cochran offered theoretical results showing that when the treatment and the control groups overlapped in their covariate distributions (i.e., age in his example), comparisons using five or six subclasses typically removed 90% or more of the bias presented in the raw comparisons. In sum, the key feature of stratification, or more precisely the mechanism for balancing data through stratification, is to make participants within a stratum as homogeneous as possible in terms of an observed covariate; then the mean outcome of the nontreated group is used as a proxy to estimate the counterfactual for the treated group. Exact stratification produces homogeneity only for discrete covariates. For instance, if gender is a covariate, stratification will produce two homogeneous strata: all females in one stratum and all males in another. However, this is not the case for a continuous covariate. With a continuous covariate x and using a quintile stratification, participants within stratum $s = 1$ are not exactly the same in terms of x . In these circumstances, we must assume that within-stratum differences on x are ignorable, and participants falling into the same stratum are "similar enough." The level of exactness on x can be improved by increasing the number of strata S , but the literature has suggested that $S = 5$ typically works well (Rosenbaum & Rubin, 1984, 1985; Rubin, 1997). 3.3 DESIGN OF THE DATA SIMULATION To help demonstrate the importance of data balancing and how conventional methods can be used to correct for bias under different settings of data generation, we designed a data simulation with a threefold purpose. First, we show how conventional correction methods work when selection bias is overt and is controlled properly. Second, we use the simulation to show conditions under which the conventional methods do not work. From this, we demonstrate the need for more sophisticated methods for balancing data. Third, we use these simulated data and their conditions as an organizing theme for the book. We 115 illustrate the reasons why new methods were developed, the specific problem each method was designed to resolve, and the contributions made by each method. Unlike using real data, the simulation approach creates artificial data based on designed scenarios. The advantage of using artificial data is that the true value of the treatment effect is known in advance. This allows us to evaluate bias directly. In addition, this simulation assumes that we are working with population data, so sampling variability is ignored. For now, we do not examine the sensitivity of estimated standard errors. However, we return to this issue in Chapter 11. The data generation is based on the following regression model: where x is a covariate or control variable, and w is a dummy variable indicating treatment conditions ($w = 1$ treated, and $w = 0$ control). The simulation creates data of y, x, w , and e according to the following known parameters: $\alpha = 10, \beta = 1.5$, and $\tau = 2$. That is, we created data using the following equation: We describe five scenarios that assume different relationships among y, x, w , and e . In each of the scenarios, the sample size is fixed at 400. Because the parameters are known, we can compare the model-estimated treatment effect with the true value $\tau = 2$ to evaluate bias for each scenario. The five scenarios are described in the following. Scenario 1: $w \perp x, x \perp e, x, w, e \sim iid, N(0, 1)$. This scenario assumes the following four conditions: (1) conditional on covariate x , the treatment variable w is independent of the error term e , (2) there is no correlation between covariate x and the error term e , (3) there is no correlation between covariate x and the treatment variable w , and (4) the error term is identically and independently distributed and follows a normal distribution with mean of zero and variance of 1. This is an ideal condition. The crucial assumption is $w \perp x$, which simulates the ignorable treatment assignment. In addition, this models the data generated from a randomized study, because these conditions are likely to be met only in data obtained from a randomized experiment. To ensure that the data generation strictly meets the assumption of $w \perp x$, we force x to be an ordinal variable taking values 1, 2, 3, and 4. The remaining scenarios simulate conditions that are not ideal and involve relaxing one or more of the assumptions included in Scenario 1. Scenario 2: $p_{w=0} \perp x, x \perp w$, and $e \sim iid, N(0, 1)$. In this scenario, the ignorable treatment assignment assumption (i.e., $w \perp x$, which is reflected by $w = 116 e\}$ is violated. The scenario indicates that conditional on covariate x , the correlation between the treatment variable w and the error term e is not equal to zero ($p_{w=0} \neq 0$). All other conditions of Scenario 2 are identical to those of Scenario 1. Thus, Scenario 2 simulates the contemporaneous correlation between the error term and the treatment variable. We expect that the estimated treatment effect will be biased and inconsistent. 1 Scenario 3: $p_{w=0} \perp x, w \perp e$, and $e \sim iid, N(0, 1)$. This scenario relaxes the Scenario 1 condition of independence between the covariate x and the treatment variable w . To simulate real situations, we also allow the covariate x to be a continuous variable to take values on more than four ordinal levels. All other conditions in Scenario 3 remain the same as those in Scenario 1. Scenario 3 simulates the condition of multicollinearity, which is a typical occurrence when using real data. In this scenario, the ignorable treatment assignment assumption still holds because x is the only source of the correlation with w , and x is used as a control variable. Thus, we can expect that the results will be unbiased, although it is interesting to see how different methods react to multicollinearity. Scenario 4: $p_{w=0} \perp x, p_{w=0} \neq 0, x \perp e$, and $e \sim iid, N(0, 1)$. This scenario differs from Scenario 1 by relaxing two of the conditions: (1) the independence between the covariate

x and the treatment variable w and (2) the independence between the treatment variable w and the error term e. Similar to Scenario 3, the covariate x is a continuous variable. All other conditions in Scenario 4 remain the same as those in Scenario 1. In addition, Scenario 4 differs from Scenario 3 in that it relaxes the assumption of independence between the treatment variable w and the error term e (i.e., it changes w e to pwe ≠ 0|x). Thus, under this scenario, the ignorable treatment assignment no longer holds, which makes Scenario 4 similar to Scenario 2. In addition, Scenario 4 simulates two data problems: (1) linear correlation between independent variables and (2) the nonignorable treatment assignment. We expect that the estimated treatment effect will be biased and inconsistent. Scenario 5: pxe ≠ 0, pwe ≠ 0, pwx ≠ 0, and e ~ iid, N(0, 1). Scenario 5 relaxes three of the conditions: (1) independence between the covariate x and the treatment variable w, (2) independence between the treatment variable w and the error term e, and (3) independence of the covariate x and the error term e. In addition, we allow the covariate x to be a continuous variable. This scenario is a further relaxation of assumptions embedded in Scenario 4 (i.e., it changes x e to pxe ≠ 0), and this is the worst-case scenario. Because the ignorable treatment assignment assumption is violated, and the problem of multicollinearity is present, we expect the estimated treatment effect will be biased and inconsistent. 117 The programming syntax for the data simulation is available at the companion webpage for this book. We employed the simple matching estimator (Abadie et al., 2004) to produce results. The algorithm by Abadie et al. (2004) is slightly different from the conventional matching estimator because it matches with replacement (see Chapter 8). Readers may replicate the analysis to verify the findings. 3.4 RESULTS OF THE DATA SIMULATION Table 3.4 presents descriptive statistics for the data and estimated treatment effects produced using the three methods under Scenario 1. The descriptive statistics show that conditions described by the design of Scenario 1 are met: that is, w is not correlated with e (pwe = .01), x is not correlated with e (pxe = .03), x is not correlated with w (pwx = .03), the mean of e is close to 0, and the standard deviation of e is close to 1. Figure 3.1 is the scatterplot of data x and y under Scenario 1. The results show that under the ideal conditions of Scenario 1, all three methods accurately estimate the treatment effect, and all biases are close to zero. Among the three methods, regression worked the best with a bias of .011, stratification worked second best with a bias of .016, and matching worked relatively the worst with a bias of .029. Note that although the three methods are technically different, each method estimated a treatment effect very close to the true parameter. The conditions of data generation under Scenario 1 are stringent. In reality, a well-designed and well-implemented randomized experiment is likely the only kind of study that could produce data in such an ideal fashion. Table 3.4 Data Description and Estimated Effects by Three Methods: Scenario 1 Figure 3.1 Scatterplot of Data Under Scenario 1 118 Table 3.5 presents descriptive statistics of the data and estimated treatment effects using three methods under Scenario 2. The descriptive statistics show that conditions described by Scenario 2 are met; that is, w is correlated with e (pwe = .61), x is not correlated with e (pxe = -.04), x is not correlated with w (pwx = .02), the mean of e is close to 0, and the standard deviation of e is close to 1. Figure 3.2 is the scatterplot of data x and y under Scenario 2. Under the conditions of Scenario 2, each of the three methods produced a biased estimate of the treatment effect. Specifically, all three methods estimated a treatment effect that was 1.13 units higher than the true value, or a 57% overestimation. In particular, the OLS result shows that the bias is in an upward direction (i.e., inflated treatment effect), when the treatment variable w is positively correlated with the error (i.e., pwe = .61). The issue of inflated treatment effect was discussed theoretically in Chapter 2, and we showed that when the error term and the independent variable are positively correlated, the OLS estimated slope is higher than the true slope (see Figure 2.1). This upward bias is exactly what happens when conditions are similar to those in Scenario 2. Table 3.5 Data Description and Estimated Effects by Three Methods: Scenario 2 119 Figure 3.2 Scatterplot of Data Under Scenario 2 Imagine a real study such as Scenario 2. In an observational data collection (where an independent variable of interest is not manipulated), suppose w is a state indicating addiction to illegal drugs, y is a measure of mental status (i.e., a high value on y indicates more psychological problems), and the data were obtained by surveying a representative sample of a population. Because the data are observational and exist naturally (i.e., representing a population) without randomization, participants who used illegal drugs were likely to have high values on y. Although x is an important covariate of y (i.e., pxy = .69), specifying only one of such covariates in the regression model is not sufficient. When we note that there is a high correlation between the treatment and the error term (i.e., pwe = .61), it becomes clear that additional covariates were omitted in the regression model. Under this condition, in addition to nonignorable treatment assignment, we encounter the problem of omitted covariates, which is also known as selection on unobservables (see Section 2.3 in Chapter 2). As a 120 result of the combination of these two problems, the estimated treatment effect (i.e., the net impact of addiction to illegal drugs on psychological problems) is biased upward. Although theory might tell us that abusing drugs causes mental problems, our data would overestimate the impact of drug abuse on mental status. This overestimation is true regardless of which one of the three analytic methods is used. It is for this reason that we can conclude that conventional correction methods are not appropriate (i.e., do not work well) in conditions such as those described for Scenario 2. Under this scenario, advanced analytic approaches must be considered. But these too are subject to bias as a result of unobserved heterogeneity, and it is advisable to assess the potential effect of hidden selection bias related to omitted variables by conducting sensitivity analyses, as discussed in Chapter 11. Table 3.6 presents descriptive statistics of the data and estimated treatment effects produced using three analytic methods under the conditions of Scenario 3. The descriptive statistics show that conditions described by Scenario 3 are met; that is, w is correlated with x (pwx = .56), x is not correlated with e (pxe = .07), w is not correlated with e (pwe = .03), the mean of e is close to 0, and the standard deviation of e is close to 1. Figure 3.3 is the scatterplot of data x and y under Scenario 3. Under Scenario 3, the ignorable treatment assignment still holds; therefore, the estimated treatment effect produced using the regression method is unbiased. Furthermore, the presence of multicollinearity does not affect the estimation of treatment effect, even though it affects the significance test; however, the significance test is irrelevant in the current discussion about the population parameters. These results confirm that among the three analytic methods, regression worked the best, with a bias of -.025 (i.e., a bias slightly downward); matching worked reasonably well, with a bias of .052; and stratification worked in a problematic way with a bias of .267 (or 13% larger than the true value). This example suggests that the methods react differently to multicollinearity and are not equally good in terms of bias correction. Table 3.7 presents descriptive statistics and estimated treatment effects using three methods under Scenario 4. The descriptive statistics show that conditions described by Scenario 4 are met; that is, w is correlated with e (pwe = .59), w is correlated with x (pwx = .55), x is not correlated with e (pxe = .06), the mean of e is close to 0, and the standard deviation of e is close to 1. Figure 3.4 is the scatterplot of data x and y under Scenario 4. When this scatterplot is compared with the scatterplot for Scenario 3, we see a more systematic concentration of data points, as all treated cases are clustered in the upper panel, and all untreated cases are clustered in the lower panel. This pattern is produced by the correlation between w and e. This scenario is more likely to occur in real applications when researchers think that they have controlled for covariates, but the available controls are not sufficient and/or relevant covariates are omitted. 121 Thus, we encounter the same problem of selection on unobservables as that in Scenario 2. These results show that all three methods are biased. Under the conditions of Scenario 4, the three methods are ranked for bias as follows: Regression produced the lowest bias of 1.647, matching produced higher bias at 1.866, and stratification produced the highest bias at 2.024. The implications are clear. When treatment assignment is not ignorable and when multicollinearity is present, no conventional method produces unbiased estimation of treatment effects. Table 3.8 presents descriptive statistics and estimated treatment effects using three methods under Scenario 5. The descriptive statistics show that conditions described by the design of Scenario 5 are met; that is, x is correlated with e (pxe = -.74), w is correlated with e (pwe = .56), w is correlated with x (pwx = .57), the mean of e is close to 0, and the standard deviation of e is close to 1. Figure 3.5 is the scatterplot of data x and y under Scenario 5. Note that Scenario 5 relaxes one assumption of Scenario 4 (i.e., it changes x e to pxe ≠ 0). Under this scenario, both w and x correlate with the error term, and the two independent variables w and x correlate with one another. Among all five scenarios, Scenario 5 assumes the weakest conditions for data generation, which would lead us to expect that the results would be the worst. Indeed, the estimated treatment effects produced with each method are biased, and the methods are ranked in terms of bias as follows: Regression produced the lowest bias (i.e., .416), matching produced the second lowest bias estimates (i.e., .492 for n = 1 and .747 for n = 4, respectively), and stratification produced the highest bias (i.e., .800). However, contrary to our expectation, all biases were lower than those produced by the methods under Scenario 4. Although Scenario 5 is the worst-case scenario, the results are not the most severe. This interesting finding indicates that data conditions often work in complicated ways, and results may not conform to our expectations. Overall, our findings underscore the importance of assumptions, the risk of biased parameter estimation when conventional corrective methods are used, and the need to develop models with greater sophistication. Table 3.6 Data Description and Estimated Effects by Three Methods: Scenario 3 122 Figure 3.3 Scatterplot of Data Under Scenario 3 3.5 IMPLICATIONS OF THE DATA SIMULATION Our data simulation, or a three-methods-and-five-scenarios design, illuminates the conditions under which common data balancing strategies correct selection bias. We purposely created challenging data environments, but they are not dissimilar to those encountered in routine program evaluation. Contending with such challenges has motivated statisticians and econometricians to seek new methods. The data simulation has at least three implications in developing these methods. Table 3.7 Data Description and Estimated Effects by Three Methods: Scenario 4 123 Figure 3.4 Scatterplot of Data Under Scenario 4 First, simple methods of data balancing work well only under ideal conditions (i.e., Scenario 1), and under ideal conditions, all three methods work equally well. When the ignorable treatment assignment assumption holds but independent variables are correlated (i.e., Scenario 3), the regression and matching methods provide unbiased estimation of treatment effect. However, estimation produced with the stratification method is problematic. When the ignorable treatment assignment assumption is violated (i.e., Scenarios 2, 4, and 5), none of the three conventional correction methods provides an unbiased estimation of the treatment effect. Second, covariance control does not automatically correct for nonignorable treatment assignment. In all five scenarios, x is highly correlated with y (i.e., pxy ranges from .69 to .92), indicating that x is an important covariate of y in all scenarios. However, there are only two scenarios under which the ignorable treatment assignment assumption holds (i.e., Scenarios 1 and 3). The data 124 simulation clearly shows that only under these two scenarios did three conventional methods produce unbiased estimates. Under all other scenarios, the three methods failed. Thus, common methods controlling for covariance do not necessarily correct for nonignorable treatment assignment. Finally, the findings suggest that we must understand data generation before running regressions and other statistical models. As Berk (2004) noted, data generation is important for inference drawn from analyses: Table 3.8 Data Description and Estimated Effects by Three Methods: Scenario 5 Figure 3.5 Scatterplot of Data Under Scenario 5 [Both statistical inference and causal inference] depend fundamentally on how the data used in the regression analysis were generated: How did 125 nature and the investigator together produce the data then subjected to a regression analysis? Was there a real intervention? Can the data be properly viewed as a random sample from a real population? How were key variables measured? There is precious little in the data themselves that bear on such matters. One must bring to the regression analysis an enormous amount of information about data, and this information will, in practice, do some very heavy lifting. (pp. 1–2) In addition, the three-methods-and-five-scenarios design offers a useful tool for thinking about the key features of new approaches for evaluation. We use this design as an organizing theme to assess the methods described in this book. 1. In the data simulation, we only have one control variable, x, whereas in practice, there are typically many control variables. When the number of control variables increases, the conventional ex post matching often fails because, under such conditions, it is difficult to match treated cases to untreated cases on multiple characteristics: This problem is known as the dimensionality of matching. A simple example helps illustrate this problem. If you use only three demographic variables (i.e., age prior to receiving treatment, gender, and race/ethnicity), you may be able to match 90% of the treated participants to the untreated participants, assuming that the sample is of a reasonably large size (e.g., N = 500) and the two groups have equal size. Even with three matching variables, chances are that some cases would be lost because they are so different that no match would be available. Depending on the nature of the sample and, of course, depending on how accurate you want matching on age to be (i.e., Should matches be exact with, say, 25.09-year-olds, or could the age be matched within an age range, such as matching participants who fall within 20 to 30 years of age?), 90% is merely a guess for the percentage of possible matches. Now suppose you add one matching variable, which is a depression scale ranging from 40 to 100. With this addition, it is likely that you will match still fewer treated participants to the untreated participants. With every added dimension, the percentage of successful matches declines. This dimensionality problem motivated researchers to develop several new approaches. These include (a) a two-step estimator that uses probabilities of receiving treatment (Heckman, 1978, 1979; Maddala, 1983), (b) propensity score matching (Rosenbaum & Rubin, 1983), and (c) the use of vector norms (Abadie & Imbens, 2002, 2006; Rubin, 1980a). These approaches to the problem of dimensionality in matching are examined in Chapters 4, 5, and 8, respectively. 2. Propensity score matching may be thought of as a slightly more complex method that combines the two conventional methods of regression and matching. That is, the analyst first creates propensity scores for all study participants, such that multiple characteristics are reduced to a one-dimensional score. The analyst 126 then matches the scores between treated and nontreated cases to create a new sample. Last, the analyst performs a secondary analysis, such as regression, on the matched sample. Note that in the second stage, many kinds of multivariate analysis may be performed (e.g., regression-type models such as the random coefficients model, multiple-group structural equation modeling, survival analysis, generalized linear models). This method of propensity score matching is described in Chapter 5. 3. Propensity score matching can also be thought of as a combination of matching and stratification. When multiple matching variables are reduced to a one-dimensional propensity score, the analyst can match treated cases to nontreated cases on the one-dimensional score and then use stratification or subclassification to estimate treatment effect. As such, regression or other types of second-stage modeling are not used (Rosenbaum & Rubin, 1984). Propensity score matching in conjunction with subclassification is described in Chapter 6. 4. Propensity scores can be used to form subclasses without matching. Using quintiles of propensity scores, the analyst can divide the sample into five equally sized subclasses, perform a stratified multivariate analysis for each stratum, and then aggregate the treatment effects across all five strata to discern whether the treatment effect for the entire sample is statistically significant. The primary advantage of this method is its capacity to analyze virtually any type of outcome measure (e.g., categorical, ordinal, or time-to-event), and the capacity to use more complicated methods, such as structural equation modeling, in within-stratum analytics. This, of course, is conditioned on a sufficiently large sample within quintiles. Propensity score subclassification is described in Chapter 6. 5. Propensity scores can be used as sampling weights and analyzed in an outcome model that takes differential information from each study subject, depending on the subject's conditional probability of falling in a treatment condition. Under this context, the analytic method can be viewed as an application of propensity scores to a regression or regression-type model. This approach shares the same advantages as those for propensity score subclassification—namely, it allows the analyst to model noncontinuous nonnormal outcome variables and perform complicated analysis such as structural equation modeling. Propensity score weighting is described in Chapter 7. 6. The analyst can also use a single method of matching to estimate the treatment effect (i.e., a method without using regression and stratification). Sophisticated approaches have been developed to improve matching as a method for estimation of the treatment effect. Abadie and Imbens (2002, 2006), 127 for example, developed matching estimators that use the vector norm (i.e., the Mahalanobis metric distance or the inverse of sample variance matrix) to estimate the average treatment effect for the treated group, the control group, and the sample, as well as similar effects for the population. The matching estimator method is described in Chapter 8. Heckman, Ichimura, and Todd (1997, 1998) developed an alternative approach for matching using nonparametric regression or, more precisely, local linear regression to match treated participants to untreated participants. This method provides a robust and efficient estimator of the average treatment effect for the treated group. Matching with local linear regression is described in Chapter 9. 7. In the data simulation, w was a dichotomous variable indicating treatment conditions: treated or nontreated. The propensity score method can easily be expanded to include multiple treatment conditions or to include the analysis of doses of treatment (Hirano & Imbens, 2004; Imbens, 2000; Rosenbaum, 2002b). This method of modeling treatment dosage is described in Chapter 10. 8. Rubin (1997) summarized three limitations of propensity score matching, one of which is that propensity score matching cannot control for unobserved selection bias. That is, propensity score matching cannot adjust for hidden selection bias, and therefore, there is no solution to Scenario 2. (We give further attention to this problem in Chapter 11, where we compare sampling properties of six models through a Monte Carlo study.) Although selection bias is hidden, the magnitude of bias may be estimated through sensitivity analysis, which was developed by Rosenbaum (2002b) and constitutes a seminal contribution to propensity score analysis. As mentioned earlier, the sensitivity analysis method is described in Chapter 11. 9. Identified by Rubin (1997), a second limitation of propensity score matching is that the method does not differentially handle covariates based on the relation to outcomes. That is, the method does not treat a covariate that is related to treatment assignment but not to outcome differently from a covariate that is related to both treatment assignment and outcome. This problem can be viewed as a variant of Scenario 4, for which there is no desirable solution currently available; however, James Robins's marginal structural modeling appears promising. We revisit this issue in Chapter 12. 10. Scenario 5 features a combination of several weak assumptions about data generation. Although the data generation is complicated, this scenario is realistic and likely to occur in practice. We revisit this scenario in Chapter 11 and underscore that—even with advanced methods—unobserved heterogeneity always holds the potential to bias effect estimations. 128 3.6 KEY ISSUES REGARDING THE APPLICATION OF OLS REGRESSION Among all statistical approaches, the OLS regression model is perhaps the most important because it not only serves as the foundation for advanced models but also is the key to understanding new approaches for the evaluation of treatment effects. We have seen conditions under which the OLS regression provides (or does not provide) unbiased estimation of treatment effects. To conclude the current chapter, we review important assumptions embedded in the OLS regression and fundamental issues with regard to the application of OLS regression. Our review follows Kennedy (2003) and Berk (2004). When applying the OLS regression model, five basic assumptions must be made about the data structure: 1. The dependent variable is a linear function of a specific set of independent variables plus a disturbance term. 2. The expected value of the disturbance term is zero. 3. The disturbances have a uniform variance and are uncorrelated with one another. 4. The observations on the independent variables are considered fixed in repeated samples. 5. The number of observations is greater than the number of independent variables, and no exact linear relationships exist between independent variables (Kennedy, 2003). These assumptions are crucial for consistent and unbiased estimation in regression models. In practice, it is important to understand the conditions under which these assumptions are violated, to be attentive to conducting diagnostic tests to detect violations, and to take remedial actions if violations are observed. From Berk (2004), regression may be used to draw causal inferences when four conditions are satisfied. First, "clear definitions of the relevant concepts and . . . good information about how the data were generated" are required (Berk, 2004, p. 101, used with permission). Causal effects involve not only the relationships between inputs and outputs, not only the relationships between observable and unobservable outcomes, but also a number of additional features related to how nature is supposed to operate. Berk recommended using Freedman's response schedules framework to infer causal effects. The available theories, including those used in economics, are mostly silent on how to characterize the role of the errors in the regression formulation. In contrast, Freedman's response schedule framework requires that errors be treated as 129 more than a nuisance to be dispatched by a few convenient assumptions. From this perspective, causal stories should include a plausible explanation of how the errors are generated in the natural world. Second, the ceteris paribus condition must be assumed. In warning that there are typically a host of ceteris paribus conditions that never exist in real life, Berk (2004) argued, "Covariance adjustments are only arithmetic manipulations of the data. If one decides to interpret those manipulations as if some confounder were actually being fixed, a very cooperative empirical world is required" (p. 115). Third, causal relationships must be anticipated in data generation and analysis. Berk (2004, p. 196) objected in part to Pearl's (2000) claim that the analyst can routinely carry out causal inference with regression analysis (or, more generally, structural equation modeling) of observational data. Last, Berk (2004) argued that credible causal inferences cannot be drawn from regression findings alone. The output from regression analysis is merely a way to characterize the conditional distribution of the response variable, given a set of predictors. Standardized coefficients do not represent the causal importance of a variable. Contributions to explained variance for different predictors do not represent the causal importance of a variable. A good overall fit does not demonstrate that a causal model is correct. Moreover, there are no regression diagnostics, he argued, through which causal effects can be demonstrated with confidence. There are no specification tests through which causal effects can be demonstrated. There are no mathematical formalisms through which causal effects can be demonstrated. In short, Berk argues that causal inference rests first and foremost on a credible response schedule. Without a credible response schedule and in the absence of a well-supported model (i.e., hypothesized relationships based on prior research and theory), the potential problems with regression-based causal models outweigh existing remedial strategies (Berk, 2004, p. 224). 3.7 CONCLUSION In this chapter, we reviewed conditions under which the fundamental assumption of ignorable treatment assignment is violated, and we discussed three conventional approaches (i.e., regression, matching, and stratification) to balance data in the presence of nonignorable treatment assignment (i.e., selection bias). In a simulation based on five scenarios of data generation, conventional methods worked well only in an ideal scenario in which the ignorable treatment assignment assumption is met. Unfortunately, these ideal conditions are most likely to be satisfied in a randomized experiment. In social and health sciences evaluations involving the use of observational data, they are unlikely to be satisfied. Thus, the findings suggest that the use of OLS regression to control for covariates and balance data in the absence of a randomized design 130 may be unwarranted in many settings where a causal inference is desired. New methods to estimate treatment effects when treatment assignment is nonignorable are needed. NOTE 1. However, in our current data simulation, we cannot see the consequence of inconsistency, because the current simulation does not look into sampling variability and sampling properties when sample size becomes extremely large. 131 CHAPTER 4 Sample Selection and Related Models This chapter describes two models: the sample selection model and the treatment effect model. Heckman's (1974, 1978, 1979) sample selection model was developed using an econometric framework for handling limited dependent variables. It was designed to address the problem of estimating the average wage of women using data collected from a population of women in which women who stayed at home—who were not in the labor market—were excluded by self-selection. Based on this data set, Heckman's original model focused on the incidental truncation of a dependent variable. Maddala (1983) extended the sample selection perspective to the evaluation of treatment effectiveness. We review Heckman's model first because it not only offers a theoretical framework for modeling sample selection but is also based on what was at the time a pioneering approach to correcting selection bias. Equally important, Heckman's model lays the groundwork for understanding the treatment effect model. The sample selection model is an important contribution to program evaluation; however, the treatment effect model is the focus of this chapter because this model offers practical solutions to various types of evaluation problems. Section 4.1 describes the main features of the Heckman model. Section 4.2 reviews the treatment effect model. Section 4.3 provides an overview of the Stata programs that are applicable for estimating the models described here. Examples in Section 4.4 illustrate the treatment effect model and show how to use this mode to solve typical evaluation problems. Section 4.5 concludes with a review of key points. 4.1 THE SAMPLE SELECTION MODEL Undoubtedly, Heckman's sample selection model is a seminal contribution to 20th-century program evaluation. The sample selection model triggered both a rich theoretical discussion on modeling selection bias and the development of new statistical procedures to address selection effects. Heckman's key contributions to program evaluation include the following: (a) He provided a 132 theoretical framework that emphasized the importance of modeling the endogenous variable; (b) his model was the first attempt to estimate the probability (i.e., the propensity score) of a participant being in one of the two conditions indicated by the endogenous dummy variable, and then used the estimated propensity score model to estimate coefficients of the regression model; (c) he treated the unobserved selection factors as a problem of specification error or a problem of omitted variables and corrected for bias in the estimation of the outcome equation by explicitly using information gained from the model of sample selection; and (d) he developed a creative two-step procedure by using the simple least squares algorithm. To understand Heckman's model, we first review concepts related to the handling of limited dependent variables. 4.1.1 Truncation, Censoring, and Incidental Truncation Due to censoring and truncation, limited dependent variables are common in social and health data. Truncation, which is an effect of data gathering rather than data generation, occurs when sample data are drawn from a subset of a larger population of interest. Thus, a truncated distribution is part of a larger, untruncated distribution. For instance, assume that an income survey was administered to a limited subset of the population (e.g., those whose incomes are above the poverty threshold). In the data from such a survey, the dependent variable will be observed only for a portion of the whole distribution. The task of modeling is to use that limited information—a truncated distribution—to infer the income distribution for the entire population. Censoring occurs when all values in a certain range of a dependent variable are transformed to a single value. Using the above example of population income, censoring differs from truncation in that the data collection may include the entire population, but below-poverty-threshold incomes are coded as zero. Under this condition, researchers may estimate a regression model for a larger population using both the censored and the uncensored data. Censored data are ubiquitous. They include (a) household purchases of durable goods, in which low expenditures for durable goods are censored to a zero value (the Tobit model, developed by James Tobin in 1958, is the most widely known model for analyzing this kind of dependent variable); (b) number of extramarital affairs, in which the number of affairs beyond a certain value is collapsed into a maximum count; (c) number of hours worked by women in the labor force, in which women who work outside the home for a low number of hours are censored to a zero value; and (d) number of arrests after release from prison, where arrests beyond a certain value are scored as a maximum (Greene, 2003). The central task of analyzing limited dependent variables is to use the truncated distribution or censored data to infer the untruncated or uncensored distribution for the entire population. In the context of regression analysis, we 133 typically assume that the dependent variable follows a normal distribution. The challenge then is to develop moments (mean and variance) of the truncated or censored normal distribution. Theorems of such moments have been developed and can be found in textbooks on the analysis of limited dependent variables. In these theorems, moments of truncated or censored normal distributions involve a key factor called the inverse Mills ratio, or hazard function, which is commonly denoted as λ. Heckman's sample selection model uses the inverse Mills ratio to estimate the outcome regression. In Section 4.1.3, we review moments for sample selection data and the inverse Mills ratio. A concept closely related to truncation and censoring, or a combination of the two concepts, is incidental truncation. Indeed, it is often used interchangeably with the term sample selection. From Greene (2003), suppose you are funded to conduct a survey of persons with high incomes and that you define eligible respondents as those with a net worth of \$500,000 or more. This selection by income is a form of truncation—but it is not quite the same as the general case of truncation. The selection criterion (e.g., at least \$500,000 net worth) does not exclude those individuals whose current income might be quite low, although they had previously accrued high net worth. Greene (2003) explained by saying, Still, one would expect that, on average, individuals with a high net worth would have a high income as well. Thus, the average income in this subpopulation would in all likelihood also be misleading as an indication of the income of the typical American. The data in such a survey would be nonrandomly selected or incidentally truncated. (p. 781) Thus, sample selection or incidental truncation refers to a sample that is not randomly selected. It is in situations of incidental truncation that we encounter the key challenge to the entire process of evaluation, that is, the departure of evaluation data from the classic statistical model that assumes a randomized experiment. This challenge underscores the need to model the sample selection process explicitly. We encounter these problems explicitly and implicitly in many data situations. Consider the following from Maddala (1983). Example 1: Married women in the labor force. This is the problem Heckman (1974) originally considered under the context of shadow prices (i.e., women's reservation wage or the minimum wage rate at which a woman who is at home might accept marketplace employment), market wages, and labor supply. Let y* be the reservation wage of a stay-at-home woman (or, homemaker) based on her valuation of time in the household. Let y be the market wage based on an employer's valuation of her effort in the labor force. According to Heckman, a woman participates in the labor force if y > y*. Otherwise, a woman is not considered a participant in the labor force. In any given sample, we have observations only on y for those women who participate in the labor force, and 134 we have no observation on y for the women not in the labor force. For women not in the labor force, we only know that y* ≥ y. In other words, the sample is not randomly selected, and we need to use the sample data to estimate the coefficients in a regression model explaining both y* and y. As explained in the following by Maddala (1983), with regard to women who are not in the labor market and who work at home, the problem is truncation, or more precisely incidental truncation, not censoring, because we do not have any observations on either the explained variable y or the explanatory variable x in the case of the truncated regression model if the value of y is above (or below) a threshold. . . . In the case of the censored regression model, we have data on the explanatory variables x for all the observations. As for the explained variable y, we have actual observations for some, but for others we know only whether or not they are above (or below) a certain threshold. (pp. 5–6) Example 2: Effects of unions on wages. Suppose we have data on wages and personal characteristics of workers that include whether the worker is a union member. A naive way of estimating the effects of unionization on wages is to estimate a regression of wage on the personal characteristics of the workers (e.g., age, race, sex, education, and experience) plus a dummy variable that is defined as D = 1 for unionized workers and D = 0 otherwise. The problem with this regression model lies in the nature of D. This specification treats the dummy variable D as exogenous when D is not exogenous. In fact, there are likely many factors affecting a worker's decision whether to join the union. As such, the dummy variable is endogenous and should be modeled directly; otherwise, the wage regression estimating the impact of D will be biased. We have seen the consequences of the naive treatment of D as an exogenous variable in both Chapters 2 and 3. Example 3: Effects of fair-employment laws on the status of African American workers. Consider a regression model (Lands, 1968) relating to the effects of fair-employment legislation on the status of African American workers y_i = αx_i + βDi + ui, where y_i is the wage of African Americans relative to that for whites in state i, Xi is the vector of exogenous variables for state i, Di = 1 if state i has a fair-employment law (Di = 0 otherwise), and ui is a residual. Here the same problem of the endogeneity of D is found as in our second example, except that the unit of analysis in the previous example is individual, whereas the unit in the current example is state i. Again, Di is treated as exogenous when in fact it is endogenous. "States in which African Americans would fare well without a fair-employment law may be more likely to pass such a law if legislation depends on the consensus" (Maddala, 1983, p. 8). Heckman (1978) observed, 135 An important question for the analysis of policy is to determine whether or not measured effects of legislation are due to genuine consequences of legislation or to the spurious effect that the presence of legislation favorable to blacks merely proxies the presence of the pro-black sentiment that would lead to higher status for blacks in any event. (p. 933) Example 4: Compulsory school attendance laws and academic or other outcomes. The passage of compulsory school attendance legislation is itself an endogenous variable. Similar to Example 3, it should be modeled first. Otherwise, the estimation of the impact of such legislation on any outcome variable risks bias and inconsistency (Edwards, 1978). Example 5: Returns of college education. In this example, we are given income for a sample of individuals, some with a college education and others without. Because the decision whether to attend college is a personal choice determined by many factors, the dummy variable (attending vs. not attending) is endogenous and should be modeled first. Without modeling this dummy variable first, the regression of income showing the impact of college education would be biased, regardless of whether the regression model controlled for covariates such as IQ (intelligence quotient) or parental socioeconomic status. Today, these illustrations are considered classic examples, and they have been frequently cited and discussed in the literature on sample selection. The first three examples were discussed by Heckman (1978, 1979) and motivated his work on sample selection models. These examples share three features: (1) The sample being inferred was not generated randomly, (2) a binary explanatory variable was endogenous rather than exogenous, and (3) sample selection or incidental truncation must be considered in the evaluation of the impact of such a dummy variable. However, there is an important difference between Example 1 and the other four examples. In Example 1, we observe only the outcome variable (i.e., market wage) for women who participate in the labor force (i.e., only for participants whose Di = 1; we do not observe the outcome variable for women whose Di = 0), whereas in Examples 2 through 5, the outcome variables (i.e., wages, the wage status of African American workers relative to that of white workers, academic achievement, and income) for both the participants (or states) whose Di = 1 and Di = 0 are observed. Thus, Example 1 is a sample selection model, and the other four examples illustrate the treatment effect model. The key point is the importance of distinguishing between these two types of models: (1) the sample selection model (i.e., the model analyzing outcome data observed only for Di = 1) and (2) the treatment effect model (i.e., the model analyzing outcome data observed for both Di = 1 and Di = 0). Both models share common characteristics and may be viewed as Heckman-type models. However, the treatment effect model focuses on program evaluation, 136 which is not the intent of the sample selection model. This distinction is important when choosing appropriate software. In the Stata software, for example, the sample selection model is estimated by the program heckman, and the treatment effect model is estimated by the program treatreg; we elaborate on this point in Section 4.3. 4.1.2 Why Is It Important to Model Sample Selection? Although the topic of sample selection is ubiquitous in both program evaluation and observational studies, the importance of giving it a formal treatment was largely unrecognized until Heckman's (1974, 1976, 1978, 1979) work and the independent work of Rubin (1974, 1978, 1980b, 1986). Recall that, in terms of causal inference, sample selection was not considered a problem in randomized experiments because randomization renders selection effects irrelevant. In nonrandomized studies, Heckman's work emphasized the importance of modeling sample selection by using a two-step procedure or switching regression, whereas Rubin's work drew the same conclusion by applying a generalization of the randomized experiment to observational studies. Heckman focused on two types of selection bias: self-selection bias and selection bias made by data analysts. Heckman (1979) described self-selection bias as follows: One observes market wages for working women whose market wage exceeds their home wage at zero hours of work. Similarly, one observes wages for union members who found their nonunion alternative less desirable. The wages of migrants do not, in general, afford a reliable estimate of what nonmigrants would have earned had they migrated. The earnings of manpower trainees do not estimate the earnings that nontrainees would have earned had they opted to become trainees. In each of these examples, wage or earnings functions estimated on selected samples do not in general, estimate population (i.e., random sample) wage functions. (pp. 153–154) Heckman argued that the second type of bias, selection

bias made by data analysts or data processors, operates in much the same fashion as self-selection bias. In their later work, Heckman and his colleagues generalized the problem of selectivity to a broad range of social experiments and discussed additional types of selection biases (e.g., see Heckman & Smith, 1995). From Maddala (1983), Figure 4.1 describes three types of decisions that create selectivity (i.e., individual selection, administrator selection, and attrition selection). In summary, Heckman's approach underscores the importance of modeling selection effects. When selectivity is inevitable, such as in observational 137 studies, the parameter estimates from a naive ordinary least squares (OLS) regression model are inconsistent and biased. Alternative analytic strategies that model selection must be explored.

4.1.3 Moments of an Incidentally Truncated Bivariate Normal Distribution

The theorem for moments of the incidentally truncated distribution defines key functions such as the inverse Mills ratio under the setting of a normally distributed variable. Our discussion follows Greene (2003). Figure 4.1 Decision Tree for Evaluation of Social Experiments Source: Maddala (1983, p. 266). Reprinted with the permission of Cambridge University Press. Suppose that y and z have a bivariate normal distribution with correlation ρ . We are interested in the distribution of y given that z exceeds a particular value a . The truncated joint density of y and z is given the truncated joint density of y and z , given that y and z have a bivariate normal distribution with means μ_y and μ_z , standard deviations σ_y and σ_z , and correlation ρ , the moments (mean and variance) of the incidentally truncated variable y are as follows (Greene, 2003, p. 781): where a is the cutoff threshold, $\phi(z) = \frac{1}{\sigma_z} \phi\left(\frac{z - \mu_z}{\sigma_z}\right)$, $\lambda(z) = \frac{\phi(z)}{\Phi(z)}$, $\lambda^*(z) = \frac{\phi(z)}{\Phi(z)} \left[1 - \Phi\left(\frac{z - \mu_z}{\sigma_z}\right)\right]$, $\delta(z) = \lambda(z) - \lambda^*(z)$, $\phi(z)$ is the standard normal density function, and $\Phi(z)$ is the standard cumulative distribution function. In the preceding equations, $\lambda(z)$ is called the inverse Mills ratio and is used in Heckman's derivation of his two-step estimator. Note that in this theorem, we consider moments of a single variable; in other words, this is a theorem about univariate properties of the incidental truncation of y . Heckman's model applied and expanded the theorem to a multivariate case in which an incidentally truncated variable is used as a dependent variable in a regression analysis.

4.1.4 The Heckman Model and Its Two-Step Estimator

A sample selection model always involves two equations: (1) the regression equation considering mechanisms determining the outcome variable and (2) the selection equation considering a portion of the sample whose outcome is observed and mechanisms determining the selection process (Heckman, 1978, 1979). To put this model in context, we revisit the example of the wage earning of women in the labor force (Example 1, Section 4.1.1). Suppose we assume that the hourly wage of women is a function of education (educ) and age (age), whereas the probability of working (equivalent to the probability of wage being observed) is a function of marital status (married) and number of children at home (children). To express the model, we can write two equations, the regression equation of wage and the selection equation of working: wage is observed if Note that the selection equation indicates that wage is observed only for those women whose wages were greater than 0 (i.e., women were considered as having participated in the labor force if and only if their wage was above a certain threshold value). Using a zero value in this equation is a normalization convenience and is an alternate way to say that the market wage of women who participated in the labor force was greater than their reservation wage (i.e., $y > y^*$). The fact that the market wage of homemakers (i.e., those not in the paid labor force) was less than their reservation wage (i.e., $y < y^*$) is expressed in the preceding model through the fact that these women's wage was not observed in the regression equation; that is, it was incidentally truncated. The selection model further assumes that u_1 and u_2 are correlated to have a nonzero correlation ρ . This example can be expanded to a more general case. For the purpose of modeling any sample selection process, two equations are used to express the determinants of outcome y_i : 139 and where x_i is a vector of exogenous variables determining outcome y_i , and ϵ_i is a latent endogenous variable. If ϵ_i is greater than the threshold value (say value 0), then the observed dummy variable $w_i = 1$, and otherwise $w_i = 0$; the regression equation observes value y_i only for $w_i = 1$; z_i is a vector of exogenous variables determining the selection process or the outcome of y_i ; $\Phi(\cdot)$ is the standard normal cumulative distribution function; and u_j and e_j are error terms of the two regression equations and assumed to be bivariate normal, with mean zero and covariance matrix Given incidental truncation and censoring of y , the evaluation task is to use the observed variables (i.e., y , x , z , and probably w) to estimate the regression coefficients β that are applicable to sample participants whose values of w equal both 1 and 0. The sample selection model can be estimated by either the maximum likelihood method or the least squares method. Heckman's two-step estimator uses the least squares method. We review the two-step estimator first. The maximum likelihood method is reviewed in the next section as a part of a discussion of the treatment effect model. To facilitate the understanding of Heckman's original contribution, we use his notations that are slightly different from those used in our previous discussion. Heckman first described a general model containing two structural equations. The general model considers continuous latent random variables and can be expressed as follows: where $X1$ and $X2$ are row vectors of bounded exogenous variables; d_i is a dummy variable defined by and 140 Heckman next discussed six cases where the general model applies. His interest centered on the sample selection model, or Case 6 (Heckman, 1978, p. 934). The primary feature of Case 6 is that structural shifts in the equations are permitted. Furthermore, Heckman allowed that was observed, so the variable can be written without an asterisk, as $y1$, and is not observed. Writing the model in reduced form (i.e., only variables on the right-hand side should be exogenous variables), we have the following equations: where Π_i is the conditional probability of $d_i = 1$, and The model assumes that $U1$ and $U2$ are bivariate normal random variables. Accordingly, the joint distribution of $V1$, $V2$, $h(V1, V2)$, is a bivariate normal density fully characterized by the following assumptions: For the existence of the model, the analyst has to impose restrictions. A necessary and sufficient condition for the model to be defined is that $\pi23 = 0 = \gamma2\beta1 + \beta2$. Heckman called this condition the principal assumption. Under this assumption, the model becomes where $\pi11 \neq 0$, $\pi12 \neq 0$, $\pi21 \neq 0$, and $\pi22 \neq 0$. With the above specifications and assumptions, the model (4.5) can be estimated in two steps: 1. First, estimate Equation 4.5b, which is analogous to solving the problem of a probit model. We estimate the conditional probabilities of the events $d_i = 1$ and $d_i = 0$ by treating $y2i$ as a dummy variable. Doing so, $\pi21$ and $\pi22$ are estimated. Subject to the standard requirements for the identification and existence of probit estimation, the analyst needs to normalize the equation 141 by and estimate 2. Second, estimate Equation 4.5a. Rewrite Equation 4.5a as the conditional expectation of $y1i$ given d_i , $X1i$, and $X2i$: Using a result of biserial correlation is estimated: where w and Φ are the density and distribution function of a standard normal random variable, respectively; Because can now be estimated, a nd Equation 4.6 can be solved by the standard least squares method. Note that refers to a truncation of y whose truncated z exceeds a particular value a (see Equation 4.1). Under this condition, Equation 4.7 Using estimated from Step 1, becomes is calculated using Now in the equation of are known, the only coefficient to be determined is thus, solving Equation 4.6 is a matter of estimating the following regression: Therefore, the parameters can be estimated by using the standard OLS estimator. A few points are particularly worth noting. First, in Equation 4.5b, $V2i$ is an error term or residual of the variation in the latent variable y_2 , after the variation is explained away by $X1i$ and $X2i$. This is a specification error or, more precisely, a case of unobserved heterogeneity determining selection bias. This specification error is treated as a true omitted variable problem and is creatively taken into consideration when estimating the parameters of Equation 4.5a. In other words, the impact of selection bias is neither through away nor assumed to be random but is explicitly used and modeled in the equation estimating the outcome regression. This treatment for selection bias connotes Heckman's contribution and distinguishes the econometric solution to the selection bias problem from that of the statistical tradition. Important implications of this modeling feature were summarized by Heckman (1979, p. 155). In addition, there are different formulations for estimating the model 142 parameters that were developed after Heckman's original model. For instance, Greene (1981, 2003) constructed consistent estimators of the individual parameter ρ (i.e., the correlation of the two error terms) and σ_{ϵ} (i.e., the variance of the error term of the regression equation). However, Heckman's model has become standard in the literature. Last, the same sample selection model can also be estimated by the maximum likelihood estimator (Greene, 1995), which yields results quite similar to those produced using the least squares estimator. Given that the maximum likelihood estimator requires more computing time, and computing speed was considerably slower than today, Heckman's least squares solution is a remarkable contribution. More important, Heckman's solution was devised within a framework of structural equation modeling that is simple and succinct and that can be used in conjunction with the standard framework of OLS regression.

4.2 TREATMENT EFFECT MODEL

Since the development of the sample selection model, statisticians and econometricians have formulated many new models and estimators. In mimicry of the Tobit or logit models, Greene (2003) suggested that these Heckman-type models might be called "Heckit" models. One of the more important of these developments was the direct application of the sample selection model to the estimation of treatment effects in observational studies. The treatment effect model differs from the sample selection model—that is, in the form of Equation 4.2—in two aspects: (1) a dummy variable indicating the treatment condition w_i (i.e., $w_i = 1$ if participant i is in the treatment condition, and $w_i = 0$ otherwise) is directly entered into the regression equation and (2) the outcome variable y_i of the regression equation is observed for both $w_i = 1$ and $w_i = 0$. Specifically, the treatment effect model is expressed in two equations: and where ϵ_i and u_i are bivariate normal with mean zero and covariance matrix Given incidental truncation (or sample selection) and that w is an endogenous dummy variable, the evaluation task is to use the observed 143 variables to estimate the regression coefficients β while controlling for selection bias induced by nonignorable treatment assignment. Note that the model expressed by Equations 4.8a and 4.8b is a switching regression. By substituting w_i in Equation 4.8a with Equation 4.8b, we obtained two different equations of the outcome regression: and This is Quandt's (1958, 1972) form of the switching regression model that explicitly states that there are two regimes: treatment and nontreatment. Accordingly, there are separate models for the outcome under each regime: For whereas for treated participants, the outcome model is nontreated participants, the outcome model is The treatment effect model illustrated earlier can be estimated in a two-step procedure similar to that described for the sample selection model. To increase the efficiency of our exposition of models, we move on to the maximum likelihood estimator. Readers who are interested in the two-step estimator may consult Maddala (1983). Let $f(\epsilon, u)$ be the joint density function of ϵ and u defined by Equations 4.8a and 4.8b. According to Maddala (1983, p. 129), the joint density function of y and w is given by the following: and Thus, the log-likelihood functions for participant i (StataCorp, 2003) are as follows: For $w_i = 1$, For $w_i = 0$, 144 The treatment effect model has many applications in program evaluation. In particular, it is useful when evaluators have data that were generated by a nonrandomized experiment and, thus, are faced with the challenge of nonignorable treatment assignment or selection bias. We illustrate the application of the treatment effect model in Section 4.4. 4.3 OVERVIEW OF THE STATA PROGRAMS AND MAIN FEATURES OF TREATREG

Most models described in this chapter can be estimated by the Stata and R packages. Many helpful user-developed programs are also available from the Internet. Within Stata, heckman can be used to estimate the sample selection model, and treatreg can be used to estimate the treatment effect model. In Stata, heckman was developed to estimate the original Heckman model; that is, it is a model that focuses on incidentally truncated dependent variables. Using wage data collected from a population of employed women in which homemakers were self-selected out, Heckman wanted to estimate determinants of the average wage of the entire female population. Two characteristics distinguish this kind of problem from the treatment effect model. The dependent variable is observed only for a subset of sample participants (e.g., only observed for women in the paid labor force), and the group membership variable is not entered into the regression equation (see Equations 4.2a and 4.2b). Thus, the task fulfilled by heckman is different from the task most program evaluators or observational researchers aim to fulfill. Typically, for study samples such as the group of women in the paid labor force, program evaluators or researchers will have observed outcomes for participants in both conditions. Therefore, the treatment membership variable is entered into the regression equation to discern treatment effects. We emphasize these differences because it is treatreg, rather than heckman, that offers practical solutions to various types of evaluation problems. The treatreg program can be initiated using the following basic syntax: treatreg depvar [indepvars], treat(depvar_t = indepvars_t) [wastep] where depvar is the outcome variable on which users want to assess the difference between treated and control groups, indepvars is a list of variables that users hypothesize would affect the outcome variable, depvar_t is the treatment membership variable that denotes intervention condition, indepvars_t is the list of variables that users anticipate will determine the selection process, and wastep is an optional specification to request an estimation using a twostep consistent estimator. In other words, absence of wastep is the default; under the default, Stata estimates the model using a full maximum likelihood. 145 Using notations from the treatment effect model (i.e., Equations 4.8a and 4.8b), depvar is y , indepvars are the vector x , and depvar_t is w in Equation 4.8a and indepvars_t are the vector z in Equation 4.8b. By design, x and z can be the same variables if the user suspects that covariates of selection are also covariates of the outcome regression. Similarly, x and z can be different variables if the user suspects that covariates of selection are different from covariates of the outcome regression (i.e., x and z are two different vectors). However, z is part of x , if the user suspects that additional covariates affect y but not w , or vice versa, if one suspects that additional covariates affect w but not y . The treatreg program supports Stata standard functions, such as the HuberWhite estimator of variance under the robust and cluster() options, as well as incorporating sampling weights into analysis under the weight option. These functions are useful for researchers who analyze survey data with complex sampling designs using unequal sampling weights and multistaged stratification. The weight option is available only for the maximum likelihood estimation and supports various types of weights, such as sampling weights (i.e., specify pweights = varname), frequency weights (i.e., specify fweights = varname), analytic weights (i.e., specify aweights = varname), and importance weights (i.e., specify iweights = varname). When the robust and cluster() options are specified, Stata follows a convention that does not print model Wald chi-square, because that statistic is misleading in a sandwich (Huber-White) correction of standard errors. Various results can be saved for postestimation analysis. You may use either predict to save statistics or variables of interest, or ereturn list to check scalars, macros, and matrices that are automatically saved. We now turn to an example (i.e., Section 4.4.1), and we will demonstrate the syntax. We encourage readers to briefly review the study details of the example before moving on to the application of treatreg. To demonstrate the treatreg syntax and printed output, we use data from the National Survey of Child and Adolescent Well-Being (NSCAW). As explained in Section 4.4.1, the NSCAW study focused on the well-being of children whose primary caregiver had received treatment for substance abuse problems. For our demonstration study, we use NSCAW data to compare the psychological outcomes of two groups of children: those whose caregivers received substance abuse services (treatment variable aodserve = 1) and those whose caregivers did not (treatment variable aodserve = 0). Psychological outcomes were assessed using the Child Behavior Checklist–Externalizing (CBCLExternalizing) score (i.e., the outcome variable external3). Variables entering into the selection equation (i.e., the z vector in Equation 4.8b) are cgrage1, cgrage2, cgrage3, high, bahigh, employ, open, sexual, provide, supervis, other, cra47a, mental, arrest, psh17a, cidl, and cgneed. Variables entering into the regression equation (i.e., the x vector in Equation 4.8a) are black, hispanic, 146 natam, chdage2, chdage3, and ra. Table 4.1 exhibits the syntax and output. Important statistics printed by the output are explained in the following. First, rho is the estimated ρ in the variance-covariance matrix, which is the correlation between the error ϵ_i of the regression equation (4.8a) and the error u_i of the selection equation (4.8b). In this example, $\rho = -.3603391$, which is estimated by Stata through the inverse hyperbolic tangent of ρ (i.e., labeled as "athrho" in the output). The statistic "athrho" is merely a middle step through which Stata obtains estimated ρ . It is the estimated ρ (i.e., labeled as rho in the output) that serves an important function.1 The value of sigma is the estimated in the above variance-covariance matrix, which is the variance of the regression equation's error term (i.e., variance of ϵ_i in Equation 4.8a). In this example, $\sigma = 12.1655$, which is estimated by Stata through $\ln(\cdot)$ (i.e., labeled as "lnsigma" in the output). As with "athrho," "lnsigma" is a middle-step statistic that is relatively unimportant to users. The statistic labeled "lmbda" is the inverse Mills ratio, or nonselection hazard, which is the product of two terms: Note that this is the statistic Heckman used in his two-step estimator (i.e., in Equation 4.7) to obtain a consistent estimation of the first-step equation. In the early days of discussing the Heckman or Heckit models, some researchers, especially economists, assumed that λ could be used to measure the level of selectivity effect, but this idea proved controversial and is no longer widely practiced. The estimated nonselection hazard (i.e., λ) can also be saved as a new variable in the data set for further analysis, if the user specifies hazard(newvarname) as a treatreg option. Table 4.2 illustrates this specification and prints out the saved hazard (variable h1) for the first 10 observations and the descriptive statistics. Second, because the treatment effect model assumes that the level of correlation between the two error terms is nonzero, and because violation of that assumption can lead to estimation bias, it is often useful to test $H_0: \rho = 0$. Stata prints results of a likelihood ratio test against " $H_0: \rho = 0$ " at the bottom of the output. This ratio test is a comparison of the joint likelihood of an independent probit model for the selection equation and a regression model on the observed data against the treatment effect model likelihood. Given that $\chi^2 = 9.47$ ($p < .01$) from Table 4.1, we can reject the null hypothesis at a statistically significant level and conclude that ρ is not equal to 0. This suggests that applying the treatment effect model is appropriate. Third, the reported model $\chi^2 = 58.97$ ($p < .0001$) from Table 4.1 is a Wald test of all coefficients in the regression model (except constant) being zero. This is one method to gauge the goodness of fit of the model. With $p < .0001$, the user can conclude that the covariates used in the regression model may be appropriate, and at least one of the covariates has an effect that is not equal to zero. 147 Fourth, interpreting regression coefficients for the regression equation (i.e., the top panel of the output of Table 4.1) is performed in the same fashion as that used for a regression model. The sign and magnitude of the regression coefficient indicate the net impact of an independent variable on the dependent variable: other things being equal, the amount of change observed on the outcome with each one-unit increase in the independent variable. A one-tailed or two-tailed significance test on a coefficient of interest may be estimated using z and its associated p values. However, interpreting the regression coefficients of the selection equation is complicated because the observed w variable takes only two values (0 vs. 1), and the estimation process uses the probability of $w = 1$. Nevertheless, the sign of the coefficient is always meaningful, and significance of the coefficient is important. For example, using the variable open (whether a child welfare case was open at baseline: open = 1, yes; open = 0, no), because the coefficient is positive (i.e., coefficient of open = .5095), we know that the sample selection process (receipt or no receipt of services) is positively related to child welfare case status. That is, a caregiver with an open child welfare case was more likely to receive substance abuse services, and this relationship is statistically significant. Thus, coefficients with p values less than .05 indicate variables that contribute to selection bias. In this example, we observe eight variables with p values of less than .05 (i.e., variables cgrage1, cgrage2, open, mental, arrest, psh17a, cidl, and cgneed). The significance of these variables indicates the presence of selection bias and underscores the importance of explicitly considering selection when modeling child outcomes. The eight variables are likely to be statistically significant in a logistic regression using the logit of service receipt (i.e., the logit of aodserv) as a dependent variable and the same set of selection covariates as independent variables. Table 4.1 Exhibit of Stata treatreg Output for the NSCAW Study 148 Source: Data from NSCAW, 2004. Table 4.2 Exhibit of Stata treatreg Output: Syntax to Save Nonselection Hazard 149 Fifth, the estimated treatment effect is an indicator of program impact net of observed selection bias; this statistic is shown by the coefficient associated with the treatment membership variable (i.e., aodserv in the current example) in the regression equation. As shown in Table 4.1, this coefficient is 8.601002, and the associated p value is .001, meaning that other things being equal, children whose caregivers received substance abuse services had a mean score that was 8.6 units greater than children whose caregivers did not receive such services. The difference is statistically significant at a .001 level. As previously mentioned, Stata automatically saves scalars, macros, and matrices for postestimation analysis. Table 4.3 shows the saved statistics for the demonstration model (Table 4.1). Automatically saved statistics can be recalled using the command "ereturn list." 4.4 EXAMPLES This section describes three applications of the Heckit treatment effect model in social behavioral research. The first example comes from the NSCAW study, and, as in the treatreg syntax illustration, it estimates the impact on child wellbeing of the participation of children's caregivers in substance abuse treatment services. This study is typical of those that use a large, nationally representative survey to obtain observational data (i.e., data generated through a nonexperimental process). It is not uncommon in such studies to use a covariance control approach in an attempt to estimate the impact of program 150 participation. Our second example comes from a program evaluation that originally included a group randomization design. However, the randomization failed, and researchers were left with a group-design experiment in which treatment assignment was not ignorable. The example demonstrates the use of the Heckit treatment effect model to correct for selection bias while estimating treatment effectiveness. The third example illustrates how to run the treatment effect model after multiple imputations of missing data. 4.4.1 Application of the Treatment Effect Model to Analysis of Observational Data Child maltreatment and parental substance abuse are highly correlated (e.g., English et al., 1998; U.S. Department of Health and Human Services [DHHS], 1999). A caregiver's abuse of substances may lead to maltreatment through many different mechanisms. For example, parents may prioritize their drug use more highly than caring for their children, and substance abuse can lead to extreme poverty and to incarceration, both of which often leave children with unmet basic needs (Magura & Laudet, 1996). Policy makers have long been concerned about the safety of the children of substance-abusing parents. Described briefly earlier, the NSCAW study was designed to address a range of questions about the outcomes of children who are involved in child welfare systems across the country (NSCAW Research Group, 2002). NSCAW is a nationally representative sample of 5,501 children, ages 0 to 14 years at intake, who were investigated by child welfare services following a report of child maltreatment (e.g., child abuse or neglect) between October 1999 and December 2000 (i.e., a multiwave data collection corresponding to the data employed by this example). The NSCAW sample was selected using a twostage stratified sampling design (NSCAW Research Group, 2002). The data were collected through interviews conducted with children, primary caregivers, teachers, and child welfare workers. These data contain detailed information on child development, social and psychological adjustment, mental health and other symptoms, service participation, environmental conditions, and protective services placements (e.g., placement in foster care or a group home). NSCAW gathered data over multiple waves, and the sample represented children investigated as victims of child abuse or neglect in 92 primary sampling units, principally counties, in 36 states. Table 4.3 Exhibit of Stata treatreg Output: Syntax to Check Saved Statistics 151 Source: Data from NSCAW, 2004. The analysis for this example uses the NSCAW Wave 2 data, or the data from the 18-month follow-up survey. Therefore, the analysis employs one-time-point data that were collected 18 months after the baseline. For the purposes of our demonstration, the study sample was limited to 1,407 children who lived at home (i.e., not in foster care), whose primary caregiver was female, and who were 4 years of age or older at baseline. We limited the study sample to children with female caregivers because females composed the vast majority (90%) of primary caregivers in NSCAW. In addition, because NSCAW is a large observational database and our research questions focus on the impact of caregivers' receipt of substance abuse services on children's well-being, it is important to model the process of treatment assignment directly; therefore, the heterogeneity of potential causal effects is taken into consideration. In the 152 NSCAW survey, substance abuse treatment was defined using six variables that asked the caregiver or child welfare worker whether the caregiver had received treatment for an alcohol or drug problem at the time of the baseline interview or at any time in the following 18 months. Our analysis of NSCAW data was guided by two questions: (1) After 18 months of involvement with child welfare services, how were children of caregivers who received substance abuse services faring? and (2) Did children of caregivers who received substance abuse services have more severe behavioral problems than did their counterparts whose caregivers did not receive such services? As described previously, the choice of covariates hypothesized to affect sample selection serves an essential role in the analysis. We chose these variables based on our review of the substance abuse literature through which we determined the characteristics that were often associated with substance abuse treatment receipt. Because no studies focused exclusively on female caregivers involved with child welfare services, we had to rely on literature regarding substance abuse in the general population (e.g., Knight, Logan, & Simpson, 2001; McMahon, Winkel, Suchman, & Luthar, 2002; Weisner, Jennifer, Tam, & Moore, 2001). We found four categories of characteristics: (1) social demographic characteristics (e.g., caregiver's age, less than 35 years, 35 to 44 years, 45 to 54 years, and above 54 years; caregiver's education, less than high school degree, high school degree, and bachelor's degree or higher; caregiver's employment status, employed/not employed; and whether the caregiver had "trouble paying for basic necessities," which was answered—yes/no), (2) risks (e.g., caregiver mental health problems—yes/no; child welfare case status—closed/open; caregiver history of arrest—yes/no; and the type of child maltreatment—physical abuse, sexual abuse, failure to provide, failure to supervise, and other); (3) caregiver's prior receipt of substance abuse treatment (i.e., caregiver alcohol or other drug treatment—yes/no); and (4) caregiver's need for alcohol and drug treatment services (i.e., measured on the World Health Organization's Composite International Diagnostic Interview–Short Form [CIDI-SF] that reports the presence/absence of the need for services and caregiver's self-report of service need—yes/no). The outcome variable is the Achenbach Children's Behavioral Checklist (CBCL4–18) that is completed by the caregivers. This scale includes scores for externalizing and internalizing behaviors (Achenbach, 1991). A high score on each of these measures indicates a greater extent of behavioral problems. When we conducted the outcome regression, we controlled for the following covariates: child's race/ethnicity (Black/non-Hispanic, White/non-Hispanic, Hispanic, and Native American), child's age (4–5 years, 6–10 years, and 11 years and older), and risk assessment by child welfare worker at the baseline (risk absence/risk presence). Table 4.4 presents descriptive statistics of the study sample. Of 1,407 153 children, 112 (8% of the sample) had a caregiver who had received substance abuse services, and 1,295 (92% of the sample) had caregivers who had not received services. Of 11 study variables, 8 showed statistically significant differences ($p < .01$) between treated cases (i.e., children whose caregivers had received services) and nontreated cases (i.e., children whose caregivers had not received services). For instance, the following caregivers were more likely to have received treatment services: those with a racial/ethnic minority status; with a positive risk to children; who were currently unemployed; with a current, open child welfare case; those who were investigated for child maltreatment types of failure to provide or failure to supervise; those who had trouble paying for basic necessities; with a history of mental health problems; those with a history of arrest; those with prior receipt of substance abuse treatment; CIDI-SF positive; and those who self-reported needing services. Without controlling for these selection effects, the estimates of differences on child outcomes would clearly be biased. Table 4.5 presents the estimated differences in psychological outcomes between groups before and after adjustments for sample selection. Taking the externalizing score as an example, the data show that the mean externalizing score for the treatment group at the Wave 2 data collection (Month 18) was 57.96, and the mean score for the nontreatment group at the Wave 2 was 56.92. The unadjusted mean difference between groups was 1.04, meaning that the externalizing score for the treatment group was 1.04 units greater (or worse) than that for the nontreatment group. Using an OLS regression to adjust for covariates (i.e., including all variables used in the treatment effect model, i.e., independent variables used in both the selection equation and the regression equation), the adjusted mean difference is -0.08 units; in other words, the treatment group is 0.08 units lower (or better) than the nontreatment group, and the difference is not statistically significant. These data suggest that the involvement of caregivers in substance abuse treatment has a negligible effect on child behavior. Alternatively, one might conclude that children whose parents are involved in treatment services do not differ from children whose parents are not referred to treatment. Given the high risk of children whose parents abuse substances, some might claim drug treatment to be successful. Now, however, consider a different analytic approach. The treatment effect model adjusts for heterogeneity of service participation by taking into consideration covariates affecting selection bias. The results show that at the follow-up data collection (Month 18), the treatment group was 8.6 units higher (or worse) than the nontreatment group ($p < .001$). This suggests that both the unadjusted mean difference (found by independent t test) and the adjusted mean difference (found above by regression) are biased because we did not control appropriately for selection bias. A similar pattern is observed for the internalizing score. The findings suggest that negative program impacts may be masked in simple mean differences and even in regression adjustment. 154 Table 4.4 Sample Description for the Study Evaluating the Impacts of Caregiver's Receipt of Substance Abuse Services on Child Developmental Well-Being 155 Source: Data from NSCAW, 2004. Notes: Reference group is shown next to the variable name. Variable name in parentheses is the actual name used in the programming syntax. AOD = alcohol or drug; CIDI-SF = Composite International Diagnostic Interview–Short Form. 156 Table 4.5 Differences in Psychological Outcomes Before and After Adjustments of Sample Selection Source: Data from NSCAW, 2004. a. Independent t tests on mean differences or t tests on regression coefficients show that none of these mean differences are statistically significant. ** $p < .01$, *** $p < .001$, two-tailed test. 4.4.2 Evaluation of Treatment Effects From a Program With a Group Randomization Design The Social and Character Development (SACD) program was jointly sponsored by the U.S. Department of Education and the Centers for Disease Control and Prevention. The SACD intervention project was designed to assess the impact of schoolwide social and character development education in elementary schools. Seven proposals to implement SACD were chosen through a peer review process, and each of the seven research teams implemented different SACD programs in elementary schools across the country. At each of the seven sites, schools were randomly assigned to receive either an intervention program or a control curriculum, and one cohort of students was followed from third grade (beginning in fall 2004) through fifth grade (ending in spring 2007). A total of 84 elementary schools were randomized to intervention and control at seven sites: Illinois (Chicago), New Jersey, New York (Buffalo, New York City, and Rochester), North Carolina, and Tennessee. Using site-specific data (as opposed to data collected across all seven sites), this example reports findings from an evaluation of the SACD program implemented in North Carolina (NC). The NC intervention was also known as the Competency Support Program, which included a skills-training curriculum, Making Choices, designed for elementary school students. The primary goal of the Making Choices curriculum was to increase students' social competence and reduce their aggressive behavior. During their third-grade year, the treatment group received 29 Making Choices classroom lessons and 8 follow157 up classroom lessons in each of the fourth and fifth grades. In addition, special in-service training for classroom teachers in intervention schools focused on the risks of peer rejection and social isolation, including poor academic outcomes and conduct problems. Throughout the school year, teachers received consultation and support (two times per month) in providing the Making Choices lessons designed to enhance children's social information-processing skills. In addition, teachers could request consultation on classroom behavior management and social dynamics. The investigators designed the Competency Support Program evaluation as a group randomization trial. The total number of schools participating in the study within a school district was determined in advance, and then schools were randomly assigned to treatment conditions within school districts; for each treated school, a school that best matched the treated school on academic yearly progress, percentage of minority students, and percentage of students receiving free or reduced-price lunch was selected as a control school (i.e., data collection only without receiving intervention). Over a 2-year period, this group randomization procedure resulted in a total of 14 schools (Cohort 1, 10 schools; Cohort 2, 4 schools) for the study: Seven received the Competency Support Program intervention, and seven received routine curricula. In this example, we focus on the 10 schools in Cohort 1. As it turned out—and as is often the case when implementing randomized experiments in social behavioral sciences—the group randomization did not work as planned. In some school districts, as few as four schools met the study criteria and were eligible for participation. When comparing data from the 10 schools, the investigators found that the intervention schools differed from the control schools in significant ways: The intervention schools had lower academic achievement scores on statewide tests (Adequate Yearly Progress [AYP]), a higher percentage of students of color, a higher percentage of students receiving free or reduced-price lunches, and lower mean scores on behavioral composite scales at baseline. These differences were statistically significant at the .05 level using bivariate tests and logistic regression models. The researchers were confronted with the failure of randomization. Had these selection effects not been taken into consideration, the evaluation of the program effectiveness would be biased. The evaluation used several composite scales that proved to have good psychometric properties. Scales from two well-established instruments were used for the evaluation: (1) the Carolina Child Checklist (CCC) and (2) the Interpersonal Competence Scale–Teacher (ICST). The CCC is a 35-item teacher questionnaire that yields factor scores on children's behavior, including social contact ($\alpha = .90$), cognitive concentration ($\alpha = .97$), social competence ($\alpha = .90$), and social aggression ($\alpha = .91$). The ICST is also a teacher questionnaire. It uses 18 items that yield factor scores on children's behavior, including aggression ($\alpha = .84$), academic competence ($\alpha = .74$), social 158 competence ($\alpha = .75$), internalizing behavior ($\alpha = .76$), and popularity ($\alpha = .78$). Table 4.6 presents information on the sample and results of the Heckit treatment effect model used to assess change scores in the fifth grade. The two outcome measures used in the treatment effect models included the ICST Social Competence Score and the CCC Prosocial Behavior Score, which is a subscale of CCC Social Competence. On both these measures, high scores indicate desirable behavior. The dependent variable employed in the treatment effect model was a change score—that is, a difference of an outcome variable (i.e., ICST Social Competence or CCC Prosocial Behavior) at the end of the spring semester of the fifth grade minus the score at the beginning of the fall semester of the fifth grade. Although "enterers" (students who transfer in) are included in the sample and did not have full exposure, most students in the intervention condition received Making Choices lessons during the third, fourth, and fifth grades. Thus, if the intervention was effective, then we would expect to observe a higher change (i.e., greater increase on the measured behavior) for the treated students than the control group students. Before evaluating the treatment effects revealed by the models, we need to highlight an important methodological issue demonstrated by this example: the control of clustering effects using the Huber-White sandwich estimator of variance. As noted earlier, the Competency Support Program implemented in North Carolina used a group randomization design. As such, students were nested within schools, and students within the same school tended to exhibit similar behavior on outcomes. When analyzing this type of nested data, the analyst can use the option of robust cluster (*) in treatreg to obtain an estimation of robust standard error for each coefficient. The Huber-White estimator only corrects standard errors and does not change the estimation of regression coefficients. Thus, in Table 4.6, we present one column for the "Coefficient," along with two columns of estimated standard errors: one under the heading of "SE" that was estimated by the regular specification of treatreg and the other under the heading of "Robust SE" that was estimated by the robust estimation of treatreg. Syntax that we used to create this analysis specifying control of clustering effect is shown in a note to Table 4.6. As Table 4.6 shows, the estimates of "Robust SE" are different from those of "SE," which indicates the importance of controlling for the clustering effects. As a consequence of adjusting for clustering, conclusions of significance testing using "Robust SE" are different from those using "SE." Indeed, many covariates included in the selection equation are significant under "Robust SE" but not under "SE." In the following discussion, we focus on "Robust SE" to explore our findings. The main evaluation findings shown in Table 4.6 are summarized next. First, selection bias appears to have been a serious problem because many variables included in the selection equation were statistically significant. We now use the 159 analysis of the ICST Social Competence score as an example. All school-level variables (i.e., school AYP composite test score, school's percentage of minority students, school's percentage of students receiving free lunch, and school's pupil-to-teacher ratio) in 2005 (i.e., the year shortly after the intervention was completed) distinguished the treatment schools from the control schools. Students' race and ethnicity compositions were also

different between the two groups, meaning that the African American, Hispanic, and White students are less likely than other students to receive treatment. The sign of the primary caregiver's education variable in the selection equation was positive, which indicated that primary caregivers of students from the intervention group had higher education than their control group counterparts ($p < .001$). In addition, primary caregivers of the treated students were less likely to have been employed full-time than were their control group counterparts. All behavioral outcomes at baseline were statistically different between the two groups, which indicated that treated students were rated as more aggressive ($p < .001$), had higher academic competence scores ($p < .01$), exhibited more problematic scores on internalizing behavior ($p < .001$), demonstrated lower levels of cognitive concentration ($p < .001$), displayed lower levels of social contact with prosocial peers ($p < .001$), and showed higher levels of relational aggression ($p < .001$). It is clear that without controlling for these selection effects, the intervention effect would be severely biased. Table 4.6 Estimated Treatment Effect Models of Fifth Graders' Change on ICST Social Competence Score and on CCC Prosocial Behavior Score 160 161 Source: Data from SACD, 2008. Notes: Syntax to create the results of estimates with robust standard errors for the "Change on ICST Social Competence": Second, we also included students' demographic variables and caregivers' characteristics in the regression equation based on the consideration that they were covariates of the outcome variable. This is an example of using some of the covariates of the selection equation in the regression equation (i.e., the x vector is part of the z vector, as described in Section 4.2). Results show that none of these variables was significant. Third, our results indicated that the treated students had a mean increase in ICST Social Competence in the fifth grade that was 0.17 units higher than that of the control students ($p < 0.1$) and a mean increase in CCC Prosocial Behavior in the fifth grade that was 0.20 units higher than that of the control students ($p < .01$). Both results are average treatment effects of the sample that can be generalized to the population, although the difference on ICST Social Competence only approached significance ($p < .10$). The data showed that the Competency Support Program produced positive changes in students' social competence, which was consistent with the study's focus on social information processing skills. Had the study analysis not used the Heckit treatment effect model, the intervention effects would have been biased and inconsistent. An independent sample t test confirmed that the mean differences on both change scores were statistically significant at a .000 level, with inflated mean 162 differences. The t test showed that the intervention group had a mean change score on ICST Social Competence that was 0.25 units higher than the control group (instead of 0.17 units higher as shown by the treatment effect model) and a mean change score on CCC Prosocial Behavior that was 0.26 units higher than the control group (instead of 0.20 units higher as shown by the treatment effect model). Finally, the null hypothesis of zero p , or zero correlation between the errors of the selection equation and the regression equation, was rejected at a significance level of .05 for the ICST Social Competence model, but it was not rejected for the CCC Prosocial Behavior model. This indicates that the assumption of nonzero ρ may be violated by the CCC Prosocial Behavior model. It suggests that the selection equation of the CCC Prosocial Behavior model may not be adequate, a topic we address in Chapter 11. 4.4.3 Running the Treatment Effect Model After Multiple Imputations of Missing Data Missing data are a ubiquitous problem in research, and missing values represent a serious threat to the validity of inferences drawn from findings. Increasingly, social science researchers are turning to multiple imputation to handle missing data. Multiple imputation, in which missing values are replaced by values repeatedly drawn from conditional probability distributions, is an appropriate method for handling missing data when values are not missing completely at random (R. A. Little & Rubin, 2002; Rubin, 1996; Schafer, 1997). The following example illustrates how to analyze a treatment effect model based on multiply imputed data sets after missing data imputation using Rubin's rule for inference of imputed data. Given that this book is not focused on missing data imputation, we ignore the description about methods of multiple imputation. Readers are directed to the references mentioned earlier to find full discussion of multiple imputation. In this example, we attempt to show the method analyzing the treatment effect model based on multiply imputed data sets to generate a combined estimation of treatment within Stata. The Stata programs we recommend to fulfill this task are called `mim` and `mimstack`; both were created by John C. Galati at the U.K. Medical Research Council and Patrick Royston at the Clinical Epidemiology and Biostatistics Unit, the United Kingdom (Galati, Royston, & Carlin, 2009). The `mimstack` command is used for stacking a multiply imputed data set into the format required by `mim`, and `mim` is a prefix command for working with multiply imputed data sets to estimate the required model such as `treatreg`. The commands to conduct a combined `treatreg` analysis look like the following: `mimstack, m(#) sortorder(varlist) istub(string) [nomj0 clear] 163 mim, cat(fit): treatreg depvar [indepvars], treat(depvar_t = indepvars_t), where m specifies the number of imputed data sets; sortorder specifies a list of one or more variables that uniquely identify the observations in each of the data sets to be stacked; istub specifies the filename of the imputed data files to be stacked with the name specified in string; nomj0 specifies that the original nonimputed data are not to be stacked with the imputed data sets; clear allows the current data set to be discarded; mim, cat(fit) informs that the program to be estimated is a regression-type model; and treatreg and its following commands are specifications one runs based on a single data set (i.e., data file without multiple imputation). For the example depicted in Section 4.4.2, we had missing data on most independent variables. Using multiple imputation, we generated 50 imputed data files. Analysis shows that with 50 data sets, the imputation achieved a relative efficiency of 99%. The syntax to run a treatreg model analyzing outcome variable CCC Social Competence change score ccscomch using 50 data files is shown in the lower panel of Table 4.7. In this mimstack command, id is the ID number used in all 50 files that uniquely identifies observations within each data set, g3scom is the commonportion name of the 50 files (i.e., the 50 imputed data files are named as g3scom1, g3scom2, . . . , and g3scom50), nomj0 indicates that the original nonimputed data set was not used, and clear allows the program to discard the current data set once estimation of the current model is completed. In the above mim command, cat(fit) informs Stata that the combined analysis (i.e., treatreg) is a regression-type model; treatreg specifies the treatment effect model as usual, where the outcome variable for the regression equation is ccscomch; the independent variables for the regression equation are agec, female, black, white, hisp, pcedu, ipovl, pcemft, and fthr; the treatment membership variable is intrl; and the independent variables included in the selection equation are agec, female, black, white, hisp, pcedu, ipovl, pcemft, fthr, discaca2, and discint2. The treatreg model also estimates robust standard error to control for clustering effect where the variable identifying clusters is schbl. Table 4.7 Exhibit of Combined Analysis of Treatment Effect Models Based on Multiple Imputed Data Files 164 Source: Data from SACD, 2008. Table 4.7 is an exhibition of the combined analysis invoked by the above commands. Results of the combined analysis are generally similar to those produced by a single-file analysis, but with an important difference: The combined analysis does not provide ρ , σ , and λ but instead shows $\hat{\alpha}$ and $\hat{\lambda}$ based on 50 files. Users can obtain ρ from the formula showing the relationship between ρ and $\hat{\alpha}$, and obtain σ by taking exponent of $\hat{\lambda}$. Alternatively, users may examine ρ , σ , and λ by checking individual files to assess these statistics, particularly if these statistics are consistent across files. If the user does not find a consistent pattern of these statistics across files, then the user will need to further investigate relations between the imputed data and the treatment effect model. 165 4.5 CONCLUSION In 2000, the Nobel Prize Review Committee named James Heckman as a corecipient of the Nobel Prize in Economics in recognition of "his development of theory and methods for analyzing selective samples" (Nobel Prize Review Committee, 2000). This chapter reviews basic features of the Heckman sample selection model and its related models (i.e., the treatment effect model and running the analysis with multiply imputed data files to handle the missing data problem). The Heckman model was invented at approximately the same time that statisticians started to develop the propensity score matching models, which we will examine in the next chapter. The Heckman model emphasizes modeling selection bias rather than assuming mechanisms of randomization work to balance data between treated and control groups. Heckman's sample selection model shares an important feature with the propensity score matching model: It uses a two-step procedure to model the selection process first and then uses the conditional probability of receiving treatment to control for bias induced by selection in the outcome analysis. Results show that the Heckman model, particularly its revised version called the treatment effect model, is useful in producing improved estimates of average treatment effects, especially when the causes of selection processes are known and are correctly specified in the selection equation. We conclude this chapter with a caveat or, perhaps, a caution. The Heckman treatment effect model appears to be sensitive to model "misspecification." It is well established that when a Heckit model is misspecified (i.e., when the predictor or independent variables are incorrect or omitted), particularly when important variables causing selection bias are not included in the selection equation, and when the estimated correlation between errors of the selection equation and the regression equation (i.e., the estimated ρ) is zero, then results of the treatment effect model are biased. The Stata Reference Manual (StataCorp, 2003) correctly states that the Heckman selection model depends strongly on the model being correct; much more so than ordinary regression. Running a separate probit or logit for sample inclusion followed by a regression, referred to in the literature as the two-part model (Manning, Duan, & Rogers, 1987)—not to be confused with Heckman's two-step procedure—is an especially attractive alternative if the regression part of the model arose because of taking a logarithm of zero values. (p. 70) Kennedy (2003) argues that the Heckman two-stage model is inferior to the selection model or treatment effect model using maximum likelihood because the two-stage estimator is inefficient. He also warns that in solving the omittedvariable problem, the Heckman procedure introduces a measurement error 166 problem, because an estimate of the expected value of the error term is employed in the second stage. Finally, it is not clear whether the Heckman procedure can be recommended for small samples. In practice, there is no definite procedure to test conditions under which the assumptions of the Heckman model are violated. As a consequence, sensitivity analysis is recommended to assess the stability of findings under the stress of alternative violations of assumptions. In Chapter 11, we present results of a Monte Carlo study that underscore this point. The Monte Carlo study shows that the Heckit treatment effect model works better than other approaches when ρ is indeed nonzero, and it works worse than other approaches when ρ is zero. NOTE 1. The relation between $\hat{\alpha}$ and ρ is as follows: using data of Table 4.1. 167 CHAPTER 5 Propensity Score Matching and Related Models This chapter describes two propensity score matching methods, greedy matching and optimal matching. Both models stem from Rosenbaum and Rubin's (1983) seminal work that defined a propensity score as the conditional probability of assignment to a particular treatment given a vector of observed covariates. When applied appropriately, these models can help solve the problem of selection bias and provide valid estimates of average treatment effects (ATEs). In this chapter, we also introduce recent advances in propensity score matching. We describe generalized boosted regression, an application used in estimating propensity scores. In addition, conducting propensity score matching with multilevel data is explored. It requires additional effort because when data are nested, controls for clustering are needed in both the model predicting propensities and the outcome model following matching. Section 5.1 provides an overview of propensity score models. The overview conceptualizes the modeling process as a one-step, two-step, or three-step sequenced analysis. Section 5.2 reviews key propositions and corollaries derived and proved by Rosenbaum and Rubin (1983, 1984, 1985). The purpose of this review is to address two key questions: (1) How do propensity score models balance data? and (2) How do propensity score models solve the problem of dimensionality that plagued the classical matching algorithm? Section 5.3 focuses on Step 1 of the analysis: the specification of a logistic regression model and the search for a best set of conditioning variables that optimizes estimates of propensity scores. We review the procedure of generalized boosted regression in this section. Section 5.4 focuses on the second analytic step, that is, resampling based on estimated propensity scores. We then review various types of matching algorithms, including greedy matching and optimal matching. Section 5.5 focuses on the third step: postmatching analysis. We review various methods that follow matching, including the general procedures for multivariate analysis following greedy matching, procedures for conducting the Hodges-Lehmann aligned rank test of the treatment effect following optimal full matching or optimal variable matching, and regression adjustment (i.e., regressing difference scores of outcome on difference scores of 168 covariates) following optimal pair matching. Section 5.6 discusses issues pertaining to multilevel data, including models predicting propensity scores with or without random effects, and postmatching analysis that controls for clustering effects formed by the original sampling or by matching. Section 5.7 summarizes key features of computing software packages that can be used to run most models described in the chapter. Section 5.8 presents examples of selected models. Section 5.9 summarizes and concludes the discussion. 5.1 OVERVIEW In 1983, Rosenbaum and Rubin published a seminal paper on propensity score analysis. That paper articulated the theory and application principles for a variety of propensity score models. Ever since this work, the propensity score method has grown at a rapid pace and moved in various directions for refinement. New models, such as the application of generalized boosted regression, have been developed to refine logistic regression and, in turn, to refine estimation of propensity scores. Other innovations include new models developed to refine matching algorithms and include optimal matching that applies developments in operations research (i.e., network flow theory) to matching. Additional new models, such as propensity score weighting, have been developed to combine propensity scores and conventional statistical methods. Novice users of propensity score methods are often puzzled by new terminologies and seemingly different techniques. It is easy to get lost when first encountering the propensity score literature. To chart the way, we summarize the process of propensity score analysis presented in Figure 5.1 as a one-step, two-step, or three-step sequence. Because the sample selection and treatment effect models described in Chapter 4 are very different from other models, Figure 5.1 only covers propensity score models presented in Chapters 5 to 9. We first examine propensity score matching as a three-step analytic process. Step 1a: Seek the best conditioning variables or covariates that are speculated to be causing an imbalance between treated and control groups. A rigorous propensity score modeling always begins with estimation of the conditional probability of receiving treatment. Data analysts fulfill the task by estimating a logistic regression model (or similar model such as probit regression or discriminant analysis) to analyze binary levels of treatment or a multinomial logit model to analyze the effects of multiple doses of treatment. Because the binary logistic regression is often used for estimating propensity scores, we show that method as Step 1a in Figure 5.1. The objective of analysis at this stage is to identify the observed covariates affecting selection bias and further to specify a functional form of the covariates for the propensity score model. Ideally, we seek an optimal estimate of propensity scores. By definition, a propensity score is a conditional probability of a study participant receiving treatment given observed covariates; hence, not only treated participants but also control participants may have nonzero propensity scores. More precisely, the propensity score is a balancing score representing a vector of covariates. In this context, a pair of treated and control participants sharing a similar propensity score are essentially viewed as comparable, even though they may differ on values of specific covariates. Figure 5.1 General Procedure for Propensity Score Analysis Step 2a: Matching or resampling. Having obtained the balancing scores (i.e., 170 the propensities), the analyst then uses the scores to match treated participants with control participants. The advantage of using the single propensity score is that it allows us to solve the problem of failure in matching based on multiple covariates. Because the common support region formed by the estimated propensity scores does not always cover the whole range of study participants, you might not find matched controls for some treated participants, and some control participants may never be used; thus, matching typically leads to a loss of study participants. Because of this characteristic, matching is referred to as resampling. Although the original sample is not balanced on observed covariates between treatment and control conditions, the resample based on propensity scores balances treatment and controls for selection bias on observed measures. The key objective at this stage is to make the two groups of participants as much alike as possible in terms of estimated propensity scores. Various algorithms have been developed to match participants with similar propensity scores. These include greedy matching, Mahalanobis metric distance matching with or without propensity scores, and optimal matching. These algorithms differentially deal with the loss of participants whose propensity scores may be so extreme as to make matching difficult. Step 3a: Postmatching analysis based on the matched samples. In principle, the new sample developed in Step 2a corrects for selection bias (on observed covariates) and violations of statistical assumptions that are embedded in multivariate models (such as the assumption embedded in regression about independence between an independent variable and the regression equation's error term). With this matched new sample, the analyst can perform multivariate analysis as is normally done using a randomized experiment. However, most multivariate analyses are permissible only for matched samples created by greedy matching. With matched samples created by optimal matching, special types of analyses are needed. These include a special type of regression adjustment (i.e., regressing difference scores of outcomes between treated and control participants on difference scores of covariates) for a sample created by optimal pair matching, a special type of regression adjustment (i.e., regressing difference scores of the Hodges-Lehmann aligned rank of the outcome variable on difference scores of the aligned rank of covariates) for a sample created by either optimal variable matching or optimal full matching, or a test of the ATE using the Hodges-Lehmann aligned rank statistic for samples created by optimal full or variable matching. Step 3b: Postmatching analysis using stratification of the propensity scores. Researchers could also perform stratification of the estimated propensity scores without conducting multivariate modeling; this stratification would also be conducted in a way similar to that in which researchers analyze treatment effects with data generated by randomized experiments, that is, by comparing the mean 171 difference of an outcome between treatment and control conditions within a stratum and then generating a mean and variance for the overall sample to gauge the sample ATE and its statistical significance. Propensity score stratification/subclassification can be performed with or without matching. The method of subclassification without matching is described below as Step 2c in the two-step process. Chapter 6 describes the method of propensity score subclassification. As indicated in Figure 5.1, propensity score analysis is also used in two-step analytic processes. Models of this type use almost identical methods to estimate propensity scores and share the same Step 1a features as the three-step models. But the two-step models skip resampling (i.e., matching). They use propensity scores in a different way. For two-step models, the main features of Step 2 are shown next. Step 2b: Multivariate analysis using propensity scores as sampling weights. As indicated earlier, this method (Hirano & Imbens, 2001; Robins & Rotnitzky, 1995; Rosenbaum, 1987) does not resample the data and, therefore, avoids undesirable loss of study participants. Use of propensity scores as weights is analogous to the reweighting procedures used in survey sampling, where adjustments are made for observations on the basis of the probabilities for inclusion in a sample (McCaffrey et al., 2004). Propensity score weighting not only overcomes the problem of loss of sample participants but also offers two kinds of estimates for treatment effects: the ATE and the ATT (i.e., average treatment effect for the treated). To be precise, this method is not matching, although it originated from the same idea of using propensity scores to control for selection bias. We describe this method in Chapter 7. Step 2c: Multivariate analysis in conjunction with propensity score stratification (subclassification). This method is similar to that of Step 3b, except that in the current setting, researchers directly do a multivariate outcome analysis without using a matching procedure (Rosenbaum & Rubin, 1983, 1984). Using quintiles, researchers first stratify the sample into five subclasses and conduct a multivariate outcome analysis for each subclass. Aggregating estimated treatment effects across all five subclasses and performing a significance test, researchers finally obtain an estimate of the average treatment effect for the entire sample and discern whether such a treatment effect is statistically significant. Other percentiles of estimated propensity scores may be used to create more than five subclasses. We describe this method in Chapter 6. Step 2d: Analysis of weighted mean differences using kernel or local linear regression (i.e., the kernel-based matching estimator developed by Heckman, Ichimura, and Todd [1997, 1998]). This method conducts a "latent" matching and combines weighting and outcome analysis into one step using nonparametric 172 regression (i.e., either a tricubic kernel smoothing technique or a local linear regression). Given that this method is categorically different from Rosenbaum and Rubin's models, we describe kernel-based matching in Chapter 9. Also indicated by Figure 5.1, propensity score analysis may be performed in one step (Step 1b), and the method is known as matching estimators (Abadie & Imbens, 2002, 2006). Although this method does not use a logistic regression to estimate propensity scores, it estimates a one-dimensional score by using vector norm. The method combines the estimation of balancing scores, matching, and outcome analysis all into one step. We describe this method in Chapter 8. On the basis of this overall process, we now move to the details to review key statistical theories, modeling principles, practice problems, and solutions for each step of the propensity score analysis approach. This chapter first focuses on the estimation of propensity scores, propensity score greedy matching, and propensity score optimal matching. 5.2 THE PROBLEM OF DIMENSIONALITY AND THE PROPERTIES OF PROPENSITY SCORES With complete data, Rosenbaum and Rubin (1983) defined the propensity score for participant i ($i = 1, \dots, N$) as the conditional probability of assignment to a particular treatment ($W_i = 1$) versus nontreatment ($W_i = 0$) given a vector of observed covariates, x_i . The advantage of the propensity score in matching, stratification, and weighting is its reduction of dimensions: The vector x may include many covariates, which represent many dimensions, and the propensity approach reduces all this dimensionality to a one-dimensional score. In conventional matching, as the number of matching variables increases, the researcher is challenged by the difficulty of finding a good match from the control group for a given treated participant. Rosenbaum (2002b) illustrated this with p covariates: Even if each covariate is a binary variable, there will be 2^p possible values of x . Suppose $p = 20$ covariates, then $2^{20} = 1,048,576$, or more than a million possible values of x . With a sample of hundreds or even thousands of participants, it is likely that many participants will have unique values of x and, therefore, can neither find matches from the control condition nor be used as a match for any treated case. Exact matching in this context often results in dropping cases and, in the presence of a large number of covariates or exceptional variation, may become infeasible. The propensity score $e(x_i)$ is a balancing measure (so called the coarsest 173 score) that summarizes the information of vector x_i in which each x covariate is a finest score. Rosenbaum and Rubin (1983) derived and proved a series of theorems and corollaries showing the properties of propensity scores. The most important property is that a coarsest score can sufficiently balance differences observed in the finest scores between treated and control participants. From Rosenbaum and Rubin (1983) and Rosenbaum (2002b), the properties of propensity scores include the following: 1. Propensity scores balance observed differences between treated and control participants in the sample. Rosenbaum (2002b, p. 298) showed that a treated and control participant with the same value of the propensity score have the same distribution of the observed covariate X . This means that in a stratum or matched set that is homogeneous on the propensity score, treated and control participants may have differing values for X (i.e., if two participants have the same propensity score, they still could differ on an observed covariate such as gender, if gender—the finest score—is included in the X vector), but the differences will be chance differences rather than systematic differences. 2. Treatment assignment and the observed covariates are conditionally independent given the propensity score; that is, This property links the propensity score to the assumption regarding strongly ignorable treatment assignment. In other words, conditional on the propensity score, the covariates may be considered independent of assignment to treatment. Therefore, for observations with the same propensity score, the distribution of covariates should be the same across the treated and control groups. Furthermore, this property means that, conditional on the propensity score, each participant has the same probability of assignment to treatment, as in a randomized experiment. 3. If the strongly ignorable treatment assignment assumption holds and $e(x_i)$ is a balancing score, then the expected difference in observed responses to the two treatment conditions at $e(x_i)$ is equal to the ATE at $e(x_i)$. This property links the propensity score model to the counterfactual framework and shows how the problem of not observing outcomes for the treated participants under the control condition (a problem discussed in Section 2.2) can be resolved. It follows that the mean difference of the outcome variable between treated and control participants for all units with the same value of the propensity score is an unbiased estimate of the ATE at that propensity score. That is, 174 4. Rosenbaum and Rubin (1983, p. 46) derived corollaries to justify three key approaches using the propensity scores. These corollaries form the foundation for all models described in Chapters 5 to 7. a. Pair matching: The expected difference in responses of treatment and control units in a matched pair with same value of propensity score $e(x)$ equals the ATE at $e(x)$, and the mean of matched pair differences obtained by this two-step sampling process is unbiased for the ATE $\tau = E(Y_1|W = 1) - E(Y_0|W = 0) = E(Y_1 - Y_0|e(x))$. b. Subclassification of propensity scores that all units within a stratum have the same $e(x)$ and at least one unit in the stratum receives each treatment condition—the expected difference in treatment mean equals the ATE at that value of $e(x)$, and the weighted average of such differences is unbiased for the treatment effect $\tau = E(Y_1|W = 1) - E(Y_0|W = 0)$. c. Covariance adjustment: Assume that the treatment assignment is strongly ignorable at the balancing score $e(x)$, and the conditional expectation of Y_i ($i = 0, 1$) is linear: $E(Y_i|W = i, e(x)) = \alpha + \beta e(x)$, then the estimator is conditionally unbiased given $e(x_i)$ ($i = 1, \dots, n$) for the treatment effect at $e(x)$ —namely, $E(Y_1 - Y_0|e(x))$, if and are conditionally unbiased estimators of α and β , such as least squares estimators. In the preceding descriptions, the propensity score $e(x)$ is defined as a predicted probability of receiving treatment in a sample where treatment assignment is nonignorable. Typically, this value is a predicted probability saved from an estimated logistic regression model. In practice, Rosenbaum and Rubin (1985) suggested using the logit of the predicted probability as a propensity score (i.e., because the distribution of approximates to normal. Note that in the literature, the quantity is also called differs from as given by the a estimated propensity score, although previous equation. In this chapter and elsewhere, except when explicitly noted, as the propensity score. We will follow the convention of referring to Readers should keep in mind that, in practice, a logit transformation of (i.e.,) may be used, and has distributional properties that may make it more desirable than . These theorems and corollaries established the foundation of the propensity score matching approach and many related procedures. For instance, the property of x_i w/o $e(x_i)$ leads to a procedure often used to check whether 175 estimated propensity scores successfully remove imbalance on observed covariates between treated and control groups. The procedure is as follows: (a) The analyst conducts a bivariate test (i.e., a Wilcoxon rank sum test—also known as the Mann-Whitney two-sample statistic, an independent sample t test, or a one-way analysis of variance [ANOVA], for a continuous covariate, or a chi-square test for a categorical covariate) using treatment condition as a grouping variable before matching; if the bivariate test shows significant difference between treated and control groups on a covariate, then the analyst needs to control the covariate by including the covariate in the model estimating propensity scores. (b) After matching on the propensity scores, the analyst performs similar bivariate tests with some adjustments (e.g., instead of using a Wilcoxon rank sum test, the analyst may use a Wilcoxon matched pairs signed rank test or absolute standardized difference in covariate means); if the postmatching bivariate tests are nonsignificant, then we may conclude that the propensity score has successfully removed group differences on the observed covariates. (c) If the postmatching bivariate tests show significant differences, the model predicting propensity scores should be reconfigured and rerun until the matching successfully removes all significant imbalances. Alternatively, the analyst can employ the normalized difference score ΔX introduced in Chapter 1 (i.e., Equation 1.1) to conduct imbalance checks before and after matching, where ΔX exceeding .25 is an indication of imbalance of X between groups. Implied in our use above, the covariates in the vector x are called conditioning or matching variables. A correct specification of covariates in the Step 1 model is crucial to the propensity score approach because the final estimation of the treatment effect is sensitive to this specification (Rubin, 1997). Many studies show that the choice of conditioning variables can make a substantial difference in the overall performance of propensity score analysis (e.g., Heckman, Ichimura, Smith, & Todd, 1998; Heckman, Ichimura, & Todd, 1997; Lechner, 2000). A correct specification of the conditioning model predicting propensity scores has two aspects: One is to include the correct variables in the model; that is, researchers should include important covariates that have theoretical relevance. To do this, we typically rely on substantive information and prior studies about predictors of receiving treatment. The other is to specify correctly the functional form of conditioning variables. This may involve introducing polynomial and interaction terms. The dilemma the analyst faces is that there is no definitive procedure or test available to provide guidance in specifying a best propensity score model, and theory often provides weak guidance as to how to choose and configure conditioning variables (Smith & Todd, 2005). This dilemma prompted the development of promising methods such as generalized boosted regression for searching for optimal propensity scores. The next section describes the development of propensity scores and strategies for dealing with 176 the specification problem. 5.3 ESTIMATING PROPENSITY SCORES As mentioned earlier, several methods for estimating the conditional probability of receiving treatment using a vector of observed covariates are available. These methods include logistic regression, the probit model, and discriminant analysis. Of these methods, this book describes only logistic regression because it is the prevailing approach. A closely related method is Mahalanobis metric distance, which was invented prior to methods for propensity score matching (Cochran & Rubin, 1973; Rubin, 1976, 1979, 1980a). A Mahalanobis metric distance per se is not a model-based propensity score. However, the Mahalanobis metric distance serves a similar function as a propensity score and is an important statistic used in greedy matching, optimal matching, and multivariate matching. Accordingly, we examine the Mahalanobis metric distance in Sections 5.4.1 and in Chapter 8. 5.3.1 Binary Logistic Regression The conditional probability of receiving treatment when there are two treatment conditions (i.e., treatment vs. control) is estimated using binary logistic regression. Denoting the binary treatment condition as W_i ($W_i = 1$, if a study case is in the treatment condition, and $W_i = 0$, if the case is in the control condition) for the i th case ($i = 1, \dots, N$), the vector of conditioning variables as X_i , and the vector of regression parameters as β , a binary logistic regression depicts the conditional probability of receiving treatment as follows: This is a nonlinear model, meaning that the dependent variable W_i is not a linear function of the vector of conditioning variables x_i . However, by using an appropriate link function such as a logit function, we can express the model as a generalized linear model (McCullagh & Nelder, 1989). Although W_i is not a linear function of x_i , its transformed variable through the logit function (i.e., the natural logarithm of odds or $\text{loge}\{P(W_i)/[1 - P(W_i)]\}$) becomes a linear function of x_i : where P denotes $P(W_i)$. Model 5.1 is estimated with the maximum likelihood estimator. To ease the 177 exposition, we now assume that there are only two conditioning variables, x_1 and x_2 . The log-likelihood function of Model 5.1 with two conditioning variables can be expressed as follows: The partial derivative of $\log l$ with respect to β maximizes the likelihood function. In practice, the problems are seldom solved analytically, and we often rely on a numerical procedure to find estimates of β . Long (1997, pp. 56–57) described three numerical estimators: the Newton-Raphson method, the scoring method, and the B-triple-H (BHHH) method. Typically, a numerical method involves the following steps: (a) Insert starting values (i.e., "guesses") of β_0 , β_1 , and β_2 in the right-hand side of Equation 5.2 to obtain a first guess of $\log l$. (b) Insert a different set of β_0 , β_1 , and β_2 into the right-hand side equation to obtain a second guess of $\log l$; by comparing the new $\log l$ with the old one, the analyst knows the direction for trying the next set of β_0 , β_1 , and β_2 . The process from Step (a) to Step (b) is called an iteration. (c) Replicate the preceding process several times (i.e., running several iterations) until the largest value of $\log l$ is obtained (i.e., the maximum log likelihood function) or until the difference in $\log l$ between two iterations is no longer greater than a predetermined criterion value, such as 0.000001. Estimated values of β_0 , β_1 , and β_2 (i.e., and β) are logistic regression coefficients at which the likelihood of reproducing sample observations is maximized. Using these estimated coefficients and applying Equation 5.1 (i.e., and), the analyst obtains the predicted replacing β_0 , β_1 , and β_2 with probability of receiving treatment (i.e., estimated propensity score) for each sample participant i . As in running ordinary least squares (OLS) regression or other multivariate models, we must be sensitive to the nature of the data at hand and the possibility of violations of assumptions. Routine diagnostic analyses, such as tests of multicollinearity, tests of influential observations, and sensitivity analyses should be used to assess the fit of the final model to the data. A number of statistics have been developed to assess the goodness of fit of the model. Unfortunately, none of these statistics indicates whether the estimated propensity scores are representative of the true propensity scores. Notwithstanding, meeting requirements embedded in these fit statistics is a minimum requirement or starting point. Details of goodness-of-fit indices for the logistic regression model can be found in textbooks on logistic regression or limited dependent variable analysis (e.g., Kutner, Nachtsheim, & Neter, 2004; Long, 1997). Here, we summarize a 178 few indices and include cautionary statements for their use. 1. Pearson chi-square goodness-of-fit test: This test detects major departures from a logistic response function. Large values of the test statistic (i.e., those associated with a small or significant p value) indicate that the logistic response function is not appropriate. However, it is important to note that the test is not sensitive to small departures (Kutner et al., 2004). 2. Chi-square test of all coefficients: This test is a likelihood ratio test and analogous to the F test for linear regression models. We can perform a chi-square test using the log-likelihood ratio, as follows: Model chi-square = 2 log likelihood of the full model – 2 log likelihood of the model with intercept only. If the model chi-square $> \chi^2(1 - \alpha, df = \text{number of conditioning variables})$, then we reject the null hypothesis stating that all coefficients except the intercept are equal to zero. As a test of models estimated by the maximum likelihood approach, a large sample is required to perform the likelihood ratio test, and this test is problematic when the sample is small. 3. Hosmer-Lemeshow goodness-of-fit test: This test first classifies the sample into small groups (e.g., g groups) and then calculates a test statistic using the Pearson chi-squares from the $2 \times g$ tables of observed and estimated expected frequencies. A test statistic that is less than $\chi^2(1 - \alpha, df = g - 2)$ indicates a good model fit. The Hosmer-Lemeshow test is sensitive to sample size. That is, in the process of reducing the data through grouping, we may miss an important deviation from fit due to a small number of individual data points. Hence, we advocate that, before concluding that a model fits, an analysis of the individual residuals and relevant diagnostic statistics be performed (Hosmer & Lemeshow, 1989, p. 144). 4. Pseudo R2: Because the logistic regression model is estimated by a non-least squares estimator, the common linear measure of the proportion of the variation in the dependent variable that is explained by the predictor variables (i.e., the coefficient of determination R2) is not available. However, several pseudo R2s for the logistic regression model have been developed by analogy to the formula defining R2 for the linear regression model. These pseudo R2s include Efron's, McFadden's, adjusted McFadden's, Cox and Snell's, Nagelkerke/Cragg and Uhler's, McKelvey and Zavoina's, count R2, and adjusted count R2. In general, a higher value in a pseudo R2 indicates a better fit. However, researchers should be aware of several limitations of pseudo R2 measures and interpret their findings with caution. The UCLA Academic Technology Services (2008) provides a 179 detailed description of each of these pseudo R2s and concludes as follows: Pseudo R-squares cannot be interpreted independently or compared across datasets: They are valid and useful in evaluating multiple models predicting the same outcome on the same dataset. In other words, a pseudo R-squared statistic without context has little meaning. A pseudo R-square only has meaning when compared to another pseudo R-square of the same type, on the same data, predicting the same outcome. In this situation, the higher pseudo R-square indicates which model better predicts the outcome. Given these basic concepts in logistic regression, how do we optimize the estimates of propensity scores in the context of observational studies? It is worth underscoring a key point mentioned previously: A good logistic regression model that meets routine requirements and standards is a necessary but insufficient condition for arriving at the best propensity scores. 5.3.2 Strategies to Specify a Correct Model—Predicting Propensity Scores A principal question then arises: What defines the "best" logistic regression? The answer is simple: We need propensity scores that balance the two groups on the observed covariates. By this criterion, a best logistic regression is a model that leads to estimated propensity scores that best represent the true propensity scores. The challenge is that the true propensity scores are unknown, and therefore, we must seek methods to measure the fit between the estimated and the unknown true scores. The literature on propensity score matching is nearly unanimous in its emphasis`

on the importance of including in models carefully chosen and appropriate conditioning variables in the correct functional form. Simulation and replication studies have found that estimates of treatment effects are sensitive to different specifications of conditioning variables. For example, Smith and Todd (2005) found that using more conditioning variables may exacerbate the common support region problem—an issue we describe in detail later. In general, a sound logistic regression model should minimize the overall sample prediction error. That is, it should minimize the overall sample difference between the observed proportion of $W_i = 1$ and $P(W_i = 1)$. McCaffrey et al. (2004) developed a procedure using generalized boosted modeling (GBM) that seeks the best balance of the two groups on observed covariates. This procedure is described in Section 5.3.4. The algorithm McCaffrey et al. invoked altered the GBM criterion in such a way that iterations stop only when the sample average standardized absolute mean difference (ASAM) in the covariates is minimized. Suffice it to say that a best logistic regression model should take covariate balance into consideration, and in practice outside of propensity score estimation, this may or may not be a crucial concern in running 180 logistic regression. Strategies for fitting the propensity score model are summarized in the following: 1. Rosenbaum and Rubin (1984, 1985) described a procedure that used higher-order polynomial terms and/or cross-product interaction terms in the logistic regression through repeatedly executing the following tasks: running logistic regression, matching, bivariate tests of covariate balances based on the matched data, and rerunning logistic regression if covariate imbalances remain. As mentioned earlier, it is a common practice to run bivariate analysis before and after matching. Using a Wilcoxon rank sum (Mann-Whitney) test, t test, chi-square, or other bivariate method, we test whether the treated and control groups differ on covariates included in the logistic regression. The propensity score matching approach aims to achieve approximately the same distribution of each covariate between the two groups. Nonetheless, even after matching, significant differences may remain between groups. When these differences exist, the propensity score model can be reformulated, or the analyst can conclude that the covariate distributions did not overlap sufficiently to allow the subsequent analysis to adjust for these covariates (Rubin, 1997). In rerunning the propensity score model, we may include either a square term of the covariate that shows significance after matching or a product of two covariates if the correlation between these two covariates is likely to differ between the groups (Rosenbaum & Rubin, 1984). 2. Rosenbaum and Rubin (1984) further recommended applying stepwise logistic regression to select variables. Note that data-driven approaches determine the inclusion or exclusion of conditioning variables based on a Wald statistic (or t statistic) and its associated p value. Thus, the estimated model that results contains only those variables that are significant at a predetermined level. Rosenbaum (2002b) suggested a similar rule of thumb: The logistic regression model should include all pretreatment covariates whose group differences met a low threshold for significance, such as $|t| > 1.5$. 3. Eichler and Lechner (2002) used a variant of a measure suggested by Rosenbaum and Rubin (1985), which was based on standardized differences between the treatment and matched comparison groups in terms of means of each variable in x , squares of each variable in x , and first-order interaction terms between each pair of variables in x . 4. Dehejia and Wahba (1999) used a procedure similar to the method recommended by Rosenbaum and Rubin (1984), but they added stratification to determine the use of higher-order polynomial terms and interactions. They first fit a logistic regression model specifying main 181 effects only, and then they stratify the sample by estimated propensity scores. Based on the stratified sample, they test for differences in the means and standard deviations of covariates between treated and control groups within each stratum. If significant differences remain, they add higher-order polynomial terms and interactions. The process continues until no significant differences are observed within a stratum. 5. Hirano and Imbens (2001) developed a procedure that fully relies on a statistical criterion to seek conditioning predictors in logistic regression and then predictors in a follow-up outcome regression. Their search for predictor variables for both the logistic regression and the follow-up regression model is, to some extent, similar to the stepwise method, but it tests a range of models by using different cutoff values of the t statistic. Because this method is innovative, important, and deserving of scrutiny, we describe it in a separate section later. In sum, a careful selection of conditioning variables and a correct specification of the logistic regression are crucial to propensity score matching. Although scholars in the field have suggested a variety of rules and approaches, there is no definitive procedure of which we are aware. Because the selection of conditioning variables affects both balance on propensity scores and the final estimate of the treatment effect, every effort must be made to ensure that the estimate of propensity scores has considered all substantively relevant factors and used observed data in a way that is not sensitive to specification errors. The method described next illustrates these points. 5.3.3 Hirano and Imbens's Method for Specifying Predictors Relying on Predetermined Critical t Values Hirano and Imbens's (2001) study was innovative in several aspects: it treated estimated propensity scores as sampling weights and conducted propensity score weighting analysis, combined propensity score weighting and regression adjustment, demonstrated the importance of carefully searching predictors for both logistic regression and outcome regression, and tested the sensitivity of specifications of critical t (i.e., statistical decisions) to the targeted outcome of estimates of treatment effectiveness. In the following description, we focus on their method using predetermined critical t values and put aside the methodology of propensity score weighting. We deal with the weighting procedure in Chapter 7. The problem of variable selection encountered by Hirano and Imbens (2001) warrants a detailed description. A total of 72 covariates were available for use in analysis: Some or all may be used in the logistic regression, and some or all may be used in the outcome regression. Hirano and Imbens viewed the inclusion of variables as a classic subset selection problem in regression (e.g., in the 182 sense of Miller, 1990). With 72 variables, it seemed that any predetermined rule for selecting covariates into either equation (i.e., logistic regression or outcome regression) was subjective and would affect the results of the other model. Therefore, instead of determining which variables to include in the equations using theoretical guidance or rules derived empirically, Hirano and Imbens ran a range of possible models using different values of the t statistic (i.e., a critical t whose value determines whether a covariate should be entered into an equation). The authors then showed estimated treatment effects under all possible combinations of models. Hirano and Imbens used the following steps to estimate the treatment effects: 1. Denoting the critical t value for inclusion of a covariate in the logistic regression as t_{prop} and the critical t value for inclusion of a covariate in the outcome regression as t_{reg} , they considered all pairs with t_{prop} and t_{reg} in the set of $\{0, 1, 2, 4, 8, 16, \infty\}$. 2. They ran 72 simple logistic regression models under a given value of t_{prop} . Each time, they used only 1 of the 72 covariates in the model. Suppose $t_{\text{prop}} = 2$. Under this critical value, if an estimated regression coefficient had observed $t < 2$, then the covariate was ruled out; otherwise, it was retained and was included in the final model of logistic regression under $t_{\text{prop}} = 2$. 3. After running all 72 simple logistic regressions under $t_{\text{prop}} = 2$, Hirano and Imbens found that only a portion of the 72 variables were significant individually. These variables then became the covariates chosen for a logistic regression predicting propensity scores under $t_{\text{prop}} = 2$. 4. Hirano and Imbens then ran outcome regressions in a fashion similar to Steps 2 and 3. That is, they ran 72 simple regressions, with each model containing only one covariate. A combined regression using all covariates that were individually significant under $t_{\text{reg}} = 2$ followed. 5. The authors then replicated Steps 2 to 4 for all other critical t values, that is, for $t_{\text{prop}} = 0, 1, 4, 8, 16$, and ∞ , and for $t_{\text{reg}} = 0, 1, 4, 8, 16$, and ∞ . 6. Finally, Hirano and Imbens calculated treatment effects under conditions defined and produced by all pairs of t_{prop} and t_{reg} . In essence, Hirano and Imbens's approach is a sensitivity analysis—that is, it tests how sensitive the estimated treatment effect is to different specifications of the logistic regression and outcome regression. Under a critical value for both t_{prop} and $t_{\text{reg}} = \infty$ (i.e., a scenario requiring an extremely large observed t for its inclusion), if none of the 72 covariates is used in the logistic regression and in the outcome regression, then the estimated treatment effect is a mean difference 183 between the treated and control patients without any control of covariates. At the other extreme, when a critical value for both t_{prop} and $t_{\text{reg}} = 0$ (i.e., no restriction is imposed on the entering criterion, because any covariate could have an observed t value greater than 0), all 72 covariates are used in both logistic regression and outcome regression. Under this scenario, the estimated treatment effect is a stringent estimation. The number of covariates used in scenarios of other paired t values (i.e., 1, 2, 4, 8, 16) varies, and a high value of critical t typically leads to a model using few covariates. Using this setup, Hirano and Imbens presented a table showing estimated treatment effects under all combinations of t_{prop} and t_{reg} . The table includes a total of 49 cells, which represent seven rows for t_{prop} and seven columns for t_{reg} , and correspond to the seven values in the predetermined set of critical t . The key finding is that there is a great range of variation among estimates of treatment effects for certain scenarios but not for others. Specifically, Hirano and Imbens found that the ranges of estimated treatment effects were the smallest for $t_{\text{prop}} = 2$ and for $t_{\text{reg}} = 2$. The range of treatment effects is $(-.062, -.053)$ under $t_{\text{prop}} = 2$ and $(-.068, -.061)$ under $t_{\text{reg}} = 2$. On the basis of this finding, the authors concluded that the true treatment effect was around $-.06$. This estimated effect was further verified and confirmed by a different estimation using a bias-adjusted matching estimator. 1 The crucial feature of Hirano and Imbens's (2001) approach is flexibility in selecting covariates and specifying models. "By using a flexible specification of the propensity score, the sensitivity to the specification of the regression function is dramatically reduced" (p. 271). Hirano and Imbens (2001) concluded their work by stating, Estimation of causal effects under the unconfoundedness assumption can be challenging where the number of covariates is large and their functional relationship to the treatment and outcome are not known precisely. By flexibly estimating both the propensity score and the conditional mean of the outcome given the treatment and the covariates one can potentially guard against misspecification in a relatively general way. Here we propose a simple rule for deciding on the specification of the propensity score and the regression function. This rule only requires the specification of two readily interpretable cutoff values for variable selection, and is therefore relatively easy to implement and interpret. However, more work needs to be done to understand its properties, and also to investigate alternative approaches to variable selection in similar problems. (pp. 273–274) 5.3.4 Generalized Boosted Modeling 184 Seeking a best logistic regression model may take a completely different route. One of the problems we have seen with logistic regression is specifying an unknown functional form for each predictor. If specifying functional forms can be avoided, then the search for a best model involves fewer subjective decisions and, therefore, may lead to a more accurate prediction of treatment probability. Generalized boosted modeling (GBM), also known as generalized boosted regression, is a method that offers numerous advantages and appears to be promising in solving the variable specification problem. McCaffrey et al. (2004) first applied the GBM approach to the estimation of propensity scores and developed a special software program for the R statistical environment. GBM is a general, automated, data-adaptive algorithm that fits several models by way of a regression tree and then merges the predictions produced by each model. As such, GBM can be used with a large number of pretreatment covariates to fit a nonlinear surface and predict treatment assignment. GBM is one of the latest prediction methods that have been made popular in the machine-learning community as well as mainstream statistics research (Ridgeway, 1999). From a statistical perspective, the breakthrough in applying boosting to logistic regression and exponential family models was made by Friedman, Hastie, and Tibshirani (2000). They showed that an exponential loss function used in a machine-learning algorithm, AdaBoost, was closely related to the Bernoulli likelihood. From this, Friedman et al. developed a new boosting algorithm that finds a classifier to directly maximize a Bernoulli likelihood. Prediction models that use a modern regression tree approach are known as GBMs. Details of the GBM approach may be found in Ridgeway (1999), Friedman (2002), and Mease, Wyner, and Buja (2007). Here, we follow McCaffrey et al. (2004) to highlight application issues in procedures using GBM to estimate propensity scores. First, GBM does not provide estimated regression coefficients such as we normally have with a maximum likelihood estimator. The key feature and advantage of the regression tree method is that the analyst does not need to specify functional forms of the predictor variables. As McCaffrey et al. (2004) pointed out, trees handle continuous, nominal, ordinal, and missing independent variables, and they capture nonlinear and interaction effects. A useful property of trees is that they are invariant to one-to-one transformations of the independent variables. Thus, "whether we use age, log(age), or age² as a participant's attribute, we get exactly the same propensity score adjustments" (McCaffrey et al., 2004, p. 408). This property explains why uncertainty about a correct functional form for each predictor variable is no longer an issue when GBM is used. Because of this, GBM does not produce estimated regression coefficients. Instead, it provides influence, which is the percentage of log likelihood explained by each input variable. The percentages of influence for all predictor variables sum to 100%. For instance, suppose there are three 185 predictor variables: age (x_1), gender (x_2), and a pretreatment risk factor (x_3). A GBM output may show that the influence of x_1 is 20%, of x_2 is 30%, and of x_3 is 50%. On the basis of this output, the analyst can conclude that the pretreatment risk factor makes the largest contribution to the estimated log-likelihood function, followed by gender and age. Second, to further reduce prediction error, the GBM algorithm follows Friedman's (2002) suggestion to use a random subsample in the estimation. In some software programs, this subsample is labeled as training data. Friedman (2002) suggested subsampling 50% of the observations at each iteration. However, programs use different specifications for the subsample size. For instance, the default of training data used by Stata boost is 80%. Third, McCaffrey et al. (2004, pp. 408–409) provided a detailed description of how GBM handles interaction terms. On the basis of their experiments, they recommended a maximum of four splits for each simple tree used in the model, which allows all four-way interactions between all covariates to be considered for optimizing the likelihood function at each iteration. To reduce variability, GBM also requires using a shrinkage coefficient. McCaffrey et al. suggested using a value of .0005, relatively small shrinkage, to ensure a smooth fit. Finally, it is noteworthy that McCaffrey et al. (2004) suggested stopping the algorithm at the number of iterations that minimized the ASAM in the covariates. 2 This is a recommendation particularly directed toward users of GBM who wish to develop propensity scores. As previously mentioned, the GBM procedure stops the algorithm at the number of iterations that minimize prediction errors. Thus, an optimal estimation of propensity scores that minimizes prediction error may not best balance the sample treated and control groups on the observed covariates. On the basis of their experience, McCaffrey et al. observed that the ASAM decreases initially with each additional iteration and reaches a minimum, and afterward the ASAM increases with additional iterations. For this reason, McCaffrey et al. suggest stopping when ASAM is minimized. 5.4 MATCHING After propensity scores are estimated, the next step of analysis often entails matching treated to control participants based on the estimated propensity scores (i.e., to proceed to the Step 2a tasks shown in Figure 5.1). Alternatively, it is possible to skip matching and move to analyzing outcome data using propensity scores as sampling weights (i.e., to proceed to the Step 2b tasks shown in Figure 5.1), conduct propensity score subclassification in conjunction with an outcome analysis (i.e., to proceed to the Step 2c), or calculate a weighted mean difference in the outcome variable using nonparametric regression (i.e., to proceed to the Step 2d tasks shown in Figure 5.1). This 186 section discusses the various methods of matching pertaining to Step 2a. The section is divided into three topics. The first describes conventional greedy matching and its related methods of matching with or without Mahalanobis metric distance. The second describes optimal matching, a method that overcomes some of the shortcomings of greedy matching. The third topic highlights key features of the fine balance procedure. 5.4.1 Greedy Matching The core idea of matching, after obtaining estimated propensity scores, is to create a new sample of cases that share approximately similar likelihoods of being assigned to the treatment condition. Perhaps the most common matching algorithm is the so-called greedy matching. It includes Mahalanobis metric matching, Mahalanobis metric matching with propensity scores, nearest neighbor matching, caliper matching, nearest neighbor matching within a caliper, and nearest available Mahalanobis metric matching within a caliper defined by the propensity score. All methods are called greedy matching. Following D'Agostino (1998) and Smith and Todd (2005), we summarize the major features of greedy algorithms next. 1. Mahalanobis Metric Matching The Mahalanobis metric matching method was invented prior to propensity score matching (Cochran & Rubin, 1973; Rubin, 1976, 1979, 1980a). To apply this method, we first randomly order study participants and then calculate the distances between the first treated participant and all controls, where the distance, $d(i, j)$, between a treated participant i and a nontreated participant j is defined by the Mahalanobis distance: where u and v are values of the matching variables for treated participant i and nontreated participant j , and C is the sample covariance matrix of the matching variables from the full set of nontreated participants. The nontreated participant, j , with the minimum distance $d(i, j)$ is chosen as the match for treated participant i , and both participants are removed from the pool. This process is repeated until matches are found for all treated participants. Because Mahalanobis metric matching is not based on a one-dimensional score, it may be difficult to find close matches when many covariates are included in the model. As the number of covariates increases, the average Mahalanobis distance between observations increases as well. This relationship is a drawback that may be overcome by using two methods that combine the Mahalanobis metric matching with propensity scores (see below). Parenthetically, it is worth noting that C is defined somewhat differently by different researchers, although these different 187 definitions all refer to the method as Mahalanobis metric matching. For instance, D'Agostino (1998) defines C as the sample covariance matrix of the matching variables from the set of control participants, while Abadie et al. (2004) define C as the sample covariance matrix of matching variables from both sets of the treated and control participants. 2. Mahalanobis Metric Matching Including the Propensity Score This procedure is performed exactly as described above for Mahalanobis metric matching, with an additional covariate, the estimated propensity score. The other covariates are included in the calculation of the Mahalanobis distance. 3. Nearest Neighbor Matching P_i and P_j are the propensity scores for treated and control participants, respectively; I_1 is the set of treated participants, and I_0 is the set of control participants. A neighborhood $C(P_i)$ contains a control participant j (i.e., $j \in I_0$) as a match for treated participant i (i.e., $i \in I_1$), if the absolute difference of propensity scores is the smallest among all possible pairs of propensity scores between i and j , as Once a j is found to match to i , j is removed from I_0 without replacement. If for each i there is only a single j found to fall into $C(P_i)$, then the matching is nearest neighbor pair matching or 1-to-1 matching. If for each i the analyst defines n participants who fall into $C(P_i)$ as matches, then the matching is 1-to- n matching. 4. Caliper Matching In the previous type of matching, there is no restriction imposed on the distance between P_i and P_j , as long as j is a nearest neighbor of i in terms of the estimated propensity score. By this definition, even if $|P_i - P_j|$ is large (i.e., j is very different from i on the estimated propensity score), j is still considered a match to i . To overcome shortcomings of erroneously choosing j , we select j as a match for i , only if the absolute distance of propensity scores between the two participants meets the following condition: where ϵ is a prespecified tolerance for matching or a caliper. Rosenbaum and Rubin (1985) suggested using a caliper size of a quarter of a standard deviation 188 of the sample estimated propensity scores (i.e., $\epsilon \leq .25\sigma_P$, where σ_P denotes standard deviation of the estimated propensity scores of the sample). 5. Nearest Neighbor Matching Within a Caliper This method is a combination of the two approaches described above. We begin with randomly ordering the treated and nontreated participants. We then select the first treated participant i and find j as a match for i , if the absolute difference of propensity scores between i and j falls into a predetermined caliper ϵ and is the smallest among all pairs of absolute differences of propensity scores between i and other j s within the caliper. Both i and j are then removed from consideration for matching, and the next treated participant is selected. The size of the caliper is determined by the investigator but typically is set as $\epsilon \leq .25\sigma_P$. Nearest neighbor matching within a caliper has become popular because multivariate analysis using the matched sample can be undertaken, if the sample is sufficiently large. 6. Nearest Available Mahalanobis Metric Matching Within Calipers Defined by the Propensity Score This method combines Mahalanobis distance and nearest neighbor matching into a single approach. The treated participants are randomly ordered, and the first treated participant is selected. All nontreated participants are then selected, within a predetermined caliper of the propensity score and Mahalanobis distances, based on a smaller number of covariates (i.e.,), are calculated between these participants and the covariates without treated participant. One j is chosen as a match for i among all candidates, if the chosen j has the shortest Mahalanobis distance from i . The closest nontreated participant and treated participant are then removed from the pool and the process repeated. According to Rosenbaum and Rubin (1985), this method produces the best balance between the covariates in the treated and nontreated groups, as well as the best balance of the covariates' squares and crossproducts between the two groups. All greedy matching algorithms described above share a common characteristic: Each divides a large decision problem (i.e., matching) into a series of smaller, simpler decisions, each of which is handled optimally. Each makes those decisions one at a time without reconsidering early decisions as later ones are made (Rosenbaum, 2002b). As such, users of greedy matching typically encounter a dilemma between incomplete matching and inaccurate matching. Taking nearest neighbor matching within a caliper as an example, we often have to make a decision between choices such as the following: While trying to maximize exact matches, cases may be excluded due to incomplete matching, or while trying to maximize cases, more inexact matching typically results (Parsons, 2001). Neither of the 189 preceding decisions is optimal. Within the framework of greedy matching, we often wind up recommending running different caliper sizes, checking the sensitivity of results to different calipers, and choosing a method that seems to be best afterward. Greedy matching is criticized also because it requires a sizable common support region to work. When we define the logit in Figure 5.2 as a propensity and set a routine common support region, we see score that greedy matching excludes participants because treated cases fall outside the lower end of the common support region (i.e., those who have low logit) and nontreated cases fall outside the upper end of the common support region (those who have high logit). These participants simply have no matches. The common support region is sensitive to different specifications of the Step 1 model used to predict propensity scores, because logistic regressions with different predictor variables and/or functional forms produce different common support regions. To solve the problem within the conventional framework of propensity score matching, the recommended procedure is for the analyst to test different models and conduct sensitivity analyses by varying the size of the common support region. The requirement of a common support region is also referred to as an overlap assumption in the literature: violations of this assumption led to development of trimming strategies. We review these strategies in Chapters 6 and 9. Addressing limitations embedded in the greedy matching led to the development of optimal matching, which has proven to have numerous advantages over greedy matching. However, before dismissing greedy matching, we want to emphasize that despite its limitations (e.g., it requires a large sample size and loses study participants because of a narrowed common support region under some settings), greedy matching, particularly its nearest neighbor matching within a caliper, has unique advantages. One of these advantages is its permission of subsequent multivariate analysis of almost any kind that allows researchers to evaluate causal effects as they do with randomized experiments. Because of this unique flexibility, greedy matching is widely applied by researchers from a range of disciplines. 5.4.2 Optimal Matching Although the application of optimal matching to propensity score analysis has a history of only about 15 years, the application has grown rapidly and fruitfully primarily for two reasons: the use of network flow theory to optimize matching and the availability of fast computing software packages that makes the implementation feasible. From Hansen (2007), matched adjustment requires analysts to articulate a distinction between desirable and undesirable potential matches and then to match treated and control participants in such a way as to favor the more desirable pairings. As such, the second task (i.e., matching itself) 190 is less statistical in nature, but completing the matching task well can substantially improve the power and robustness of matched inference (Hansen, 2004; Hansen & Klopfer, 2006). Rosenbaum (2002b, pp. 302–322) offers a comprehensive review of the theory and application principles of optimal matching. Hansen developed an optmatch that performs optimal matching in R and is available free with R. Hansen's optmatch package is by far the fastest matching package. Ming and Rosenbaum (2001) proposed using SAS Proc Assign, which is a reasonable alternative for individuals who prefer programming in SAS. Haviland et al. (2007) provided an excellent example of applying optimal matching to analysis of group-based trajectories, and their accessible work outlined important concerns and strategies for conducting optimal matching. Furthermore, Haviland et al. present the material about optimal matching in an accessible way. The central ideas of optimal matching are described in the following. Figure 5.2 Illustration of Common Support Region Using Hypothetical Data As we said earlier, all greedy matching algorithms share a common characteristic: Each method divides a large decision-making problem into a series of smaller, simpler decisions, each of which is handled optimally. Decisions are made one at a time without reconsidering early decisions as later ones are made. In this sense, greedy matching is not optimal. Taking a numerical example, let's consider creating two matched pairs from four participants with the following propensity scores: .1, .5, .6, and .9. A greedy matching would first pick up the second and third participants to form the first pair, because their propensity score distance is smallest and the two participants look most similar (i.e., $|.5 - .6| = .1$) among the four; next, the greedy matching would use the first and last participants to form the second matched pair. By doing so, the total distance on propensity scores from the two 191 pairs are $|.5 - .6| + |.1 - .9| = .9$. An optimal matching, described in this section, would form the following two pairs: Use the first and second participants to form the first pair, and use the third and fourth participants to form the second pair. By doing so, none of the two pairs created by the optimal matching is better than the first pair created by the greedy matching, because the distance for each pair is larger than .1. However, the total distance from the optimal matching is $|.1 - .5| + |.6 - .9| = .7$, which is better than the total distance of the greedy matching (i.e., .9). It is from this example that we see the importance of conducting optimal matching. To facilitate discussion, we first introduce the notation used by Rosenbaum (2002b). Initially, we have two sets of participants: The treated participants are in a set A and the controls are in a set B , with $A \cap B = \emptyset$. The initial number of treated participants is $|A|$ and the number of controls is $|B|$, where $| \cdot |$ denotes the number of elements of a set. For each $a \in A$ and each $b \in B$, there is a distance, $\delta(a, b)$, with $0 \leq \delta(a, b) \leq \infty$. The distance measures the difference between a and b in terms of their observed covariates, such as their difference on propensity scores or Mahalanobis metrics. Matching is a process to develop S strata $\{A_1, \dots, A_S; B_1, \dots, B_S\}$ consisting of S nonempty, disjoint participants of A and S nonempty, disjoint for B and subsets of B , so that By this definition, a matching process produces S matched sets, each of which contains $|A_1|$ and $|B_1|$, $|A_2|$ and $|B_2|$, \dots , and $|A_S|$ and $|B_S|$. Note that, by definition, within a stratum or matched set, treated participants are similar to controls in terms of propensity scores. Depending on the structure (i.e., the ratio of the number of treated participants to control participants within each stratum) the analyst imposes on matching, we can classify matching into the following three types: 1. Pair matching: Each treated participant matches to a single control or a stratification of $\{A_1, \dots, A_S; B_1, \dots, B_S\}$ in which $|A_s| = |B_s| = 1$ for each s . 2. Matching using a variable ratio or variable matching: Each treated participant matches to, for instance, at least one and at most four controls. Formally, this is a stratification whose ratio of $|A_s|/|B_s|$ varies. 3. Full matching: Each treated participant matches to one or more controls, and similarly each control participant matches to one or more treated participants. Formally, this is a stratification of $\{A_1, \dots, A_S; B_1, \dots, B_S\}$ in which the minimum of $(|A_s|, |B_s|) = 1$ for each s . Optimal matching is the process of developing matched sets $\{A_1, \dots, A_S; B_1, \dots, B_S\}$ with a size of (α, β) in such a way that the total sample size of 192 propensity scores is minimized. Formally, optimal matching minimizes the total distance Δ defined as where $\omega(|A_s|, |B_s|)$ is a weight function. Rosenbaum (2002b) defined three choices among weight functions: (1) the proportion of α treated participants who fall in set s or $\omega(|A_s|, |B_s|) = |A_s| / \alpha$, (2) the proportion of the b control participants who fall in set s or $\omega(|A_s|, |B_s|) = |B_s| / \beta$, and (3) the proportion of the sum of treated and control participants who fall in set s or $\omega(|A_s|, |B_s|) = (|A_s| + |B_s|) / (\alpha + \beta)$. For each of the three weight functions, the sum of weights equals 1, and the total distance Δ is truly a weighted average of the distance $\delta(A_s, B_s)$. The actual choice of weight function in an application is not so important. Of greater importance is that optimal matching develops matched sets (i.e., the challenge is to create S sets and identify which controls are matched to which treated participants) in such a way that the matching optimizes or minimizes the total distance for a given data set and prespecified structure. How does optimal matching accomplish this goal? Suffice it to say that this method achieves the goal by using a network flow approach (i.e., a topic in operations research) to matching. Rosenbaum (2002b) provides a detailed description of the optimal matching method. A primary feature of network flow is that it concerns the cost of using b for a as a match, where a cost is defined as the effect of having the pair of (a, b) on the total distance defined by Equation 5.6. Standing in sharp contrast to greedy matching, optimal matching identifies matched sets in such a way that the process aims to optimize the total distance, and decisions made later take into consideration decisions made earlier. Indeed, later decisions may alter earlier ones. From an application perspective, the structure imposed on optimal matching (i.e., whether you want to run a 1-to-1 pair matching, a matching with a constant ratio of treated to control participants, a variable matching with specifications of the minimum and maximum number of controls for each treated participant, or a full matching) affects both the level of bias reduction and efficiency. In this context, efficiency is defined as the reciprocal of the variance, and therefore, a high level of efficiency is associated with a low variance. Haviland et al. (2007) review this topic in detail and distill from it two practice implications: (1) There are substantial gains in bias reduction from discarding some controls, yet there is little loss in efficiency from doing so, provided multiple controls are matched to each treated participant, and (2) there are substantial gains in bias reduction from permitting the number of matched controls to vary from one treated participant to another, yet there is little loss in efficiency from doing so if imbalance is not extreme. From this, they draw three general principles for making decisions about matching structure: 193 1. Having two controls for each treated participant is more efficient than matched pairs (i.e., a 1-to-1 match). 2. Having a large number of controls yields negligible gains in efficiency. 3. Having some variation in the number of matched controls among all strata S does not greatly harm efficiency. In practice, the selection of a matching structure should be based on the number of treated participants and controls. Sometimes the decision is implied by the structure of the data. For example, assume you want to evaluate data generated by a quasi-experimental design or a randomized experiment in which randomization has failed and the number of treated participants is close to the number of controls. Given this scenario, a 1-to-1 pair matching is probably the only choice. On the other hand, suppose you have conducted a case-cohort design (e.g., a design described by Joffe & Rosenbaum, 1999) and the ratio of control to treated participants is large, perhaps on the order of 3:1 or 4:1. Under such conditions, there is a range of possible choices to specify the structure of optimal matching, and the choice of structure will have an impact on bias reduction and efficiency. A common practice under such conditions is to test different structures and then compare results (i.e., estimates of treatment effects) among matching schemes. Methods in this field of matching are rapidly evolving, and we recommend also consulting the literature. Hansen (2004), for example, found that in the context of a specific application, variable matching with a specific structure worked best; that is, each treated participant was matched to at least $.5(1 - \gamma)$ controls and at most $2(1 - \gamma)$ controls, where represents the proportion of treated participants in the sample. We cannot be sure that a variable matching approach with such a structure will continue to be a best choice in other data sets; however, it should clearly be explored as a viable option. Finally, it is important to note that pair matching generated by an optimalmatching algorithm is different from the pair matching generated by greedy matching, particularly so for a 1-to-1 or 1-to- n nearest neighbor matching within a caliper. The primary reason for these differences is that greedy matching is not an optimal process, and decisions made earlier affect the level of optimizations accomplished later. On the other hand, even though the researcher might use an optimal-matching algorithm such as optmatch to conduct a pair matching, Rosenbaum (2002b, pp. 310–311) showed that such pair matching is not generally optimal, especially when compared with full matching based on the same data. 5.4.3 Fine Balance The matching procedures implemented by the greedy and optimal algorithms share a key feature: Treated participants are matched with control participants 194 on a single propensity score to balance a large number of covariates. An innovative method called fine balance does not require individually matching on the propensity score (Rosenbaum, Ross, & Silber, 2007). Because this method is a bit different, we briefly highlight its main ideas here. Fine balance refers to exactly balancing a nominal variable, often a variable with many discrete categories, without trying to match individuals on this variable. Fine balance employs a principle similar to the network optimization algorithm (see Section 5.4.2) to create a patterned distance matrix, which is passed to a subroutine that optimally pairs the rows and columns of the matrix. In an illustrative example, Rosenbaum et al. (2007) addressed the problem of matching on a nominal variable with 72 categories (i.e., study participants had 9 possible years of diagnosis and 8 geographic locations, which were considered substantively important for achieving an exact balance). Using the fine balance method, an exact balance on the 72 categories and close individual matches on a total of 61 covariates based on a propensity score are obtained. The key idea of fine balancing is that the nominal variable is balanced exactly at every level. Rosenbaum et al. used SAS Proc Assign to implement the fine balance strategy. As a matching tool, fine balance is used in conjunction with other matching tools, such as propensity scores, minimum distance matching, or the Mahalanobis metric distance. It exemplifies the growth and innovation in the field. In practice, the choice is not which one tool to use, but rather it has become whether or not to apply a particular tool in conjunction with other tools. For a discussion of when fine balance is implied, see Rosenbaum et al. (2007). 5.5 POSTMATCHING ANALYSIS This section describes postmatching procedures (i.e., Step 3a shown in Figure 5.1) for a three-step analysis of propensity scores. Step 3b shown in Figure 5.1 (i.e., subclassification following a greedy matching) involves a procedure similar to subclassification without conducting greedy matching. We defer the discussion about postmatching subclassification to Chapter 6. Because methods applied at Step 2 for matching vary, methods of postmatching analysis also vary. We describe five methods: (1) multivariate analysis after greedy matching, (2) checking covariate imbalance before and after optimal matching, (3) outcome analysis for an optimally matched sample using the Hodges-Lehmann aligned rank test, (4) regressing the difference scores of an outcome on difference scores of covariates based on an optimal pair-matched sample, and (5) regression using the Hodges-Lehmann aligned ranks of both outcomes and covariates based on an optimally matched sample. 5.5.1 Multivariate Analysis After Greedy Matching The property described in 4(c) under Section 5.2 (i.e., Corollary 4.3 of 195 Rosenbaum & Rubin, 1983) is the theoretical justification for conducting multivariate analysis after greedy matching. The impetus for developing propensity scores and matching is that observational data are usually not balanced, and hence, we cannot assume that treatment assignment is ignorable. After matching on the estimated propensity scores, at least the sample is balanced on observed covariates (between treated and control participants), and therefore, we can perform multivariate analyses and undertake covariate adjustments for the matched sample as is done in randomized experiments. In theory, regression, or any regression-type models, may be used at this stage to estimate ATEs by using a dichotomous explanatory variable indicating treatment conditions. Many studies have used this approach to adjust and estimate ATEs. For instance, following a caliper matching, Morgan (2001) conducted a regression analysis to estimate the impact of Catholic schools on learning. H. L. Smith (1997) conducted a variance-components analysis (also known as a hierarchical linear modeling) based on a matched sample generated by a random order, nearest available pair-matching method. In a sample of Medicare eligible patients, he sought to estimate the effects of an organizational innovation on mortality within hospitals. Guo et al. (2006) conducted a survival analysis (i.e., Kaplan-Meier product limit estimates) after nearest neighbor matching within a caliper. They were interested in estimating the impact of caregivers' use of substance abuse services on the hazard rate of child maltreatment rereport. 5.5.2 Computing Indices of Covariate Imbalance It is often desirable to check covariate balance before and after optimal matching. Haviland et al. (2007) developed the absolute standardized difference in covariate means, d_X for use before matching and d_{Xm} for use after matching. This measure is similar to ASAM in the literature. Before matching, d_X is used to check imbalance on covariate X . It is estimated using the following formula: where MX_t and MX_c are the means of X for treated and potential control groups, respectively. Denoting the standard deviations of the treated and potential control groups as S_{Xt} and S_{Xc} , we compute the overall standard deviation as Note that the statistic d_X is similar to the normalized difference score Δ_X introduced in Chapter 1 (Equation

1.1). The difference lies in the denominator: dX uses while ΔX uses The value dX_m for the level of imbalance on covariate X after matching is estimated by 196 In this equation, subscript c denotes the control group, and MX_c denotes the unweighted mean of means of the covariate X for the controls matched to treated participants. This covariate X can be computed by the following method: After matching, each treated participant i in stratum s is matched to ms_i controls, $j = 1, \dots, ms_i$. The number of treated participants in stratum s is ns , and the total number of treated participants in the whole sample is $n+$. The values of a covariate X have a subscript t or c for treated or control group, a subscript s for the stratum, a subscript i to identify the treated participant, and a subscript j for controls to distinguish the ms_i controls matched to treated participant i . Thus, X_{csij} denotes the value of X for the j th control who matches to treated participant i , $j = 1, \dots, ms_i$. Denoting M_{cs_i} the mean of the ms_i values of the covariate X for the controls matched to treated participant i , and MX_c the unweighted mean of these means, we have dX and dX_m can be interpreted as the difference between treated and control groups on X in terms of the standard deviation unit of X . Note that dX and dX_m are standardized measures that can be compared with each other. Typically, one expects to have $dX > dX_m$, because the need to correct for imbalance before matching is greater, and the sample balance should improve after matching. Taking the data reported by Haviland et al. (2007, Table 4, p. 256) as an example, before optimal matching, the dX of the covariate "peer-rated popularity" is 0.47, meaning that the treated and control groups are almost half a standard deviation apart on peer-rated popularity, whereas, after optimal matching, the dX_m of the same covariate is 0.18, meaning that the difference between the two groups is 18% of a standard deviation for peer-rated popularity; indeed, matching improves balance. An illustrative example of computing dX and dX_m is presented in Section 5.8.2. Graphic approaches to checking covariate balances also have been developed (e.g., Haviland et al., 2007). The R program `twang` (Ridgeway et al., 2013) and `MatchIt` (Ho et al., 2004) produce useful figures for checking imbalances before and after matching.

5.5.3 Outcome Analysis Using the Hodges-Lehmann Aligned Rank Test After Optimal Matching After optimal matching, we usually want to estimate the ATE and perform a 197 significance test. In this section, we describe these procedures for a matched sample created by full matching or variable matching. Methods for a sample created by optimal pair matching are described in the next section. The sample ATE may be assessed by a weighted average of the mean differences between treated and control participants of all matched sets, as where i indexes the b matched strata, N the total number of sample participants, n_i the number of treated participants in the i th stratum, m_i the number of controls in the i th stratum, and the mean responses corresponding to the control and treated groups in the i th stratum. The significance test of the ATE may be performed by the Hodges-Lehmann aligned rank test (Hodges & Lehmann, 1962). Lehmann (2006, pp. 132–141) described this test in detail. Its major steps include the following: 1. Compute the mean of the outcome for each matched stratum i and then create a centering score for each participant by subtracting the stratum's mean from the observed value of the outcome. 2. Sort the whole sample by the centering scores in an ascending order and then rank the scores; the ranked score is called aligned rank and is denoted as k_{ij} ($j = 1, \dots, N_i$), where i indicates the i th stratum, j the j th observation within the i th stratum, and N_i the total number of participants in the i th stratum. 3. For each stratum i , compute 4. Across strata, calculate where is the sum of the aligned ranks for the treated participants within the i th stratum. Note that the subscript s in all the preceding equations indicates treatment participants. 5. Finally, calculate the following test statistic Z^* : The Z^* statistic follows a standard normal distribution. Using Z^* , the analyst 198 can perform a significance test of a nondirectional hypothesis (i.e., perform a two-tailed test) or a directional hypothesis (i.e., perform a one-tailed test). Social and behavioral sciences researchers and, indeed, policy makers, are often interested in effect sizes. An exact measure of the size of the treatment effect under the current setting is yet to be developed. However, we recommend using dX_m defined by Equation 5.8 as an approximation of effect size for postmatching analysis. The dX_m statistic applied to an outcome variable, according to Haviland et al. (2007), is similar to Cohen's d . To approximate an effect size, simply calculate dX_m for the outcome variable after matching. An example of outcome analysis after optimal full or variable matching is presented in Section 5.8.3. 5.5.4 Regression Adjustment Based on Sample Created by Optimal Pair Matching After obtaining a matched sample using optimal pair matching, an ATE is estimated using a special type of regression adjustment developed by Rubin (1979). The basic concept of regression adjustment is straightforward: regressing the difference scores between treated and control participants on the outcome variable on the difference scores between treated and control participants on the covariates. Following Rubin (1979), we write $\alpha_i + W_i(X)$ to denote the expected value of the outcome variable Y given a covariate matrix X in the population P_i , where i denotes treatment condition ($i = 1$, treated; $i = 0$, control). The difference in expected values of Y for P_1 and P_0 units with the same value of X is $\alpha_1 - \alpha_0 + W_1(X) - W_0(X)$. When P_1 and P_0 represent two treatment populations such that the variables in X are the only ones that affect Y and have different distributions in P_1 and P_0 , then this difference is the effect of the treatment at X . If $W_1(X) = W_0(X) = W(X)$ for all X , the response surfaces are then parallel, and $\alpha_1 - \alpha_0$ is the treatment effect for all values of the covariates X . The regression adjustment is undertaken as follows: 1. Take the difference scores on the outcome variable Y between treated and control participants $Y = Y_1 - Y_0$. 2. Take the difference scores on the covariate matrix X between treated and control participants $X = X_1 - X_0$. 3. Regressing Y on X , we obtain the following estimated regression function: then is the estimated ATE. We can use the observed t statistic and p value 199 associated with to perform a significance test (two-tailed or one-tailed). An example of outcome analysis after optimal pair matching is presented in Section 5.8.4. 5.5.5 Regression Adjustment Using Hodges-Lehmann Aligned Rank Scores After Optimal Matching The method described in Section 5.5.3 is bivariate—that is, it analyzes ATE on the outcome between treated and control participants where control of covariates is obtained through optimal matching. Analysts could also evaluate ATE in conjunction with covariance control (i.e., regression adjustment) to conduct a multivariate analysis, which is the procedure suggested by Rosenbaum (2002a, 2002b). In essence, this method is a combination of the Hodges-Lehmann aligned rank method and regression adjustment. The central idea of this approach is summarized as follows: create aligned rank scores of the outcome variable, create aligned rank scores for each of the covariates, and then regress (perhaps using robust regression) the aligned rank scores of the outcome on the aligned rank scores of covariates; the residuals are then ranked from 1 to N (i.e., the total number of participants in the sample) with average ranks for ties. By following this procedure, the sum of the ranks of the residuals for treated participants becomes the test statistic (i.e., Equation 5.10). At this stage, you may use the Hodges-Lehmann aligned rank test to discern whether the treatment effect is statistically significant after the combination of optimal matching and regression adjustment. 5.6 PROPENSITY SCORE MATCHING WITH MULTILEVEL DATA Social and health sciences researchers are increasingly encountering grouped or multilevel data in which, for instance, students are nested within classrooms and classrooms are nested within schools. Analyzing such data requires special treatment because most multivariate models assume independent observations of the study sample, and grouped data clearly violate this assumption. For the past two decades, progress has been made toward developing efficient and robust statistical methods to correct for violations of the independence assumption. This section describes the application of propensity score analysis to multilevel data. We focus on propensity score matching, although the same approaches can be applied to propensity score weighting and subclassification with appropriate adjustments. In the context of observational studies, the need for conducting a multilevel analysis in conjunction with propensity score modeling can be seen in the failure of a multisite randomized trial, where randomization has been compromised by attrition, treatment noncompliance, or both (Barnard, Frangakis, Hill, & Rubin, 2003), or by a small number of clusters used in 200 randomization such that the randomization mechanism cannot fully balance data. Selection bias due to either individual- or cluster-level covariates often exists in such data, and as a consequence, a multilevel model such as a linear mixed model suffers from violations of assumptions. For instance, the linear mixed model assumes zero means of intercept- and slope-random effects, as well as zero correlations of random effects with covariates at the individual and cluster levels. When selection bias exists, these assumptions are prone to violation, and multilevel models may be affected by the same endogeneity problems as those encountered in the analysis of single-level data using OLS regression. In this section, we review statistical approaches developed to analyze multilevel data, primarily the linear mixed model and the Huber-White estimator of robust standard errors. We then discuss the problem of selection bias in the context of failures of multisite cluster trials and explain the need for conducting a multilevel analysis in conjunction with the use of propensity scores. Next, we move to the description of approaches for estimating propensity scores based on multilevel observed variables. Finally, we present general procedures using estimated propensity scores, that is, postmatching outcome analysis in which researchers may consider correction of additional sources of clustering produced by matching. 5.6.1 Overview of Statistical Approaches to Multilevel Data The need for multilevel modeling stems from research questions as well as the special structure of data sets being analyzed. In the social and health sciences, researchers often need to test how micro-level characteristics interact with macro-level characteristics and, therefore, to test the joint effects of two-level characteristics on an outcome variable. When analyzing such data, researchers inevitably encounter a statistical problem known as clustering. When study subjects at a lower level are nested within a unit at the higher level, such as students who are nested within classrooms, the sample data often show clustering of the outcome variable—that is, students within the same classroom share similarities on the outcome variable. Alternatively, we can think of this as a nonzero autocorrelation on values of the outcome variable among students in the same classrooms. In essence, when autocorrelation or clustering is present, information coming from the same unit (such as n students from the same classroom) tends to be more alike than information from independent units (such as a data set of n unrelated students). The presence of autocorrelation is not by chance, because these students share the same physical classroom and the same teacher. Their common experiences are likely to produce correlated scores. Technically, when clustering or autocorrelation is present, some information in the nested data set is redundant (Allison, 1995). A regression model failing to adjust for clustering tends to produce a biased standard error for the coefficient of interest. Typically, the error is smaller than it should be, raising the risk of 201 spuriously observing a significant finding. The level of autocorrelation or clustering is often assessed by an intraclass correlation coefficient (ICC) and is computed by an unconditional ANOVA with random effects (Raudenbush & Bryk, 2002). Through such modeling, one obtains the between-group and within-group variances of the outcome variable. The ICC can be computed by dividing the between-group variance by the sum of the between-group and within-group variances. The coefficient measures the proportion of variance in the outcome variable that is between groups: A high ICC indicates that much variation in the outcome variable is due to groups (i.e., more information in the individual-level data is redundant) and signifies the need to use a multilevel model. There is no clear cutoff value of ICC found from the literature that warrants a control of clustering effects, but in general, whenever the ICC is greater than zero, estimating the design effect of clustering becomes important in determining whether a correction is warranted At the present time, efforts toward correction for clustering effects are made in two distinct categories of statistical approaches: the linear mixed model and Huber-White estimator of robust standard errors. The core features of these models are summarized below. The linear mixed model, also known as a random effects model and a hierarchical linear model (HLM), was originally developed by Laird and Ware (1982). In the social and behavioral sciences, Raudenbush and Bryk (2002), Goldstein (2010), and Snijders and Bosker (1999) are widely cited. The linear mixed model in the form of Laird and Ware (1982) may be expressed as follows: where β is a fixed effects vector, X is a matrix of independent variables, Z is a design matrix of random effects, u is a random effects vector, and e is a random vector. The model assumes that the data vector y is normal and independently distributed with mean vector $\mu = X\beta + Z\gamma$ and variance-covariance matrix $V = ZGZ' + R$, $E(u) = 0$, $E(e) = 0$, $V(u) = G$, and $V(e) = R$. To model V , one needs to define the design matrix Z and specify G and R . Denote the parameters of G and R of the general linear model as q ; the -2 log-likelihood function using the maximum likelihood method (R. C. Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006) is Solving the score equations, the maximum likelihood estimators of the fixed-effect parameters are $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$. The crucial feature of the model is its inclusion of the component Zu ; that is, with a user-specified design matrix Z , the model contains intercept and/or slope random effects (i.e., the inclusion of 202 random-effect vector u). Without this term, the model reduces to a linear fixed-effect model or regression. Random effects are a set of extra heterogeneity and usually do not have important substantive meaning; by having them in the multilevel modeling, researchers correct for biases triggered by clustering and make the significance tests more accurate. The second type of multilevel model was originally developed by Huber (1967) and later by White (1980) independently. Instead of adding random effects to a fixed-effect model, this approach corrects for biases induced by clustering by directly estimating robust standard errors. The method has been modified since its invention to accommodate more complicated data situations, but in general, the term robust standard error estimator can be used interchangeably with sandwich estimator and marginal approach. Important variants of the Huber-White approach include the generalized estimating equation (GEE) estimator (Zeger, Liang, & Albert, 1988) and two marginal models developed for survival analysis: the VLW model (Wei et al., 1989) and the LWA model (Lee et al., 1992). The principle of the robust estimator of standard errors is illustrated by the following algorithm employed in most statistical computing packages, such as Stata (StataCorp, 2007). Suppose the variance estimator assuming independent observations is where is the estimator of variance provided by a fixed-effect model, L is the likelihood function, b is the parameter vector, and u_j (a row vector) is the When clustering effects are present, contribution from the j th observation to the observations denoted by j are not independent; however, the observations can be divided into M groups G_1, G_2, \dots, G_M that are independent, and then the robust estimator of variance becomes where is the contribution of the k th group to In this context, application of the robust namely, k variance formula involves using a different decomposition of $1, \dots, M$, rather than $u_j, j = 1, \dots, N$. Because the three terms of the robust estimator of variance make the estimator look like a sandwich, the method is called "sandwich estimator." The crucial term in the correction lies in the middle of the estimator (i.e., the "meat" of the "sandwich"), or by having this term, bias induced by clustering is corrected. The GEE estimator allows users to specify different types of middle terms, called specifying working correlation matrices (Hardin & Hilbe, 2003). In the following description of applying propensity score matching to multilevel modeling, we will focus on the first type of correction method, that is, propensity score analysis with the linear mixed modeling. This is a core method found in the social and behavioral sciences literature. Thoemmes and West (2011) provide a comprehensive review of corrective methods using 203 propensity scores for nonrandomized designs with clustered data. Others are also beginning to address the issue. See, for example, Hong and Raudenbush (2006); Hong and Yu (2008); Arpino and Mealli (2011); Gadd, Hanson, and Manson (2008); Grieswold, Localio, and Mulrow (2010); and Kim and Seltzer (2007). Our discussion follows Thoemmes and West (2011). 5.6.2 Perspectives Extending the Propensity Score Analysis to the Multilevel Modeling When selection bias exists in observational data that are multilevel, the usual correction of clustering effects cannot remove the bias due to the same problem of endogeneity found from a single-level analysis. Thoemmes and West (2011) considered two perspectives to extend propensity score models to multilevel data. In the first case, a propensity score analysis attempts to approximate a multisite randomized trial in which units are randomized within individual clusters. A typical example of this type of clustering is the occurrence of selection bias at the school level, that is, schools having different academic resources and different policies about the retention of poorly performing students (Hong & Raudenbush, 2006). In this context, treatment effects within cluster and their generalization across clusters become the focus of the study. Hong and Raudenbush show that in this setting, one still can apply the counterfactual framework or potential outcome model to evaluate treatment effects, but additional assumptions must be made. Specifically, the strongly ignorable treatment assignment assumption for single-level analysis needs to be modified. When data are multilevel, the ignorability assumption implies that the potential outcome should ideally be invariant to cluster membership, cluster composition, and the treatment assignments of other participants and that treatment delivery should be identical across clusters. Hong and Raudenbush show that this assumption might be unrealistic in many applied research contexts. To relax the strongly ignorable treatment assignment assumption, Hong and Raudenbush (2006) propose allowing cluster-specific effects on the potential outcome to be additive: "The observed group composition and agent allocation . . . are viewed as random events that are exchangeable" (p. 1850). Under this relaxed assumption, the cluster-specific effects are regarded as random effects that have an expected value of zero in the population. Small and nonsignificant variance components are then interpreted as evidence that the treatment effect is relatively homogeneous across clusters. This perspective justifies an outcome analysis that is multilevel and specifies random effects associated with clusters. The second perspective, considered by Thoemmes and West (2011), treats clustering as an incidental feature of the design. A typical example is the clustering of randomly selected members of unacquainted individuals waiting to complete forms in a state unemployment office. These individuals are offered 204 the opportunity to participate in a job-seeking skills program delivered in a group setting. In this case, the treatment is assumed to be implemented without variation. The focus of the evaluation is to estimate the average treatment effect for the population of individuals, controlling for the potential nuisance effect of incidental clustering. The propensity score perspective in this case attempts to approximate a single-level randomized experiment on individuals who happen to be clustered. With regard to the ignorability assumption, Thoemmes and West (2011) propose, "In this case, it is assumed that there are no variations in treatment implementation across clusters and that selection is invariant across clusters. The multilevel analysis addresses incidental effects of the clustering of participants into groups. As in the approximation of the multisite randomized trial, the assumptions of no between-cluster interference and strongly ignorable treatment assignment given person-level covariates are needed" (p. 519). 5.6.3 Estimation of the Propensity Scores Under the Context of Multilevel Modeling When applying propensity score modeling to clustered data, both the procedure for estimating propensity scores and the outcome analysis must be modified to take clustering effects into consideration. In this subsection, we describe five approaches that modify the logistic regression estimating propensity scores: (1) the single-level model that pools all participants prior to estimation of parameters and ignores clustering, (2) fixed effects regression models in which single-level logistic regression is used for each individual cluster, (3) a multilevel logistic regression with a narrow inference space, (4) a multilevel logistic regression with a broad inference space, and (5) a single cluster-level logistic regression. The first four models come from Grieswold et al. (2010), Kim and Seltzer (2007), and Thoemmes and West (2011), and the last model is added by the authors and is based on their own research experiences. Single-level model. In this approach, both the vectors of person-level variables X and cluster-level variables W are specified in a one-level logistic regression. The model may contain possible interactions of W and X but does not include random effects associated with clusters. The model is specified as follows: where $\logit(e^{(x, w)})$ is the estimated logit of the propensity score, β_0 is an is a vector of regression coefficients and predictor variables for intercept, the P person-level variables (potentially including person-level interactions and polynomial terms), is a vector of regression coefficients and predictor variables for the Q cluster-level variables (potentially including cluster-level interactions and polynomial terms), and is a vector of regression 205 coefficients and all possible interactions between person- and cluster-level variables. Fixed effects model. This is an alternative approach to the single-level model. The model includes a set of dichotomous variables and interaction terms of cluster dichotomous variables and person-level variables. Conceptually, one may view the estimation of propensity scores in this setup as running a separate logistic regression for each cluster. The model is specified as follows: where C is a set of dichotomous variables, one for each cluster, with the reference group as the ($C + 1$)th cluster, and the remaining regression coefficients and variables are defined as in Equation (5.11). The fixed effects model has a similar setup to the single-level model, "with the difference being that the indicator variable C in the fixed effects model allows for estimation of intercepts and possibly regression slopes of the X variables for every single cluster. As a result, the fixed effects model theoretically allows for estimation of unbiased within cluster regression slopes, regardless of presence or absence of any cluster-level covariates" (Thoemmes & West, 2011, p. 522). It is worth noting that this method has two limitations: It requires very large sample sizes within clusters, and the estimated propensity scores across different clusters are not comparable (Thoemmes & West, 2011). The fixed effects model cannot be applied to a cluster randomization when all individuals in C – K clusters are in the treatment condition and all individuals in K clusters are in the control condition; in this circumstance, the cluster indicator variables are perfectly linear functions of the treatment conditions, and the dependent variable of the logistic regression is perfectly predictable by the cluster indicators. Multilevel model with a narrow inference space. The logistic regression predicting propensity scores may take into consideration clustering effects and be specified as a hierarchical generalized linear model with random effects. However, predicted propensity scores can be based on fixed effects only or on both fixed and random effects. This distinction is also referred to as the distinction between narrow (subject-specific) and broad (population-average) inference spaces. According to Thoemmes and West (2011), the narrow inference space model captures propensity score models that mimic a randomized multisite trial (i.e., the first perspective discussed in Section 5.6.2), whereas the broad inference space model captures propensity score models that mimic a randomized individual trial with incidental clustering (i.e., the second perspective discussed in Section 5.6.2). The multilevel logistic regression with a narrow inference space includes both fixed and random effects, as 206 where is a vector of regression coefficients and person-level covariates, is a vector of regression coefficients and cluster-level covariates, is a vector of all possible interaction terms between person- and cluster-level covariates, is a random effect component that influences the intercept of each is a vector of random effect components that influence each cluster j , and of the regression slopes of person-level predictors. In running empirical data, there may be fewer than P random slope effects that are included in the model, because some random effects may not be statistically significant and are excluded. Multilevel model with a broad inference space. The model to estimate propensity scores based on the broad inference space is identical to that with a narrow inference space, except that the model does not include random effects, as Single cluster-level model. In social and health sciences research, researchers often apply a group or cluster randomization to avoid spillover effects or other types of contamination to ensure that the stable unit treatment value assumption (SUTVA) holds. The Social and Character Development (SACD) program discussed in Chapter 1 (i.e., Example 6) is such a case. Evaluating programs generated by group randomization is often challenging, because the unit of analysis is a cluster—such as a school—and the sample sizes may be so small as to compromise randomization. Randomization fails in this context primarily because the number of clusters being randomized is extremely small, and as a consequence, it cannot sufficiently balance data. When randomization fails at the cluster level (which can be tested on pretreatment covariates between treatment and control conditions), it is important to predict propensity scores directly using cluster-level variables. On the basis of our research experiences with the SACD and other evaluations, we suggest predicting propensity scores using a single cluster-level logistic regression; at least, it is desirable to test this model and compare results of the model with other models. The single cluster-level model can be expressed as w h e r e e W denotes cluster-level variables, typically collected from a pretreatment point, and the model is simply run based on J clusters. 207 5.6.4 Multilevel Outcome Analysis After obtaining estimated propensity scores using one or several models described in the previous subsection, one may employ these scores to conduct a greedy matching. Because the nesting structure of the data mimics a randomized multisite trial or a randomized individual trial with incidental clustering, the outcome analysis should also be multilevel to correct for the clustering effects. Failure to correct for clustering effects leads to a biased estimation of standard errors associated with regression coefficients and inaccurate tests of statistical significances (Guo, 2005; Raudenbush & Bryk, 2002). A typical linear mixed model of the outcome analysis may look as follows (Thoemmes & West, 2011): where is the estimated average causal treatment effect over all clusters and Z_{ij} is the treatment assignment variable ($1 =$ treatment, $0 =$ control) for participant i in cluster j . The model may include additional covariates at the person level (i.e., X variables) and at the cluster level (i.e., W variables), the interactions of person- and cluster-level variables, and random slopes, depending on research interests and tests of statistical significance of these effects. The correction of clustering effects after matching warrants special consideration. Although it may not happen in all data sets and the occurrence varies by research design and study phenomenon, in some studies, there may exist two sources of clustering after matching: clustering due to multilevel data, in which study participants are nested within clusters, and clustering due to matching, in which participants from the same matched set share similar values on the outcome variable. The second source of clustering is a design effect that is added by matching. By a matching design, study participants are all similar in terms of the estimated propensity scores or the joint distributions of predictor variables used in the logistic regression. This similarity may lead to a high level of covariance on the outcome variable, due to known and unknown reasons. Whether or not this second source of clustering is present is not self-evident and should be regularly tested out. That is, researchers need to use the ID of matched sets as a cluster variable to check ICCs. If the ICC from this source is nonnegligible, the outcome analysis should be modified to control a matching design effect. In the postmatching outcome analysis, if both sources of clustering are present, special procedures are needed. The two sources of clustering after matching make the data set crossclassified rather than nested. In contrast to nesting, a cross-classified data structure is more complex and defined by lower-level units that are crossclassified by two higher-level units. In typical two-way cross-classified data, the so-called within-cell observations are classified into a matrix having J rows (i.e., J clusters/schools embedded in the original data) and K columns (i.e., K 208 matched sets produced by matching). To correct for the two sources of clustering, one may run a cross-classified random effect model (CCREM). CCREM is a rich model that provides numerous modeling possibilities (Raudenbush & Bryk, 2002, chap. 12). In general, with CCREM, researchers first run a Level 1 or "within-cell" model, in which they have a unique set of observations that are nested within each cell of the cross-classification. Next, the researchers run various models (depending on the research purpose and availability of data) to examine fixed effects of a Level 2 row predictor and a Level 2 column predictor, randomly varying effects of a Level 2 row and a Level 2 column predictor, and fixed and random effects of additional row and column predictors. The CCREM that we propose to use in the current context takes a simpler form and does not have all possible parameters. The conditional model we propose to use will work for most matched data sets, and the estimation procedure is available in popular software packages (e.g., Stata, SAS, and HLM). In this model, we include random main effects for original clusters and for matched sets. Denoting an outcome for subject i who is a member of cluster j and matched set k as Y_{ijk} , the treatment assignment variable as Z ($1 =$ treatment, $0 =$ control), P person-level variables as X_p ($p = 1, 2, \dots, P$), Q cluster-level variables as W_q ($q = 1, 2, \dots, Q$), we express a general CCREM as With this setup, the model provides the following parameters: θ_{00} is the model intercept, θ_{0i} is the estimated average treatment effect over all clusters and matched sets, P fixed effects θ_{0p} (one for each X_p at the person level), Q fixed effects θ_{0q} (one for each W_q at the cluster level), and two random effects: b_{00j} is the main random effect associated with original cluster j , and c_{00k} is the main random effect associated with matched set k . By including the two random effects and b_{00j} and c_{00k} , the model adequately controls for clustering effects induced by the multilevel structure of the original data and by matching. Raudenbush (1993) developed an algorithm by combining Lindley and Smith's (1972) two types of exchangeabilities into a single procedure and using an expectation-maximization (EM) algorithm of a full or restricted maximum likelihood estimator to estimate the model. Goldstein (1987) described a trick analysts can use (i.e., creating an artificial Level 3 unit in which both original clusters and matched sets are nested) so that they can use a software procedure designed to estimate a linear mixed model (e.g., `xtmixed` of Stata) to estimate CCREM. It is worth noting that modeling multilevel data with propensity scores is a 209 rapidly growing area. The current discussion does not attempt to provide a comprehensive review of all issues pertaining to the topic. Instead, it focuses on main methodological challenges and useful strategies developed to address these challenges. We hope the discussion of key issues provides readers with a general framework and contributes also to understanding additional nuances that bear consideration when conducting propensity score analysis with multilevel data. 5.7 OVERVIEW OF THE STATA AND R PROGRAMS Currently, no commercial software package offers a comprehensive procedure for implementing all matching models described in this chapter. In SAS, Lori Parsons (2001) developed several macros (e.g., the GREEDY macro does nearest neighbor within-caliper matching). Several user-developed programs available in Stata and R allow users to undertake most tasks described in this chapter. On the basis of our experience with Stata, we found `psmatch2` (Leuven & Sianesi, 2003), `boost` (Schonlau, 2007), `imbalance` (Guo, 2008b), and `hdgesl` (Guo, 2008a) to be especially useful. In the following section, we provide an overview of the main features of these programs. In addition, the overview includes the R program `optmatch` (Hansen, 2007), which we found to be a comprehensive program for conducting optimal matching. To our knowledge, there is no current package available in either Stata or R that can be used to conduct the nonbipartite matching. To obtain from the Internet a user-developed program in Stata, you may use the `findit` command followed by the name of the program (e.g., `findit psmatch2`) and then follow online instructions to install the program. All user-developed programs contain a help file that offers basic instructions for running the program. The `psmatch2` program implements full Mahalanobis matching and a variety of propensity score matching methods (e.g., greedy matching and propensity score matching with nonparametric regression). Table 5.1 exhibits the syntax and output of three `psmatch2` examples: nearest neighbor matching within a caliper of .250P, Mahalanobis matching without propensity scores, and Mahalanobis matching with propensity scores. In our nearest neighbor matching example, we used propensity scores (named `logit3`) estimated by a foreign program. In other words, we input propensity scores into `psmatch2`. This `psmatch2` step is often needed if users want to use programs such as R-`gbm` or other software packages to estimate propensity scores. Alternatively, users can specify the names of the conditioning variables and let the program estimate the propensity scores directly. A few cautionary statements about running `psmatch2` are worth mentioning. When one treated case is found, several nontreated cases—each of which has 210 the same value of propensity score—may be tied. In a 1-to-1 match, identifying which of the tied cases was the matched case depends on the order of the data. Thus, it is important to first create a random variable and then sort data using this variable. To guarantee consistent results from session to session, users must control for the seed number by using a set seed command. For nearest neighbor and Mahalanobis matching, the literature (e.g., D'Agostino, 1998) has suggested the use of nonreplacement. That is, once a treated case is matched to one nontreated case, both cases are removed from the pool. Nonreplacement can be done in nearest neighbor matching in `psmatch2` by using `nonreplacement` descending. However, this command does not work for Mahalanobis matching. In the matched sample created by `psmatch2` Mahalanobis, it is possible that one control case can be used as a match for several treated cases. To perform nonreplacement for Mahalanobis, users need to examine matched data carefully, keep one pair of the matched and treated cases in the data set, and delete all pairs that used the matched control more than once. The `boost` program estimates boosted regression for the following link functions: Gaussian (normal), logistic, and Poisson. Table 5.2 exhibits the syntax and output of `boost`. Following `boost`, the analyst specifies the names of the dependent variable and independent variables used in the regression model, the name of link function distribution (logistic), and other specifications of the model. In our example, we specified a maximum number of iterations of 1,000, a training data set of 80%, the name of saved predicted probability as `p`, a maximum of four interactions allowed, a shrinkage coefficient of .0005, and a request for showing the influence of each predictor variable in the output. Table 5.1 Exhibit of Stata `psmatch2` Syntax and Output Running Greedy Matching and Mahalanobis Metric Distance 211 212 213 Source: Data from NSCAW, 2004. Table 5.2 Exhibit of Stata `boost` Syntax and Output Running Propensity Score Model Using GBM 214 Source: Data from Hofferth et al., 2001. Information about running `gbm` in R and the program developed by McCaffrey et al. (2004) can be found at . The imbalance program is used to produce the covariate imbalance statistics `dx` and `dxm` developed by Haviland et al. (2007), and `hdgesl` is a program that performs the Hodges-Lehmann aligned rank test. Both programs are available on the Internet and can be downloaded through executing `findit` command within Stata. A feature of `hdgesl` is that it saves the mean of the outcome variable and the number of participants for treated and control groups within each matched set for future analysis. To run `optmatch`, the analyst needs to first install the free statistical software package R from `www.r-project.org`. After installing and starting R, to obtain the `optmatch` package, go to the "Packages" menu in R, choose the "Load package" feature, select `optmatch` from the list, type `library(optmatch)` at the R command prompt, and type `help(fullmatch)` for instructions. Table 5.3 exhibits syntax and the output of running `optmatch` for creating the propensity scores within the program (i.e., by using R's `glm` function). The example also shows how to perform a full matching, to request a calculation of mean distance and total distance based on the matched sets, and finally to request output showing the 215 stratum structure (i.e., the number of matched sets associated with all possible ratios of treated to control participants after matching). Table 5.3 Exhibit of R Syntax and Output Running Logistic Regression and Full Matching 216 Source: Data from Hofferth et al., 2001. 5.8 EXAMPLES We now present examples to illustrate the various models described in this chapter. Each example is based on analyses of observational data from recent studies. These examples represent substantive interests in three topic areas. The first study is an evaluation of caregivers' use of substance abuse services on the hazard rate of child maltreatment. Technically, the study focuses on a sample of families referred to child welfare, and the outcome is defined as a subsequent report—a rereport—of child maltreatment after referral. This study used a large nationally representative sample and longitudinal data to assess the causal effect of substance abuse services on child welfare outcomes, which is an issue with important implications for both policy makers and child welfare workers (Section 5.8.1). The second study is a causal study of the impacts of poverty and multigenerational dependence on welfare on children's academic development. This study also used a nationally representative sample and longitudinal data to test important hypotheses derived from theoretical models (Sections 5.8.2–5.8.4). The third study is an evaluation of a school-based intervention aimed at promoting children's social competence and decreasing their aggressive behavior. This intervention was originally designed as using a group randomization trial, but in practice the randomization did not work (Sections 5.8.5–5.8.6). Methodologically, these examples illustrate most models depicted in this chapter. Section 5.8.1 demonstrates greedy matching followed by a survival analysis. Section 5.8.2 demonstrates optimal matching and evaluation of 217 covariate imbalance before and after matching. Section 5.8.3 demonstrates the Hodges-Lehmann aligned rank test after optimal matching. Section 5.8.4 demonstrates regression adjustment after optimal pair matching. Section 5.8.5 illustrates greedy matching with multilevel data—note that the outcome analysis in this example employs two models: hierarchical linear modeling assuming the nesting structure of data and cross-classified random effects model assuming cross-classified data. And Section 5.8.6 compares the GBM algorithm developed by Rand Corporation to a user-developed GBM algorithm available in Stata. 5.8.1 Greedy Matching and Subsequent Analysis of Hazard Rates This study uses sample and conditioning variables similar to those used in the example in Section 4.4.1. It analyzes a subsample of 2,758 children from the panel data of the National Survey of Child and Adolescent Well-Being (NSCAW). The primary interest of the study is whether caregivers' use of substance abuse services reduces the likelihood of having a rereport of child maltreatment. Thus, the dependent variable is the timing of a maltreatment rereport 18 months after the baseline interview; study participants who did not have a rereport at the end of the 18-month window are defined as censored. As described in Section 4.4.1, the study subsample was limited to children who lived at home (e.g., they were not in foster care) and whose primary caregiver was female. The study was limited to female caregivers because they constitute the vast majority (90%) of primary caregivers in NSCAW. We conducted a three-step analysis. At Step 1, we used conditioning variables to develop propensity scores. At Step 2, we used nearest neighbor matching within-caliper and Mahalanobis metric matching to

create various matched samples. At Step 3, because the timing to the first maltreatment rereport involves censoring (i.e., we knew only the timing of rereport within an 18month window, and those who did not have rereport by the end of the study window are censored), we conducted a Kaplan-Meier product limit analysis to assess differences on survivor functions between the treated participants (i.e., caregivers who receive substance abuse treatment between baseline and the 18th month) and controls (those who did not receive such services in the same period). The matched samples are a 1-to-1 match (i.e., each treated case matches to only one nontreated case in the resamples). The 1-to-1 match for the rereport analysis was a "three by two by two" design. That is, we used three logistic regression models (i.e., each model specified a different set of conditioning variables to predict the propensity scores of receiving treatment), two matching algorithms (i.e., nearest neighbor within a caliper and Mahalanobis), and two matching specifications (i.e., for nearest neighbor, we used two different specifications on caliper size, and for Mahalanobis, we used one with and one without propensity score as a covariate 218 to calculate the Mahalanobis metric distances). Hence, we tested a total of 12 matching schemes. The design using multiple matching schemes was directly motivated by the need to compare results among varying methods and to test the sensitivity of study findings to various model assumptions. We defined the logit or $\log\left(\frac{1 - e^{-x}}{e^{-x}}\right)$ rather than the predicted probability e^{-x} as a propensity score, because the logit is approximately normally distributed. Table 5.4 presents sample descriptive statistics and three logistic regression models. Among the 2,758 children, 10.8% had female caregivers who received substance abuse treatment and the remaining 89.2% had female caregivers who did not receive treatment services. Bivariate chi-square tests showed that most variables were statistically significant ($p < .05$) before matching, indicating that the covariate distributions were not sufficiently overlapped between the treated and nontreated participants in the original sample. Clearly, the treatment group, overall, had many more problems with substance abuse and greater exposure to risk-related conditions. The three logistic regression models differ in the following ways: Logistic 1 contains all predetermined covariates except four variables measuring service needs, Logistic 2 adds the four service need variables, and Logistic 3 drops the variable "child welfare worker (CWW) report of need for service" because we determined that this variable was highly correlated with the actual delivery of treatment and therefore was not an appropriate covariate for matching. Table 5.5 describes the 12 matching schemes and numbers of participants for the resamples: Schemes 1 to 4 were based on Logistic 1, Schemes 5 to 8 were based on Logistic 2, and Schemes 9 to 12 were based on Logistic 3. Within each set of schemes using the same logistic regression, the first two schemes used nearest neighbor matching within a caliper (i.e., one used a caliper size that is a quarter of the standard deviation of the propensity scores or .25oP, and the other employed a more restrictive or narrowed caliper of 0.1), and the next two schemes used Mahalanobis metric matching (i.e., one did not use and the other used the propensity score as a matching covariate). The use of different caliper sizes shows the dilemma we encountered in matching: While a wide caliper results in more matches and a larger sample (i.e., NScheme 1 > NScheme 2, NScheme 5 > NScheme 6, and NScheme 9 > NScheme 10), inexact matching may occur as indicated by large distances on the propensity score between the treated and nontreated cases. We included both caliper sizes in the analysis to test the sensitivity of findings to varying caliper sizes. Note that the sample sizes were the smallest when using Schemes 5 and 6: Using the same nearest neighbor within a caliper of .25oP, the sample size dropped from 564 for Scheme 1 (based on Logistic 1) to 328 for Scheme 5 (based on Logistic 2), which indicated that adding the four need variables greatly restricted successful matching and reduced sample size. Yet further runs indicated that the resample sizes were most sensitive to the inclusion of the variable "CWW report of 219 need." Logistic 3 retained three need variables from Logistic 2 but dropped "CWW report of need," which increased the sample size from 328 for Scheme 5 to 490 for Scheme 9. Because different matching schemes produce different resamples, it is important to check covariate distributions after matching and to examine sensitivity of the results to different resampling strategies. Table 5.6 presents this information. Among these 12 matching schemes, only two matching methods (Schemes 5 and 6) successfully removed all significant differences of covariate distributions between treated and nontreated groups. However, because of the problem of nontrivial reduction in sample size noted earlier, these matching methods did not produce resamples representative of the original sample. Table 5.4 Sample Description and Logistic Regression Models Predicting Propensity Scores (Example 5.8.1) 220 221 Source: Guo, Barth, and Gibbons (2006, Table 1, p. 372–373). Reprinted with permission from Elsevier. Note: Reference group is shown next to the variable name. AOD = alcohol or drug; CIDI-SF = Composite International Diagnostic Interview–Short Form; CWW = child welfare worker. * $p < .05$, ** $p < .01$, *** $p < .001$. All schemes using Mahalanobis matching (Schemes 3, 4, 7, 8, 11, and 12) could not remove significant differences between treated and nontreated groups. Testing the schemes this way suggests that the Mahalanobis approach may not be a good method for matching that involves a large number of matching covariates, such as is the case for this study. Furthermore, using the propensity score as an additional matching variable (Schemes 4, 8, and 12) did not help. Table 5.5 Description of Matching Schemes and Resample Sizes (Example 5.8.1) 222 Source: Guo, Barth, and Gibbons (2006, Table 2, p. 375). Reprinted by permission from Elsevier. Table 5.6 Results of Sensitivity Analyses (Example 5.8.1) 223 Source: Guo, Barth, and Gibbons (2006, Table 3, p. 375). Reprinted by permission from Elsevier. a. Thirty-five study participants were eliminated from the analysis because of missing data. Among the rest of the matching schemes that used nearest neighbor within calipers, only Schemes 9 and 10 successfully removed the significant differences between groups, although the variable CWWREP remains significant in these samples. Both schemes were based on Logistic 3, which excludes CWWREP as a matching covariate. This exclusion was defined because a closer look at the distribution of CWWREP indicated that the 224 CWWREP variable was highly correlated with the dependent variable of the logistic regression, that is, the dichotomous variable for receipt of substance abuse services. In the original sample, 95.5% of the non–service users had a zero need for service use identified by child welfare workers, whereas 74.8% of the service users had an identified need. This is almost certainly likely to have occurred because child welfare workers who observed phenomena such as positive drug tests are more likely to conclude that a caregiver is involved with substance abuse treatment. The presence of a high correlation between CWWREP and the dependent variable of logistic regression prevents the use of CWWREP as a conditioning variable. Thus, we conclude that among the three logistic regressions, Logistic 3 is the best. Table 5.6 also presents results of the survival analysis, that is, the 85th percentile of survivor function associated with each scheme and significance tests on the null hypothesis about equal survivor functions between groups. We report the 90th percentile (instead of the 85th percentile) for Schemes 4, 11, and 12, because the proportion of survivors for groups in these schemes is greater than 85% by the end of the study window. In this analysis, the 85th percentile indicates the number of months it takes for the remaining 15% of study participants to have a maltreatment rereport, and the smaller the number, the sooner the rereport and the greater the risk of maltreatment recurrence. As the statistics show, for the original sample of 2,723 children, it took 7.6 months for 15% of the children whose caregivers used substance abuse services to have a rereport, while it took 13.6 months for 15% of the nontreated children to have a rereport, and the difference is statistically significant ($p < .0001$). Thus, children whose caregivers used substance abuse services were more likely to have a child maltreatment rereport than children whose caregivers did not use the substance abuse treatment services. All matching schemes showed differences in survivor functions in the same direction; that is, the treated group had a higher hazard for rereport than the nontreated group. This finding is consistent across different matching methods, indicating that children of service users have a higher likelihood of being rereported than children of non–service users. The remaining question is whether this difference is statistically significant. Because we know that the methods of nearest neighbor matching within a caliper using Logistic 3 (i.e., Schemes 9 and 10) are the only ones that meet the assumption about ignorable treatment assignment on the covariates, and the group difference on survivor functions is statistically significant in these schemes (i.e., p value is .02), we can conclude that the difference between groups is statistically significant. Children of substance abuse service users appear to live in an environment that elevates risk for maltreatment and, compared with the children of caregivers who were non–service users, warrant continued protective supervision. Figure 5.3 is a graphic representation of the survivor curves comparing the original sample with the resample of Scheme 9. The figure shows that (a) the 225 gap between treated and nontreated groups was slightly wider for the resample of Scheme 9 than for the original sample between months 8 and 12, and (b) by the end of the study window, the proportion of children remaining in a "noreport" state was slightly higher in the original sample than in the Scheme 9 resample. The analysis, based on the original sample without controlling for heterogeneity of service receipt, masked the fact that substance abuse treatment may be a marker for greater risk and an indication for the continued involvement of families with child welfare services. Figure 5.3 Survivor Functions: Percentage Remaining No Rereport (Example 5.8.1) Source: Guo, Barth, and Gibbons (2006, Figure 1, p. 377). Reprinted by permission from Elsevier. 226 In sum, the propensity score matching analysis of the rereport risk enabled an analysis of observational data, when experimental data were unavailable or could not be made available, that provides evidence that substance abuse treatment is not generating safety for children of service users. Additional confirmatory analyses and discussion of these findings are available elsewhere (Barth, Gibbons, & Guo, 2006). 5.8.2 Optimal Matching We now present an example of optimal matching. The example uses the same data and research questions as those presented in Section 2.8.5, a study that investigates intergenerational dependence on welfare and its relation to child academic achievement. For this illustration, we report findings that examine one domain of academic achievement: the age-normed "letter-word identification" score of the Woodcock-Johnson Revised Tests of Achievement (Hofferth et al., 2001). A high score on this measure indicates high achievement. The score is defined as the outcome variable for this study and has the dual virtues of being standardized and representing a key concept in the welfare reform on children's educational attainment. Table 5.7 shows the level of intergenerational dependence on welfare in this sample. Of 1,003 children whose welfare status was compared with that of their caregivers, 615 or 61.3% remained in the same status of not using welfare, 114 or 11.4% showed upward social mobility (i.e., their caregivers used welfare between ages 6 and 12, but the next generation did not use welfare from birth to their current age in 1997), 154 or 15.4% showed downward social mobility (i.e., their caregivers did not use welfare, but the next generation used welfare at some point in their lives), and 120 or 12.0% remained in the same status of using welfare as their caregivers. Thus, the overall level of intergenerational dependence on welfare, as defined within a two-generation period, was 12.0%. Table 5.7 Status of Child's Use of AFDC by Status of Caregiver's Use of AFDC in Childhood (Example 5.8.2) Source: Data from Hofferth et al., 2001. Note: $p < .001$, chi-square test; each percentage (in parentheses) is obtained by dividing the 227 observed frequency by the sample total of 1,003. AFDC = Aid to Families With Dependent Children. On the basis of the research question, we classified the study participants into two groups: those who ever used AFDC from birth to current age in 1997 and those who never used AFDC during the same period. Thus, this dichotomous variable defines the treatment condition in the study: those who ever used AFDC versus controls who never used AFDC. To assess the treatment effect (i.e., child's use of AFDC) on academic achievement, the analyst must control for many covariates or confounding variables. For the purpose of illustration, we considered the following covariates: current income or poverty status, measured as the ratio of family income to poverty threshold in 1996; caregiver's education in 1997, which was measured as years of schooling; caregiver's history of using welfare, which was measured as the number of years (i.e., a continuous variable) caregiver used AFDC between ages 6 and 12; child's race, which was measured as African American versus non-African American; child's age in 1997; and child's gender, which was measured as male versus female. Table 5.8 shows sample descriptive statistics, an independent sample t test on the ATE, and an OLS regression evaluating the ATE. The p values associated with covariates were provided by the Wilcoxon rank sum (Mann–Whitney) test. As the table shows, except for child's gender, the difference on each covariate between treated and control groups is statistically significant. The treated group tended to be those who were poorer in 1996 ($p < .000$), whose caregivers had a lower level of education in 1997 ($p < .000$), and whose caregivers used AFDC for a longer time in childhood ($p < .000$). In addition, the treated group had a larger percentage of African Americans ($p < .000$), and the treated group was on average older in 1997 ($p < .001$) than the control group. Without controlling for these covariates, the study's estimate of treatment effect would be biased. The table also presents two estimates of ATE. One estimate was derived from the independent sample t test, which shows that the treated group on average has a mean letter-word identification score that was 9.82 points lower than that of the control group ($p < .000$). The other estimate was obtained from the OLS regression with robust estimates of standard errors to control for the clustering of children within families. This second estimate is the model most studies would use, which shows that controlling for these covariates, the treated group on average has a letter-word identification score that is 4.73 points lower than that of the control group ($p < .01$). On the basis of Chapters 2 and 3, we could say with reasonable confidence that both estimates are likely to be biased and inconsistent. Our next step was to conduct a propensity score analysis. At Step 1, we used GBM (i.e., Stata boost) to estimate the propensity scores of receiving treatment. The GBM showed that the ratio of family income to poverty in 1996 had the strongest influence on the likelihood function (86.74%), caregiver's use of 228 AFDC in childhood had the next strongest influence (6.03%), and the remaining influences included caregiver's education in 1997 (3.70%), child's race (2.58%), and child's age in 1997 (0.95%). Gender was not a significant predictor shown by the Wilcoxon rank sum test; therefore, the GBM did not use child's gender. Figure 5.4 shows the box plots and histograms of the estimated propensity scores by treatment status. As the figure indicates, the two groups differ from each other on the distribution of estimated propensity scores. The common support region is especially problematic because the user group's 25th percentile is equal to the nonuser group's upper adjacent value. The two groups share a very narrow common support region. If we applied nearest neighbor matching within a caliper or other types of greedy matching, it is likely that the narrow common support region would produce a nontrivial loss of matched participants. In addition, greedy matching identifies matches for each treated participant in a nonoptimal fashion. On the basis of these concerns, we decided to use optimal matching. Table 5.8 Sample Description and Results of Regression Analysis (Example 5.8.2) Source: Data from Hofferth et al., 2001. Note: AFDC = Aid to Families With Dependent Children. a. Independent sample t test, two-tailed. b. Wilcoxon rank sum (Mann-Whitney) test. * $p < .05$, ** $p < .01$, *** $p < .001$. 229 Figure 5.4 Distribution of Estimated Propensity Scores (Example 5.8.2) Before running optimal matching, we need to carefully examine the ratio of treated participants to control participants to determine matching schemes. In this sample, we have 274 treated participants and 729 controls or a ratio of treated participants to controls of approximately 0.38:1. With these data, a 1-to1 or 1-to-2 pair matching scheme is feasible. However, optimal pair matching generally is not optimal, especially when compared with full matching based on the same data (see Section 5.4.2), and a 1-to-1 matching will make 729 – 274 = 455 controls nonusable, and a 1-to-2 matching will make 729 – (274 \times 2) = 181 controls nonusable. For purposes of illustration, we decided to run a 1-to-1 pair matching. We also found that with these data, a full matching or variable matching is permissible. Using principles and considerations described in Section 5.4.2, we used the following matching schemes: (a) full matching; (b) 230 Variable Matching 1, which uses at least one and at most four controls for each treated participant; (c) Variable Matching 2, which uses at least two and at most four controls for each treated participant; (d) Variable Matching 3, which uses Hansen's equation (i.e., it specifies that the minimum number of matched controls is $.5(1 - \lambda) = .5(1 - .273)/.273 = 1.33$, and the maximum number of matched controls is $2(1 - \lambda) = 2(1 - .273)/.273 = 5.32$); (e) Variable Matching 4, which uses at least two and at most seven controls for each treated participant; and (f) Pair Matching 1-to-1. We ran optmatch in R to implement optimal matching with these schemes. Table 5.9 presents optimal matching results. Two useful statistics produced by optmatch are shown in the table: stratum structure and total distance. The stratum structure is a count of matched sets in terms of the ratio of the number of treated participants to the number of controls. For instance, the full matching produced a total of 135 matched sets; of these, 1 set had 22 treated participants and 1 control, 40 sets had 1 treated participant and 1 control, 20 sets had 1 treated participant and 2 controls, and so forth. Note that one set contains 1 treated participant and 245 controls. Total distance is the sum of differences on propensity scores between treated and control participants over all matched sets, and thus, total distance is an overall measure of the closeness of matching, with a small number indicating a close match. Using total distance, we found that among all six schemes, full matching offered the best approach, and pair matching was the second best. Of the four variable-matching schemes, Variable Matching 3 using Hansen's equation worked the best. It is worth noting that the 1-to-1 pair matching did not use 455 controls, which means there was a loss of 45.4% of study cases. Although pair matching showed closeness in matching, it eliminates study cases, which exerts an undesirable impact on study power and representativeness of the analytic sample. Our data confirmed previous findings that full matching is the best. With matched samples, we always want to know how well matching has reduced bias. The level of bias reduction can be shown by a comparison between the absolute standardized differences in covariate means before and after matching (i.e., a comparison between dx and dxm). Table 5.10 presents this information. Full matching greatly reduces covariate imbalance on all variables except gender. Taking the ratio of family income to poverty in 1996 as an example, before matching, the treated and control groups differ on this variable by more than 100% of a standard deviation, whereas after full matching, the standard bias is only 4% of a standard deviation. Nearly all matching schemes reduce bias for almost all variables to some extent, but some do more, and some do less. In terms of covariate balancing, Table 5.10 confirms that full matching worked the best and pair matching the second best. Table 5.9 Results of Optimal Matching (Example 5.8.2) 231 Source: Data from Hofferth et al., 2001. Table 5.10 Covariate Imbalance Before and After Matching by Matching Scheme (Example 5.8.2) 232 233 Source: Data from Hofferth et al., 2001. Note: Absolute standardized difference in covariate means, before (dx) and after matching (dxm). AFDC = Aid to Families With Dependent Children. 5.8.3 Post–Full Matching Analysis Using the Hodges-Lehmann Aligned Rank Test The Hodges-Lehmann aligned rank test can be used on matched samples created by full matching or variable matching. Because none of the variable matching schemes showed satisfactory bias reduction, we ruled out these matched samples for further analysis. Table 5.11 presents results of the post–full matching analysis. Table 5.11 Estimated Average Treatment Effect on Letter-Word Identification Score in 1997 With Hodges-Lehmann Aligned Rank Test (Matching Scheme: Full Matching) (Example 5.8.3) 234 Source: Data from Hofferth et al., 2001. As the table shows, we used full matching and found that children who used AFDC had a letter-word identification score in 1997 that was, on average, 1.97 points lower than those who had never used AFDC; the difference was statistically significant at a .05 level. We used the Hodges-Lehmann test to gauge the statistical significance. The study also detected an effect size of .19, which is a small effect size in terms of Cohen's (1988) criteria. 5.8.4 Post–Pair Matching Analysis Using Regression of Difference Scores A regression of difference scores may be performed on matched samples created by optimal pair matching. Based on the pair-matched sample, we first calculated difference scores between treated and control cases for each pair on all study variables (i.e., on the outcome variable and all covariates). We then regressed the difference score of the outcome on the difference scores of covariates. In addition, note that our model includes a correction for clustering effects (i.e., children are nested within caregivers) that we accomplished by using robust estimates of standard errors. Table 5.12 presents results of the post–pair matching analysis of regression adjustment. As explained in Section 5.5.4, the intercept of a difference score regression indicates the ATE of the sample. The estimated intercept from this model is -3.17 ($p < .05$). Thus, using pair matching and regression adjustment, the study found that, on average, children who used AFDC had a letter-word identification score in 1997 that was 3.17 points lower than do children who never used AFDC; this finding was statistically significant. 5.8.5 Multilevel Propensity Score Analysis This example illustrates the application of a multilevel propensity score analysis to data that are generated by a group randomization. This is the same example discussed in Chapter 1 (i.e., Example 6). In the current presentation, we use the North Carolina data of the Social and Character Development (SACD) project to illustrate the importance of controlling for clustering effects when estimating propensity scores and conducting the postmatching outcome analysis. The example uses the same data and research questions presented in Section 4.4.2. The outcome variable for the current analysis is a change score, and there are two of them: students' change in the fourth grade on Prosocial Behavior from the Carolina Child Checklist (CCCPROS) and students' change 235 in the third grade on Relational Aggression from the Carolina Child Checklist (CCCRAG). In North Carolina, the SACD project randomized schools within two school districts to assign treatment conditions. Included in the current analysis are students from six schools receiving treatment and six schools not receiving treatment. Because the number of schools was very small, the group randomization failed to balance data. As such, the study sample was imbalanced on numerous covariates at both the person and school levels. Had these imbalances been ignored and the failed randomization been treated as a "successful randomization," the evaluation of SACD treatment effects would have been biased and suffered from low internal validity. In these situations, evaluators must consider using a multilevel propensity score analysis. Table 5.12 Regressing Difference Score of Letter-Word Identification on Difference Scores of Covariates After Pair Matching (Example 5.8.4) Source: Data from Hofferth et al., 2001. * $p < .05$, one-tailed test. In this kind of analysis, the first task is to estimate propensity scores for the sample students. Clearly, the nature of group randomization makes the study students nested within schools and a multilevel analysis is called for. Of the five models depicted in Section 5.6.3, the fixed effects model was out of consideration in the SACD analysis, because the dependent variable of the logistic regression (receiving vs. not receiving treatment) was perfectly predictable by the cluster indicators; the multilevel logistic regression with either a narrow or a broad inference space did not work well due also to group randomization. Indeed, the multilevel logistic regression did not converge. As a consequence, we used the single-level model (i.e., the model pools together both person- and school-level variables but ignores the clustering structure) and the single cluster-level model (i.e., the model employs school-level variables only and runs at the school level with an N of 12). A nearest neighbor withincaliper matching that specified a caliper size of a quarter of the standard 236 deviation of the estimated propensity scores was used to create the matched sample. Tables 5.13 and 5.14 present the results of the multilevel propensity score analysis for students' change on the CCCPROS score in the fourth grade. After a listwise deletion of missing data, the original sample comprised 554 students from 12 schools. The ICC of the original sample was 0.185, suggesting that 18.5% of the variation in the change score of CCCPROS lies between schools, which is nontrivial and warrants a control in the outcome analysis. As Table 5.13 indicates, three variables from the single-level model were statistically significant (i.e., race [African American], school Adequate Yearly Progress at baseline, and school percentage of minority students at baseline); none of the variables from the school-level model was significant, primarily due to the small sample size at the school level ($N = 12$). Before matching, five variables were statistically significant from bivariate imbalance checks. After matching, none of these variables from the matched sample based on the single-level model was significant, but two of them from the matched sample based on the school-level model remained significant. The results suggest that the singlelevel model works better than the school-level model in terms of balancing the observed covariates. After matching on the propensity scores estimated by the single-level model, we found that the ICC using school as a clustering variable was 0.183, and the ICC using the matched set as a clustering variable was 0.075. For this data set, the clustering of outcome within matched sets was not high, but was greater than .05. Table 5.14 presents results of the outcome analysis using four models: (1) two-level HLM based on the original sample, (2) two-level HLM based on the matched sample using propensity scores estimated by the single-level model, (3) two-level HLM based on the matched sample using propensity scores estimated by the school-level model, and (4) CCREM based on the matched sample using propensity scores estimated by the single-level model. Note that the CCREM was based on the matched sample using the single-level model, because it is this model that removes all imbalances. On the basis of the coding scheme, we know that a positive sign on the coefficient of treatment variable indicates a beneficial treatment effect. Results from all four models indicate that the SACD treatment produces beneficial and statistically significant effects on students' change on the CCCPROS score in the fourth grade ($p < .05$), including the original sample without controlling for selection bias. However, the model correcting for selection bias and both sources of clustering effects (i.e., the CCREM) shows a larger treatment effect than the original sample with no control of the selection bias: The difference between the two models is $0.256 - 0.231 = 0.025$, or 10.8% larger. Tables 5.15 and 5.16 present the results of the multilevel propensity score analysis for student change on CCRAG score in the third grade. The tables have the same setup as Tables 5.13 and 5.14. The original sample comprises 237 587 students from 12 schools. The ICC of the original sample was 0.078, suggesting that 7.8% of the variation in the change score of CCCRAG lies between schools. Table 5.15 shows that three variables from the single-level model were statistically significant, and none of the variables from the schoollevel model was significant. Before matching, five variables were statistically significant from bivariate imbalance checks. After matching, two of these variables from the matched sample using the single-level model were significant, but none of them from the matched sample using the school-level model was significant. The results suggest that the school-level model works better than the single-level model in terms of balancing the observed covariates. After matching using the propensity scores estimated by the school-level model, we found that the ICC using school as a clustering variable was 0.051, and the ICC using the matched set as a clustering variable was 0.191. Note that the clustering within matched sets is nontrivial. Because the matching based on the school-level model worked better than the matching based on the single-level model (i.e., no significant covariates remained significant from the school-level model), it was important to control for the clustering of both sources by running CCREM based on the matched sample derived from the school-level model. Table 5.16 presents results of the outcome analysis using four models. All models are equivalent to those presented in Table 5.14 except that the CCREM is based on the matched sample using propensity scores estimated by the schoollevel model. On the basis of the coding scheme, we know that a positive sign of the coefficient of treatment variable indicates a beneficial effect. Results from all four models indicate a negative sign of the treatment variable, indicating a detrimental effect of the treatment. However, only the effect from the original sample that does not control for selectivity is statistically significant ($p < .05$); all models using propensity score matching show a nonsignificant difference between treated and nontreated students. Of the three propensity score models, CCREM may be considered the final model due to findings discussed above. It is important to note that the propensity score analysis reveals categorically a different finding than the original uncorrected model: There is no significant difference between treated and nontreated students, rather than a finding that suggests intervention had an unanticipated negative effect. Table 5.13 Propensity Score Models and Imbalance Check for the Change of "CCCPROS" in the Fourth Grade (Example 5.8.5) 238 Source: Data from SACD, 2008. Note: AYP = Adequate Yearly Progress. ** $p < .01$, *** $p < .001$, two-tailed test. Table 5.14 Estimated Coefficients by Multilevel Model for the Change of "CCCPROS" in the Fourth Grade (Example 5.8.5) 239 Source: Data from SACD, 2008. * $p < .05$, + $p < .1$, one-tailed test for the treatment effect and two-tailed for all other variables. Table 5.15 Propensity Score Models and Imbalance Check for the Change of "CCCRAG" in the Third Grade (Example 5.8.5) 240 Source: Data from SACD, 2008. Note: AYP = Adequate Yearly Progress. ** $p < .01$, *** $p < .001$, two-tailed test. Table 5.16 Estimated Coefficients by Multilevel Model for the Change of "CCCRAG" in the Third Grade (Example 5.8.5) 241 Source: Data from SACD, 2008. Note: AYP = Adequate Yearly Progress. * $p < .05$, + $p < .1$, one-tailed test for the treatment effect and two-tailed for all other variables. 5.8.6 Comparison of Rand-gbm and Stata's boost Algorithms In this final example, we want to show results of a study serving solely a methodological purpose: comparison of the GBM procedure developed by McCaffrey et al. (2004) to the general GBM algorithm. As discussed in Section 5.3.4, GBM aims to minimize sample prediction error; that is, the algorithm stops iterations when the sample prediction error is minimized. This is the standard setup for both Stata's boost program and R's gbm program. McCaffrey et al. (2004) altered this procedure by stopping the algorithm at the number of iterations that minimizes the ASAM in covariates (thus we refer to this 242 algorithm as Rand-gbm). In other words, the Rand-gbm is a procedure tailored more specifically to propensity score analysis. We are interested in the following question: To what extent do results from Rand-gbm differ from those produced by a regular GBM algorithm? To answer this question, we compared the Rand-gbm with Stata's boost. We have applied different data sets to conduct the comparisons and obtained almost the same findings. Although we believe that our findings are sufficiently important to be shared with readers, we should emphasize that our comparison is not a Monte Carlo study, and results may not hold in other settings or types of data. We performed our comparison using data from an evaluation of a schoolbased intervention in which group randomization failed. For more information about the study sample and data, see Section 4.4.2. We used boosted regression to estimate propensity scores. Figure 5.5 presents distributions of estimated propensity scores by Rand-gbm and Stata's boost. Using the same set of conditioning variables, Rand-gbm and Stata's boost produced quite different propensity scores. The Rand-gbm propensity scores, shown in the histograms in Figure 5.5, have a high level of dispersion with a range between .2 and .8. In contrast, propensity scores estimated by Stata's boost have a low level of dispersion and concentrate on the range between .4 and .6. As a result, the two methods produced very different box plots: Rand-gbm does not show much overlapping of the propensity scores between treated and control participants while Stata's boost does. Figure 5.5 Comparison of Estimated Propensity Scores Generated by Rand-gbm and Those Generated by Stata's boost (Example 5.8.6) 243 Using the two sets of propensity scores, we then conducted pair-matching and postmatching analysis. We were interested in whether the two sets of propensity scores would produce different results in the subsequent analyses. Table 5.17 shows results of the covariate-imbalance check. As the last two columns of Table 5.17 (i.e., the statistics of dxm) show, even though the two sets of propensity scores have different distributions, both corrected for (or failed to correct for) imbalance in very similar fashion. Based on the pair-matched samples, we then analyzed difference score regression (Table 5.18). Results show that the treatment effect (i.e., constant of the regression model) estimated by Rand-gbm is 0.15 and that estimated by Stata's boost is 0.13. Both effects are not statistically significant. Thus, the two sets of data provide approximately the same findings. Given the results, we conclude that Rand-gbm and Stata's boost do not lead to different results on covariate control and estimates of treatment effects, although this finding needs to be verified in future studies. Table 5.17 Comparison of Covariate Imbalance Before and After Matching Between Rand-gbm and Stata's boost (Example 5.8.6) 244 Source: Data from SACD, 2008. Note: Absolute standardized difference in covariate means, before (dx) and after matching (dxm). Table 5.18 Regressing Difference Score of Outcome (i.e., Change of Academic Competence in Third Grade) on Difference Scores of Covariates After Pair Matching: Comparison of Results Between Rand-gbm and Stata's boost (Example 5.8.6) Source: Data from SACD, 2008. * $p < .05$, + $p < .10$, two-tailed test. 245 5.9 CONCLUSION Before we conclude this chapter, we need to point out some limitations of propensity score matching. According to Rubin (1997), propensity scores (a) cannot adjust for unobserved covariates, (b) work better in larger samples, and (c) do not handle a covariate that is related to treatment assignment but is not related to outcome, in the same way as a covariate with the same relation to treatment assignment but strongly related to outcome. Rubin recommended performing sensitivity analysis and testing different sets of conditioning variables to address the first limitation. Michalopoulos et al. (2004) also noted that propensity scores correct less well for studies in which the treated and nontreated groups are not from the same social context/milieu and, therefore, are not exposed to the same ecological influences. This is a special case of being unable to adjust for unobserved covariates common in social service and other program evaluations that compare across service jurisdictions. It is also worth saying again that propensity score matching is a rapidly growing field of study, and many new developments are still in a testing stage. Additional problems as well as strategies may be identified as researchers move to new developments. We agree that even randomized clinical trials are imperfect ways of determining the result of treatment for every member of the population—treated or not. Nor can propensity score methods provide definitive answers to questions of treatment effectiveness. Multiple methods for estimating program effects are indicated for use within and across studies. Researchers using propensity score matching should be cautious about these limitations and make efforts to warrant that interpretation of study results does not go beyond the limits of data and analytical methods. Nonetheless, from this chapter we have seen that propensity score matching is a promising approach that offers a growing evidentiary base for observational studies facing violations of the unconfoundedness assumption and selection biases. NOTES 1. We describe the bias-adjusted matching estimator in Chapter 8. 2. The calculation of ASAM will be described in Section 5.5.2. 246 CHAPTER 6 Propensity Score Subclassification Propensity score matching has the advantage of addressing the dimensionality problem in data balancing, and it has been increasingly employed in research in the social and health sciences. The greedy matching approach provides for the analysis of virtually any type of outcome, including categorical, ordinal, or censored measures. But greedy matching tends to compromise sample sizes. Optimal full or variable matching partially addresses this problem but is limited to analyses of continuous outcome variables. Subclassification resolves some of these problems. This chapter describes the subclassification approach, a method that has been widely employed in biomedical and epidemiological research and that permits researchers to conduct any kind of multivariate outcome analysis while retaining a large portion of the original sample size. Because the overlap assumption is embedded in most propensity score models, this chapter also discusses this assumption explicitly and presents a method that addresses its violation—that is, trimming data at the lower and upper regions of estimated propensity scores to address the limited overlap of propensity scores between treatment groups. In the social and health sciences, researchers often need to evaluate complex relationships between variables, such as the mediating, moderating, and nonrecursive effects of variables. Structural equation modeling (SEM) is the dominant approach to investigations of this type. However, SEM, in its original form, suffers from the same problem of endogeneity as that of regression and cannot be applied directly to observational studies. In this chapter, we discuss emerging ideas regarding the integration of SEM and propensity score analysis into one model, primarily how to conduct SEM by using propensity score subclassification. The chapter also describes a recently developed method that uses subclassification and multilevel analysis to model treatment effect heterogeneity. Section 6.1 provides an overview of propensity score subclassification. The overview describes steps of stratifying a study sample using user-defined quantiles, conducting balance checks, and employing equations to aggregate treatment effects estimated by each subclass to obtain an estimated treatment effect for the entire sample. In addition, we discuss a process for discerning whether the overall treatment effect for a sample is statistically significant. 247 Section 6.2 reviews the overlap assumption and a trimming approach to address limited overlap in empirical data (Crump, Hotz, Imbens, & Mitnik, 2009). Section 6.3 presents a framework for integrating propensity score analysis and SEM into one model and steps in using propensity score subclassification to conduct SEM. The main methodological issue the integrated model addresses is the stability of model parameters across subclasses and, hence, how to test a series of hypotheses to execute a multigroup comparison. Section 6.4 describes the stratification-multilevel method that models treatment effect heterogeneity. Section 6.5 presents examples of selected models. Section 6.6 provides a summary and conclusion. 6.1 OVERVIEW The central idea of employing subclassification to balance data was developed by Cochran (1968) and formulated even before the development of propensity score analysis. Section 3.1 in Chapter 3 describes the basic ideas of Cochran's framework. Suppose x is a variable known to cause selection bias. Researchers can stratify the entire sample by subclasses of x such that within a stratum, study participants are homogeneous on x , and therefore, the estimation of treatment effect using the standard estimator (i.e., evaluation of the difference on the outcome between treated and untreated participants within the stratum) is net of the influence of x , and the selection bias due to x is removed. Repeating this process for all strata, researchers can obtain

an estimate of average treatment effect (ATE) for the entire sample, and this ATE is bias free on x . A subclassification formed in this way is called exact subclassification on x . Researchers may perform exact subclassification by using more than one variable. Suppose race and gender are two covariates that make the treated and untreated groups imbalanced, and race is operationalized as having five groups. With five race and two gender groups, researchers can form 10 subclasses such that in each subclass, participants are all identical on a given race and gender identity. By calculating treatment effect within each subclass and aggregating ATE over all 10 subclasses, the sample ATE is net of selections on race and gender and is unbiased. This in practice is feasible. However, exact subclassification suffers from the same dimensionality problem as matching. When the number of covariates increases, as noted by Cochran, the number of subclasses or strata grows exponentially. For instance, if all covariates were dichotomous variables, then there would be 2^p subclasses for p covariates. If p is moderately large, then some strata might contain no units, and many strata might contain either treated or control units but not both, which would make it impossible to estimate a treatment effect in that stratum (Rosenbaum & Rubin, 1984). It is from here that we see the advantage of using a coarsest balancing score (i.e., the propensity score) rather than several finest scores (i.e., the observed covariates) to stratify. The advantage offered by a 248 propensity score is the reduction of dimensionality in subclassification—that is, with an estimated propensity score, researchers only need to stratify the sample based on one score. This was the contribution made by Rosenbaum and Rubin (1983) in their seminal paper. Rosenbaum and Rubin have proven that a propensity score $e(x)$ is a balancing score $b(x)$ that best represents all covariates or vector x and reduces the multiple variables into one score on which researchers can successfully perform subclassification. Corollary 4.2 of Rosenbaum and Rubin (1983) argues for the balancing property of propensity score analysis with subclassification: Suppose treatment assignment is strongly ignorable. Suppose further that a group of units is sampled using $b(x)$ such that: (i) $b(x)$ is constant for all units in the group, and (ii) at least one unit in the group received each treatment. Then, for these units, the expected difference in treatment means equals the average treatment effect at that value of $b(x)$. Moreover, the weighted average of such differences, that is, the directly adjusted difference, is unbiased for the treatment effect, when the weights equal the fraction of the population at $b(x)$. (p. 46) Specifically, a propensity score subclassification comprises a process of five intuitive steps: (1) sort the sample by estimated propensity scores in an ascending order; (2) divide the sample into K strata using quantiles (quintiles, deciles, or other) of the estimated propensity scores; (3) evaluate the treatment effect by calculating mean difference of outcome and variance of difference between treated and control participants within each stratum, or by running a multivariate analysis of outcomes within each stratum as one does for samples generated by a randomized experiment; (4) estimate the mean difference (ATE) for the whole sample (i.e., all K strata combined) through aggregating; and (5) test whether the sample difference on outcome is statistically significant. Technically, let $0 = c_0 < c_1 < c_2 < \dots < c_K = 1$ be boundary values; let B_{ik} be the indicators defined as $B_{ik} = 1$ if $c_{k-1} < e(x_i) < c_k$, $B_{ik} = 0$ otherwise, and where i is the index of observation ($i = 1, \dots, n_k$, n_k is the number of observations in stratum k), k is the index of stratum ($k = 1, \dots, K - 1$), and $e(x_i)$ is the propensity score for i . Now the average treatment effect within stratum k can be evaluated by applying the standard estimator to stratum k , or $\tau_k = E[Y_{1i} - Y_{0i} | B_{ik} = 1]$. With sample data, the stratum's ATE is estimated by where and $\omega = 1$ or 0 indicating treatment or control status. Imbens and Wooldridge (2009) explain the condition under which Corollary 4.2 of Rosenbaum and Rubin (1983) regarding the constant propensity score property holds, that is, if K is sufficiently large and the differences $c_k - c_{k-1}$ are small; under this condition, "there is little variation in the propensity score within a stratum or block, and one can analyze the data as if the propensity score constant, and thus as if the data within a block were generated by a completely randomized experiment (with the assignment probabilities constant within a stratum, but varying between strata)" (p. 33). The average treatment effect (ATE) for the whole sample is then estimated by using the weighted average of the within-stratum estimates: where N is the total number of participants. The variance of the sample ATE is estimated by the following formula: Using—that is, by taking a square root of variance, one obtains a standard error of—one then can perform a significance test of a nondirectional (i.e., perform a two-tailed test) or a directional (i.e., perform a one-tailed test) hypothesis. Estimating ATE and testing its statistical significance for the entire sample using propensity score subclassification are widely used by researchers in both the social and health sciences. However, ATE is not the only statistic that can be estimated through subclassification. The method can be extended to estimation of other treatment effects, such as the average treatment effect for the treated (ATT). For simplicity of exposition, we focus on ATE in this book. Equations 6.1 and 6.2 are evaluations of ATE using means and may be performed after stratifying the sample based on quantiles of propensity scores (i.e., performing the analysis of Step 2c shown by Figure 5.1) or after propensity score matching and then stratifying the matched sample based on quantiles of propensity scores (i.e., performing the analysis of Step 3b shown by Figure 5.1). Regardless of whether you use matching, propensity score subclassification conducted in this manner (i.e., computing means) is essentially a bivariate analysis of outcome and does not control for covariates of outcome. The advantage of propensity score subclassification is that the method can be employed in conjunction with a multivariate outcome analysis. The method accommodates most types of multivariate models such as ordinary least squares (OLS) regression, structural equation modeling, survival analysis, multinomial logit model, and ordered logistic regression. In other words, the analysis conducted in Step 2c of Figure 5.1 encompasses most types of multivariate models. Suppose the OLS regression within stratum k is $Y_{ik} = \alpha_k + \tau_k W_{ik} + \beta_k' X_{ik} + e_{ik}$; then, aggregating estimated treatment effects over all k strata, one obtains an estimated ATE for the entire sample. The formula for the weighted 250 average of the within-stratum estimates is After running the regression model, one also obtains the standard error for each stratum. The sample overall variance can be obtained in a similar fashion as that for Equation 6.2. That is, one first computes variance of by then aggregating variances over all k strata, taking squares of one obtains the variance for the entire sample, as Next, taking the square root of the sample variance one obtains the standard error for the entire sample. Finally, one can perform a . The standard significance test of the overall treatment effect by using error of the treatment effect within each stratum is that associated with the dichotomous treatment variable W , as long as one specifies such a treatment indicator in a multivariate model and is normally reported by any software package. This could be the standard error of the coefficient of the treatment indicator from a survival model (e.g., the Cox proportional hazards model), S E M, multilevel modeling, or a multinomial logit model. Because these multivariate models assume that the test statistic is subject to a standard normal distribution, it is valid to apply the above procedure to perform a z test to discern whether the sample-aggregated treatment effect is statistically significant, assuming that the sample size of each stratum is sufficiently large and running the multivariate outcome analysis within stratum is feasible. It is from here that we see the attractive property of propensity score subclassification—that is, it permits most multivariate outcome analysis and allows researchers to further control for covariates by performing a multivariate, rather than a bivariate, analysis. Compared to greedy matching, subclassification generally retains the original sample size in the outcome analysis. The ATE and variance estimators shown by Equations 6.1 to 6.4 are the most popular ones in the applications of propensity score subclassification. There are revised versions of the variance estimator when researchers need to deal with more complex research issues. For instance, Rosenbaum and Rubin (1984) employed the estimator of Mosteller and Tukey (1977) to calculate standard errors of directly adjusted probabilities of survival. When analyzing data with a complex survey design, researchers need to incorporate sampling weights into the analysis. Zanutto (2006) provides a survey-weighted version of 6.1 to estimate ATE and suggests an approximate variance estimator developed by Lohr (1999) that is analogous to Equation 6.4. There are a few important issues worth discussing, particularly caveats users 251 need to keep in mind when conducting propensity score subclassification. These issues and caveats are summarized below. 1. The crucial condition for a successful propensity score subclassification is to make the propensity score constant within a stratum, or to meet the requirement of making K sufficiently large and the differences $c_k - c_{k-1}$ small (Imbens & Wooldridge, 2009). In practice, what is an optimal size of K ? Cochran (1968) uses one variable age to stratify and finds that a subclassification of five groups efficiently removes 90% of the bias. Rosenbaum and Rubin (1984) extend this finding to propensity score subclassification and advocate five subclasses. In practice, this is also the choice in many published applications (Ahmed et al. 2006; Austin, 2009; Cao, 2010; D'Agostino, 2007; Leow, Marcus, Zanutto, & Boruch, 2004; Lunceford & Davidian, 2004; Perkins, Tu, Underhill, Zhou, & Murray, 2000; Rubin, 2001; Zanutto, 2006). However, variation exists. For instance, based on theoretical work and data simulation, Lunceford and Davidian (2004) show that subclassification using quintile (i.e., forming five strata) may not necessarily remove bias, and a correct specification of a regression model for the outcome analysis is essential; ultimately, the number of strata should be determined by the extent to which the subclassification achieves covariate balances. In studying whether 401(k) eligibility increased saving, Benjamin (2003) used 10 groups to subclassify a sample and showed that the subclassification sufficiently balanced each covariate by using t tests or z tests (for binary covariate) within each stratum. Harder, Stuart, and Anthony (2010) used 10 groups initially but made adjustments subsequently; specifically, they found that the first four groups had a small number of observations in the control group; they then collapsed these four groups into one and created seven groups in the final analysis. When deciding the number of optimal strata, researchers encounter a dilemma: Using more subclasses creates better homogeneity within subclasses and reduces bias in treatment effect estimates, but it also results in fewer observations in each subclass and therefore less precise treatment estimates (Du, 1998). Ultimately, researchers need to test different numbers of strata and choose one that optimizes the estimation of treatment effect while balancing covariates within the stratum. 2. Researchers need be aware of the assumptions embedded in Equations 6.2 and 6.4 exercise caution when performing significance tests that assume a standard normal distribution. Tu and Zhou (2003) argue that the validity of statistical inference based on propensity score subclassification hinges not only on the assumption of the normality of the point estimator τ but also on several rather implicit assumptions: (1) The cut points of the subclasses are fixed, (2) the responses are independent across all the subclasses, and (3) within each subclass, the responses from the treated and control subjects are independent. 252 These assumptions, as shown by Du (1998), are prone to violation in empirical data because the subclassification is based on orders of estimated propensity scores and hence introduces an ordered statistical structure into the problem. Consequently, the resulting subclassification destroys the original independent data structure (both within and between subclasses) and the variance estimator 6.2 or 6.4 becomes incorrect. Tu, Perkins, Zhou, and Murray (1999) further show that even if the independent structure were maintained, the variance formula for a stratified random sample should not be used in propensity score—based inferences because it also fails to account for the uncertainty associated with the estimation of the propensity score. On the basis of these findings, Tu and Zhou (2003) developed a procedure to use bootstrapped confidence intervals for significance tests of overall treatment effects. It should be noted, however, that previous studies (Agodini & Dynarski, 2001; Benjamin, 2003) have found that the variance estimator shown by either Equation 6.2 or 6.4 can be a reasonable approximation. 3. Testing for covariate balance after subclassification is crucial, and researchers should routinely perform within-stratum balance checks; if imbalances of covariates remain, they should consider strategies correcting for imbalances. This process typically requires resubclassifying the sample. Recent systematic reviews of studies applying propensity score methods have demonstrated that the methods are inconsistently used and frequently poorly applied in the medical literature (Shah, Lalpaci, Hux, & Austin, 2005; Weitzen, Lapan, Toledano, Hume, & Mor, 2004). Failure to conduct a balance check after a propensity score correction, according to Austin's (2008) critical appraisal of propensity score matching in the medical literature between 1996 and 2003, is a common pitfall. Although the critique was primarily made on propensity score matching, researchers should be cautious and avoid a similar mistake when conducting propensity score subclassification. As discussed in Section 5.3 in Chapter 5 for matching, researchers can use the procedure suggested by Rosenbaum and Rubin (1984, 1985) to employ high-order polynomial terms and/or cross-product interaction terms in the logistic regression predicting the propensity score, and the procedure may be repeatedly executed as follows: running logistic regression, subclassification, balance check within stratum, and rerunning logistic regression if covariate imbalances remain. In the rerunning, users may include a square term of the covariate that shows an imbalance after propensity score correction or a product of two covariates if the correlation between these two covariates is likely to differ between groups. 4. Like most propensity score methods, subclassification also assumes overlap of estimated propensity scores between treated and untreated groups. When this assumption is violated, users often obtain strata that contain either 253 treated or untreated units but not both, and this happens most often in the stratum of lowest propensity scores and/or the stratum of highest propensity scores. To address limited overlap, researchers can perform a trimming procedure to delete observations at the lower and upper regions of the estimated propensity scores from the analysis. Crump et al. (2009) suggest a formula to determine the exact proportion of observations to be trimmed. They also offer a rule of thumb for trimming, that is, discarding all units with estimated propensity scores outside the range $[0.1, 0.9]$, and this rule of thumb approximates the formula for optimal trimming. Because this procedure is so important for most propensity score methods and works well for subclassification analysis, we describe the overlap assumption and the Crump et al. method in the next section. 6.2 THE OVERLAP ASSUMPTION AND METHODS TO ADDRESS ITS VIOLATION The propensity score models developed by Rosenbaum and Rubin (1983) are based on the strongly ignorable treatment assignment assumption. This assumption consists of two components: (1) the treatment assignment is independent of the outcomes, conditional on the observed covariates, or, and (2) the probability of assignment is bounded away from 0 and 1, or $0 < e(x) < 1$, if one uses a propensity score or conditional probability of receiving treatment as a proxy of the assignment probability. The second component implies that for propensity score analysis that assumes unconfoundedness, the estimated propensity scores for all units should be greater than zero and less than 1; more precisely, it should be as follows: For some $c > 0$, and all $x \in X$, $c < e(x) < 1 - c$. In practice, researchers often find that the second component of the assumption does not hold; that is, the overlap in the covariate distributions in the treated and control subpopulations is not sufficient. Even if the supports of the two covariate distributions are identical, there may be parts of the covariate space with limited numbers of observations for either the treatment or control group. Such areas of limited overlap can lead conventional estimators of average treatment effects to produce substantial bias and large variances (Crump et al., 2009). In practice, researchers often discard units for which there is no close counterpart in the subsample with the opposite treatment. However, this ad hoc approach lacks theoretic support and statistical rationale. It is inefficient. To address the limited overlap problem, Crump et al. (2009) develop a systematic approach to deal with the problem. They proved that the limited overlap can be efficiently addressed by discarding observations with propensity scores outside an interval $[\alpha, 1 - \alpha]$, with the optimal cutoff value α determined by the marginal distribution of the propensity score. The Crump et al. method is 254 not tied or limited to a specific estimator, has some optimality properties, and is straightforward to implement in practice. To implement the trimming procedure, one first estimates the propensity score; in the second step, one solves for the that satisfies smallest value where N denotes the total number of observations, and is the estimated propensity score for observation i . Crump et al. have shown that may be approximated to .1, and a rule of thumb for trimming is to discard all observations with estimated propensity scores outside the range $[0.1, 0.9]$. The Crump et al. (2009) method is based on a rigorous mathematic derivation and has the attractive feature of easy implementation. When conducting propensity score subclassification, using a conventional ad hoc approach, one would discard an entire stratum that only contains treated (or control) units without justification for achieving the semiparametric efficiency bound as that calculated by Hahn (1998). Using the Crump et al. procedure, one would calculate based on the marginal distribution of the estimated propensity scores, or simply choose = .1, to trim. The method solves the problem of limited overlap and makes subclassification feasible. Moreover, it has a clear theoretical justification. We highly recommend applying this method to address the limited overlap problem in propensity score subclassification. An example applying the Crump et al. procedure is shown in Section 6.5. We discuss the overlap assumption and the trimming strategy developed by Crump et al. (2009) in the setting of propensity score subclassification, because subclassification is often conducted in conjunction with trimming to help address limited overlap problems. The overlap assumption is embedded in all propensity score models assuming ignorability. The Crump et al. trimming method can be applied to many PSA models, such as those using matching or weighting procedures. 6.3 STRUCTURAL EQUATION MODELING WITH PROPENSITY SCORE SUBCLASSIFICATION This section switches to a slightly different topic: running SEM with propensity score analysis. Subclassification is not the only approach that permits an integrated analysis of propensity scores with SEM. The same integrated SEM model can be implemented by propensity score weighting (see Chapter 7). Because SEM has been widely used in the social and behavioral sciences and there is growing interest in the conjoint use of SEM and propensity score analysis, we present a general framework for developing an integrated model in this section. After discussing background information and general issues, we 255 focus on using subclassification to conduct SEM in observational studies. 6.3.1 The Need for Integrating SEM and Propensity Score Modeling Into One Analysis The past four decades have witnessed a rapid expansion of the use of SEM in research ranging from education to behavioral genetics to developmental psychology to sports medicine. During this time, SEM has evolved from a technique restricted to the highest echelon of statisticians (e.g., Jöreskog, 1971) to a practical tool used and valued by a broad scientific audience. SEM is a "class of methodologies that seeks to represent hypotheses about the means, variances, and covariances of observed data in terms of a smaller number of 'structural' parameters defined by a hypothesized underlying model" (Kaplan, 2000, p. 1). SEM enables researchers to study complex relationships, including those when some variables are latent. The underlying relationship, or structure, among variables is formulated as a model and is often articulated graphically through path diagrams. SEM permits multiple indicators of latent variables and the ability to estimate and test hypothesized relationships while controlling for random and systematic measurement error. Taking account of measurement error contrasts with the common practice of treating covariates, whether single variables or scales, as if they are free of error. Another valuable attribute of SEM is its capacity to compare hypothesized models to the empirical data, which leads to evaluating how well or to what extent the empirical data "fit" a theorized model. If the fit is acceptable, then the model is not rejected and the hypothesized relationships between variables are considered consistent with the data. Alternatively, a poor fit raises questions about the appropriateness of a model, even if problems are not evident with individual coefficients. Whereas analysis using SEM would have once required highly trained specialists, modern SEM software is widely available and easy to use. However, the increasing popularity of SEM and the availability of userfriendly SEM software warrant caution because SEM, like other procedures, can be abused. Although SEM applies to experimental as well as observational data, observational applications are more common. Regardless of the statistical procedure, use of observational data raises challenges in making causal inferences. Across many possible issues, we will focus on the selectivity problem in intervention studies using SEM. The literature on selectivity in latent variable SEM is sparse and underdeveloped (Bollen, 1989; Kaplan, 1999; B. O. Muthén & Jöreskog, 1983). Too often, a dummy treatment variable is specified as exogenous in evaluation models, but in fact it is not exogenous. In this context, the determinants of incidental truncation or sample selection must be explicitly modeled and selection effects considered when estimating causal impacts on outcomes (Heckman, 1978, 1979). The strongly ignorable treatment assignment 256 assumption is prone to violation in observational studies, and the presence of endogeneity leads to biased and inconsistent estimation of coefficients (Berk, 2004; Imbens, 2004; Rosenbaum & Rubin, 1983). Furthermore, typical covariance control does not automatically correct for nonignorable treatment assignment, as shown by the data simulation example of Section 3.5 in Chapter 3. Here we must draw a distinction between statistical modeling techniques and research design. Using an observational research design raises problems that occur due to selectivity and the failure to randomize confounding effects. Statistical modeling must take account of these problems to ensure valid inferences. If it does not, then biased estimates are likely. Randomized controlled trials tend to eliminate selectivity problems, but true experimental designs are not always possible, practical, ethical, or even desirable in the social and health sciences. Because evaluation continues to rely heavily on quasi-experimental and observational research designs, researchers have increasingly sought methods to control selection and confoundedness. Propensity score analysis concentrates on controlling for the factors that differentiate who receives treatment and who does not in an attempt to eliminate selectivity and its impact on the outcome. For instance, through propensity score matching or subclassification (Rosenbaum & Rubin, 1983), researchers create the groups or stratum similar in terms of observed covariates; through propensity score weighting (Rosenbaum, 1987), researchers correct for both systematic and chance imbalance. The focus on controlling selectivity in propensity score analysis is one of its great strengths. However, applications are differentially developed. In practice, propensity score analysis assumes that all variables have negligible measurement error. This is true for the variables that predict treatment assignment as well as for the variables that predict the outcome itself. Given variables such as academic performance, motivation, and educational aspirations, the implicit assumption of negligible measurement error is implausible and can contribute to biased assessments. In addition, propensity score models are often overidentified and assume a subset of variables influences treatment while having no direct effect on the outcome variable. These overidentification constraints are not tested, although they could be, and testing would lend greater plausibility to models. Finally, when propensity score methods lead to a comparison of groups, the comparison might be restricted to mean comparisons, whereas a much more comprehensive comparison of parameters across groups is possible and could be informative. Insufficient attention to latent variables, measurement error, testing overidentification, and comparing parameters across groups are the weaknesses of propensity score analysis, but attention to these is among the strengths of SEM. Inadequate consideration of selectivity is a weak point of latent variable SEM, but it is the key strength of propensity score analysis. It would seem 257 natural, therefore, to combine propensity score analysis and SEM methods to take advantage of their complementary strengths. Despite ongoing and increasing use of propensity score and SEM methods, few attempts have been made to combine propensity score and SEM into a single approach—with one notable exception: Kaplan's (1999) work. We describe next two examples that illustrate the need for an approach that integrates propensity score analysis and SEM. Example 1: Assessing Programs in Cluster-Randomized Trials—SACD Evaluation Presented in Chapter 1 (i.e., Example 6), this example illustrates the importance of conducting propensity score analysis, but here we present more detail to underscore the need for an integrated model using both propensity score modeling and SEM. Compared to descriptive designs, randomized controlled trials are seldom implemented in the social sciences, and random assignment in social experiments is prone to compromise. A case in point was the program evaluation of seven Social and Character Development (SACD) school-based programs. During 2003–2008, the Institute of Education Sciences (IES) and the Centers for Disease Control and Prevention (CDC) initiated a collaborative effort to conduct an impact evaluation of SACD programs (SACD Research Consortium, 2010). The SACD intervention project was designed to assess the effectiveness of schoolwide social and character development education programs. These programs were intended to help schools promote positive behaviors and reduce negative behaviors in elementary school students. IES used a peer-review process to select seven proposals to implement unique SACD programs in elementary schools across the country. The SACD intervention used a cluster randomization design. That is, at each school district of the seven sites, schools were randomly assigned to receive either an intervention program or routine services control curricula, and one cohort of students was followed from third grade (beginning in fall 2004) through fifth grade (ending in spring 2007). A total of 84 elementary schools were randomized to intervention and control at seven sites. Evaluating programs generated by a cluster randomization design is often challenging because the unit of analysis is a cluster (e.g., a school), and sample sizes are often so small—such as the number of schools within a school district—that randomization may fail to produce balance. To ensure rigorous program evaluation, the SACD Research Consortium (2010) identified 27 covariates as key control variables. Using the absolute standardized difference in covariate means (ASAM; Haviland et al., 2007; McCaffrey, et al. 2004) and statistical significance test using the Wilcoxon rank sum (Mann–Whitney) method, the authors of this book found great imbalances 258 on these 27 covariates. Note that ASAM indicates the absolute mean difference on a covariate between treated and control groups, where the mean difference is standardized at the sample standard deviation unit. An ASAM above .20 suggests that the between-groups difference on the covariate may be nonignorable, and the two groups may be imbalanced on that covariate. For instance, the SACD program site in North Carolina (i.e., Competence Support Program) had a between-group difference on the covariate "child race—Black" that yielded an ASAM of .46, which meant the treatment and control groups were almost half a standard deviation apart on Black race. The Wilcoxon rank sum test indicates that this level of difference is statistically significant ($p < .001$). All seven programs participating in the SACD intervention had imbalanced covariates. Of the 27 covariates, each of the seven programs had between two and five covariates with ASAM scores of .20 or greater, and the number of significant covariate differences (i.e., $p < .05$ shown by the Wilcoxon rank sum test) ranged from four to eight instances across the seven programs. The balance checks indicated that the cluster randomization was compromised, and the treated and control groups differed in significant ways. If these selection effects were ignored, then the evaluation findings would be biased. It is precisely at this intersection of design (i.e., failure of randomization) and data analysis that an integrated approach of propensity score modeling and SEM becomes critical. An SEM or other form of analysis that fails to control for these differences between the experimental and control groups risks confounding program effects with selectivity effects. In addition to a routine covariance analysis (e.g., regression or multilevel modeling or growth curve analysis), a promising alternative to data analysis would be propensity score analysis. However, none of the existing propensity score models permits users to address research questions that are best answered by SEM. For instance, SACD evaluators needed to address whether and in what ways the implementation of the SACD intervention generated differences between the treated and control students on key constructs, including student-reported domains of altruistic behavior, engagement with learning, and problem behavior. In this context, the best analytic method would be a revised confirmatory factor analysis (CFA) model (i.e., a submodel of SEM) to test factor invariance between groups, where the revision would incorporate a propensity score approach (e.g., propensity score weighting) and an SEM approach (i.e., CFA testing factor invariance between groups) into one model. As another example, the SACD evaluators often needed to test the mediating role of the intervention. That is, evaluators needed to test whether an exogenous variable could exert direct and indirect effects on an outcome variable and, if so, in what ways and how much of the total effect was due to the mediation of SACD intervention. Again, the existing propensity score models do not permit direct tests of mediating effects, and a SEM model used to test mediating effects without controlling for selection suffers from the selection bias problem. 259 Therefore, it is crucially important to integrate SEM and propensity score analysis into a single approach that will enable researchers to test for mediating effects and to simultaneously control for selection bias. Example 2: Assessing the Impact of Poverty on Academic Achievements The call for an integrated model also comes from research concerning causality. Consider again Example 2 of Chapter 1. A substantial body of research in the field concerning poverty and children's academic achievement implies causal relation and shows that both exposure to poverty and participation in welfare programs strongly influence child development. In general, growing up in poverty adversely affects a child's life prospects, and the sequelae of living in poverty increase in severity with greater exposure over time (Duncan et al., 1998; Foster & Furstenberg, 1999; P. K. Smith & Yeung, 1998). Most prior inquiries in this field applied a multivariate analysis (e.g., multiple regression or regression-type models) to samples of nationally representative data such as the Panel Study of Income Dynamics (PSID) data or administrative data, although a few studies used a correction method such as propensity score matching (e.g., Votruba-Drzal, 2006; Yoshikawa et al., 2003). Using a multivariate approach with this type of data poses a fundamental problem. First, the majority of the literature regarding the impact of poverty on children's academic achievement assumes a causal perspective (i.e., poverty is the cause of poor academic achievement), whereas a regression model or covariance control approach is not robust in handling endogeneity bias without explicit corrective procedures. Second, PSID is an observational survey without randomization; therefore, researchers must consider selection bias when using PSID data to assess causal effects. Sections 5.8.2 to 5.8.4, Section 7.3.1, and Sections 10.6.1 and 10.6.2 of this book present examples using optimal matching, propensity score weighting, and models of treatment dosages to establish causal linkages between poverty and achievements. Although these propensity score models help correct for selection bias and better answer questions concerning causality, none of the models allows researchers to answer the key question raised within the literature on the impact of poverty. Namely, what is the mediating role of the child use of welfare in the influence of the caregiver's use of welfare on child academic achievement? Answering this type of question is critical to research and policy decisions because, as suggested by the literature, if empirical data confirm that child use of welfare serves as a mediator, then interventions directly manipulating child use of welfare are worth trying to improve academic outcomes. Figure 6.1 shows an example of a mediating model in which child academic achievement is modeled as a latent variable with four achievement indicators (i.e., y_1 to y_4). As underscored by Kaplan (1999), conceptualizing achievement as a latent variable and then assessing the treatment effect on the 260 latent variable rather than on any individual indicator is an important strategy: "The focus on a factor rather than on any one of its indicators is because each indicator alone may be an unreliable measure of the corresponding construct. This lack of reliability in the outcome measure will result in biased standard errors of treatment effects" (p. 467). Currently, no established procedure allows researchers to test this conceptual model. Integrating propensity score analysis and SEM into one approach will enable researchers to test this model directly and efficiently. Figure 6.1 A Sample Conceptual Model Depicting the Mediating Role of a Child's Use of a Welfare Program (AFDC) in the Influence of Caregiver's Use of Welfare Program in Caregiver's Childhood on Child Academic Achievement The central idea of using propensity score subclassification to test the conceptual model would involve two steps. First, one would use propensity score subclassification to test similarities or divergences of the mediating role of the child use of welfare among groups with different propensities of using the welfare program. Second, one would evaluate, through a multigroup comparison among five propensity score subclasses, the total effect and direct effect of the 261 caregiver's use of welfare on child achievement and the indirect effect of the caregiver's use of welfare via the mediator of child use of welfare. 6.3.2 Kaplan's (1999) Work to Integrate Propensity Score Subclassification With SEM Currently, few studies integrate propensity scores with SEM. A notable exception is Kaplan's (1999) pioneering work extending the propensity score adjustment for the analysis of group differences in a MIMIC model. Kaplan held that integrating the propensity score and SEM into one model was important for two reasons. First, it is rarely the case in social behavioral research that any single measure of a construct is a reliable and valid measure of that construct. Therefore, researchers need to take advantage of SEM's latent-variable approach to incorporate multiple indicators into their analyses. Second, when the focus is on outcomes, measurement errors in outcome variables result in biased standard errors of treatment effects. Such bias in standard errors can be mitigated by incorporating multiple measures into a latent outcome, thus yielding accurate tests of treatment effects. Kaplan (1999) first used logistic regression to estimate propensity scores, and then he used quintiles (i.e., the 20th, 40th, 60th, and 80th percentiles) of propensity scores to divide the sample into five strata. Finally, Kaplan conducted subclassification analysis with SEM with a multigroup comparison (Jöreskog, 1971). The SEM that Kaplan used was a MIMIC model (i.e., multiple-indicators-multiple-causes model; Jöreskog & Goldberger, 1975). Using this model, Kaplan developed procedures to perform a multigroup comparison (i.e., compare the MIMIC model across five strata) to test factor invariance of the path coefficient of interest. The Kaplan model is an innovative development that combines SEM with propensity score subclassification. However, the model needs to be extended to a more general context. Specifically, Kaplan did not evaluate the magnitude of the overall treatment effect and whether such an effect was significant. Assessing the magnitude and significance of treatment effect is often the most important objective in program evaluation; therefore, it's important to develop formulas to aggregate the five path coefficients of interest, to aggregate the five standard errors associated with these coefficients, and to discern whether the overall treatment effect is significant. More generally, other propensity score approaches, such as propensity score weighting, may also be applied to the development of an integrated approach. In Chapter 7, we show how to apply propensity score weighting to conduct SEM. 6.3.3 Conduct SEM With Propensity Score Subclassification As depicted earlier, a special strength of SEM is that it allows researchers to evaluate the mediating role (i.e., the direct and indirect effects of an exogenous 262 variable on an endogenous variable via a mediator) in a latent-variable model. Under many research settings, latent-variable analysis is important to answering key research questions. We can use either propensity score subclassification or weighting to develop an integrated approach to test a mediation model. The analysis begins with the estimation of propensity scores for all study participants by using a logistic regression. The estimation should be based on important covariates that predict selection in the literature. After obtaining estimated propensity scores, researchers can use quantiles to divide the sample into strata and then follow Kaplan's (1999) procedure to compare key structural coefficients (i.e., those reflecting direct and indirect effects) among the strata. This is a routine process in conducting a multigroup SEM; essentially, the process involves testing a hierarchy of hypotheses that constrain model parameters in a systematic way. To estimate a treatment effect for the entire sample and test whether an overall treatment effect is statistically significant, researchers can employ Equations 6.3 and 6.4 to aggregate the path coefficients and their standard errors over all strata. Specifically, one first conducts propensity score subclassification to stratify the study sample into subclasses. Next, one performs balance check to ensure that covariate biases are sufficiently removed for most subclasses. One then runs SEM separately for each stratum. Finally, one aggregates treatment effects and standard errors over all strata by applying Equations 6.3 and 6.4. For a mediating model, researchers can obtain each structural path coefficient and its variance for the entire sample by applying Equations 6.3 and 6.4 and calculating the direct and indirect effects. The direct effect is the path coefficient from the exogenous variable to the endogenous variable (e.g., y_2 , sample in Figure 6.1), and the indirect effect is the product of two path coefficients (e.g., y_{11} , sample β_{21} , sample in Figure 6.1). The standard error of the indirect effect can be estimated by applying the following formula that uses the first-order Taylor series method (Krull & MacKinnon, 2001): or by using the bootstrapping methods for estimating variability of indirect effects proposed in Bollen and Stine (1990). 6.4 THE STRATIFICATION-MULTILEVEL METHOD In this section, we describe propensity score subclassification for modeling treatment effect heterogeneity. Developed by Xie et al. (2012), this is a stratification-multilevel (SM) method. The description of modeling effect heterogeneity serves two purposes. First, as shown in Section 2.8.1 in Chapter 2, modeling treatment effect heterogeneity is embedded in many substantive 263 research questions and designs of various types of observational studies. Researchers may find the latest development on this topic helpful. Second, the SM method provides a good example addressing the clustering issues presented by multilevel data. The current section is also an extension of Section 5.6 and shows the adjustment researchers may consider when analyzing multilevel data with propensity score subclassification. With regard to the first purpose, we show in Section 2.8.1 in Chapter 2 that the conventional use of interactions of the treatment indicator by covariates is not the best way to model treatment effect heterogeneity. Problems of using such interactions are depicted in that section. Xie et al. (2012) recommend focusing on the interaction of the treatment effect and the propensity score as one useful way to study effect heterogeneity. Although this is not the only way, it is a clearly interpretable way for exploring effect heterogeneity, because propensity scores summarize the relevance of

a full range of effects from the observed covariates. On the basis of this rationale, Xie and his colleagues (2012) developed three methods to model effect heterogeneity—namely, the SM method, the matching-smoothing (MS) method, and the smoothing-differencing (SD) method. All three methods basically employ the interaction of treatment by a propensity score. We focus on the SM method in this book. Readers who are interested in MS and SD methods are referred to Xie et al. (2012) for details. With regard to the second purpose, we discussed the need to control for clustering effects when applying propensity score matching to multilevel data in Section 5.6. We mentioned in that section that the multilevel modeling approaches for matching can also be extended to propensity score subclassification and weighting, but the extension may require adjustments. SM provides an example of such an adjustment. Recall that the purpose of subclassification is to create homogeneous subclasses such that the propensity score differences within a stratum can be ignored. As a consequence, study subjects within the stratum, after subclassification, tend to be homogeneous on the propensity scores. The process of subclassification, however, not only balances the data on covariates but also would make study subjects within a stratum alike on values of the outcome variable. Hence, there is a clustering effect caused by the subclassification. Just like matching, the clustering or autocorrelation on the outcome variable within the stratum is a design effect. Analysis should consider controlling for this effect. It is this consideration that leads Xie et al. (2012) to use a multilevel modeling. Note that in Subsection 5.6.4 in Chapter 5, we discussed an approach of CCREM to correct for two sources of clustering following matching, because in that context, the grouping from the two sources is not nesting but cross-classifying. The SM method, however, considers only the clustering effect due to subclassification and assumes an absence of multilevel data in the original sample. Formally, there are four steps in conducting an SM analysis to model treatment effect heterogeneity. The first two steps aim to stratify the study 264 sample using propensity scores. They are similar to those described in this chapter. The four steps are shown as follows: 1. Estimate propensity scores for all units using logistic regression or a probit model. 2. Construct propensity score strata (or ranges of the propensity score) where there are no significant differences in the average values of covariates and the propensity score between the treatment and control groups. In other words, test balance within each stratum. 3. Estimate propensity score stratum-specific treatment effects within strata. This requires a Level 1 model of the treatment effects. Let Y_{ij} be the outcome variable, d_{ij} the treatment indicator variable ($d_{ij} = 1$, treated; $d_{ij} = 0$, nontreated), and p_{ij} the propensity score, for the i th observation in the j th stratum, where $i = 1, 2, \dots, N$, and $j = 1, 2, \dots, J$. If we have five strata, $J = 5$. The Level 1 model can be expressed as follows: where δ_j , β_j are the treatment effect and regression coefficient associated with the propensity score within stratum j , respectively, and ϵ_{ij} is the residual term for the i th observation within stratum j . 4. Evaluate a trend across the strata using variance-weighted least squares regression of the strata-specific treatment effects, obtained in Step 3, on strata rank at Level 2. Xie et al. (2012) emphasize that the main research objective here is to look for a systematic pattern of heterogeneous treatment effects across strata. In the interest of simplicity and preserving statistical power, the authors suggest modeling the heterogeneity pattern mainly as a linear function across strata ranks. For the sake of controlling for clustering effect, as explained earlier, the authors suggest running a Level 2 model as follows: where the Level 1 slopes δ_j are regressed on propensity score rank indexed by j , ϕ represents the Level 2 slope, and η_j is the random effect for the j th stratum. The model assumes normality of η_j . Note that ϕ is the key coefficient of this model: It is a cross-level interaction, or an interaction of treatment indicator by propensity score rank—an innovative approach the authors proposed for modeling treatment effect heterogeneity. ϕ can be interpreted as the change in the treatment effect with each one-unit change to a higher propensity score stratum; significance of this coefficient indicates 265 the existence of effects heterogeneity, that is, treatment effect varies by the level (i.e., the rank) of propensity score stratum. The Stata program `hte` (Jann et al., 2010) is developed to estimate the SM model. To install the program, the analyst simply types "`ssc install hte`" in Stata. Users need also install the Stata program `pscore` to run `hte`. An empirical example studying the effects of college attendance on women's fertility can be found in Xie et al. (2012, pp. 326–334). Results show that the effects of attending college on women's number of children are not homogeneous and vary by the propensity of attending a college. Additional examples of modeling treatment effect heterogeneity using this type of methods can be found from Xie and Wu (2005) and Brand and Xie (2010). 6.5 EXAMPLES 6.5.1 Stratification After Greedy Matching To illustrate the calculation of the ATE of the sample and its significance test based on Equations 6.1 and 6.2 following a greedy matching (i.e., a bivariate analysis of outcome differences across treatment conditions, or performing analysis of Step 3b shown by Figure 5.1), we use an example provided by Perkins et al. (2000). On the basis of propensity score stratification, Perkins et al. reported means and standard errors of an outcome variable as in Table 6.1. Table 6.1 Estimating Overall Treatment Effect After Stratification (Example 6.5.1) Source: Perkins, Tu, Underhill, Zhou, and Murray (2000, Table 2). Reprinted by permission of John Wiley & Sons, Ltd. Applying Equation 6.1 to these data, the sample ATE is 266 Applying Equation 6.2 to these data, the variance and standard error of the ATE are Because the mean difference between treatment groups for the whole sample (i.e., the average sample treatment effect) is not statistically significant at the level of $\alpha = .05$. Note that in the above calculation, indicates a variance of a difference between two mean scores; such a variance, as a matter of fact, is simply the sum of the variance for the control group and the variance for the treatment group. 6.5.2 Subclassification Followed by a Cox Proportional Hazards Model This example illustrates the propensity score subclassification followed by a multivariate outcome analysis. The outcome analysis employs a Cox proportional hazards model. The example used the same data and research question as those presented in Section 5.8.1, except that the outcome analysis replaced the Kaplan-Meier product limit analysis with Cox regression. Included in this study are 2,758 children from the panel data of the National Survey of Child and Adolescent Well-Being (NSCAW). After a listwise deletion of missing data, the study sample comprises 2,723 children. Recall that the study is interested in whether caregivers' use of substance abuse services is predictive of the timing of a maltreatment rereport 18 months after the baseline interview. Study participants who did not have a rereport at the end of the 18-month window were censored. Of the 2,723 children included in the study, 294 (10.8%) had female caregivers who received substance abuse treatment, and the remaining 2,429 (89.2%) had female caregivers who did not receive treatment services. Table 6.2 presents sample descriptive statistics and the coefficients of the logistic regression predicting the propensity scores. Bivariate chi-square tests indicate that the two groups are not balanced on most covariates, as many 267 covariates are statistically significant ($p < .05$). Using the normalized difference (Imbens & Wooldridge, 2009), the following covariates have a difference that exceeds .25, indicating that selection bias exists: case status open, child age group of 0 to 2 years old, caregiver mental health problem, caregiver arrested, caregiver received alcohol or drug (AOD) treatment, presence of risk, presence of the Composite International Diagnostic Interview—Short Form (CIDI-SF), and caregiver report of need. The logistic regression employed the third model of Section 5.8.1, and covariates that are statistically significant at the level of .01 are the same covariates whose normalized difference exceeds .25. Sorting the study participants by estimated propensity scores in an ascending order and using quintiles, the study divides the sample into five subclasses. The number of observations in each subclass, from Subclass 1 to Subclass 5, is 545, 545, 543, 546, and 544, respectively. As we have indicated, it is important to check balance on observed covariates between the two groups before the study can move to the next step. All covariates presented in Table 6.2 were then checked by using the same chi-square tests, but this time the bivariate tests were performed within subclasses. The results were not positive. As is often the case, the two groups did not balance on certain covariates, particularly at the two ends of the five subclasses. Note that all participants in Subclass 1 had low propensity scores and that all participants in Subclass 5 had high scores. As such, some covariates show statistical significant differences between the two treatment groups—that is, they were not balanced between treatment conditions; alarmingly, in the highest and lowest subclasses, some participants fell uniquely into one category of certain covariates. The results of the balance check are summarized as follows, and the summary reports significant differences only: In Subclass 1, all participants on "risk assessment" fall into the category of "absence," and on "caregiver report of need," they fall into the category of "no need"; in Subclass 2, all participants on "risk assessment" fall into the category of "absence," and the chi-square test on "caregiver report of need" is statistically significant ($p < .000$); in Subclass 3, the chi-square test on "AOD treatment receipt" is statistically significant ($p < .01$); in Subclass 4, the chi-square test on "caregiver report of need" is statistically significant ($p < .05$); and in Subclass 5, the chi-square tests on the following covariates are statistically significant: "case status" ($p < .05$), "caregiver mental health problem" ($p < .05$), "risk assessment" ($p < .001$), "CIDI-SF" ($p < .01$), and "caregiver report of need" ($p < .05$). Table 6.2 Sample Description and Logistic Regression Predicting Propensity Scores (Example 6.5.2) 268 269 Source: Data from NSCAW, 2004. Note: Reference group is shown next to the variable name. AOD = alcohol or drug; CIDI-SF = Composite International Diagnostic Interview—Short Form. * $p < .05$, ** $p < .01$, *** $p < .001$, two-tailed test. Given this amount of imbalance, the study cannot proceed using five subclasses. The study then investigates the possibility of conducting a 10-group subclassification and inquires whether a subclassification with more groups will balance the data in an improved fashion. By design, the 10-subclass scheme would encounter the same problem as that for the 5-subclass scheme—that is, at 270 the lower and/or higher regions of propensity scores, participants may uniquely fall into one category on some covariates. However, if the imbalance occurs only in, say, the first three subclasses and the last two subclasses, then a new subclassification combining the first three subclasses into one and the last two subclasses into one—that is, a new scheme of seven-subclass grouping—might sufficiently balance the data. Using deciles (i.e., every one increment of the 10th percentile), the study divides the sample into 10 groups with the following number of participants for each subclass (the following Ns are for Subclass 1 to Subclass 10, respectively): 272, 273, 270, 275, 272, 271, 274, 272, 271, and 273. Results of the balance check for each of the 10 subclasses are summarized as follows, and the summary reports significant differences only: In Subclass 1, all caregivers did not use substance abuse treatment, and all participants on "risk assessment" fall into the category of "absence" and on "caregiver report of need" fall into the category of "no need"; in Subclass 2, all participants on "risk assessment" fall into the category of "absence" and on "caregiver report of need" fall into the category of "no need"; in Subclass 3, all participants on "risk assessment" fall into the category of "absence" and on "caregiver report of need" fall into the category of "no need"; in Subclass 4, all participants on "risk assessment" fall into the category of "absence," and the chi-square test on "caregiver report of need" is statistically significant ($p < .001$); in Subclass 5, all participants on "risk assessment" fall into the category of "absence," and the chi-square test on "caregiver age" is statistically significant ($p < .05$); in Subclass 6, the chi-square tests on "AOD treatment receipt" and on "caregiver report of need" are statistically significant ($p < .01$ and $p < .05$, respectively); in Subclass 7, the chi-square tests on "caregiver marital status," "poverty," and "caregiver report of need" are statistically significant ($p < .05$ for all three tests); and in Subclasses 8 to 10, all covariates are balanced and no significant results are detected. Because all first seven subclasses show imbalance one way or another, the new grouping using 10 subclasses does not balance data either. The findings suggest that the data set violates the overlap assumption. To address the problem of limited overlap in this data set, the study follows Crump et al. (2009) and pursues a trimming strategy. Specifically, the study searches for a value of that is between 0 and 1/2, and the value is the smallest among all possible conditions that make Equation 6.5 hold. Doing so, the study finds the value for this data set to be .079. Then the study trims/discards participants whose estimated propensity scores are below .079 and above .921 (i.e., $1 - .079 = .921$) from the analysis. The trimming reduces the sample size from 2,723 to 745, or a reduction of 72.6% of the original sample. The sample reduction through trimming is necessary to address the limited overlap problem, but the loss of sample participants is huge. It clearly reduces the study's external validity. This is a limitation of trimming. However, the number of observations lost through this process is smaller than 271 that through greedy matching where a loss of 82.3% is observed (see Section 5.8.1). Using the newly trimmed sample ($N = 745$) and quintiles, the study creates five new subclasses. Results of the balance check, shown in Table 6.3, are good. Only three covariates remain statistically significant (i.e., "child age" in Subclass 1, "poverty" in Subclass 3, and "CIDI-SF" in Subclass 5), and all other covariates are not statistically significant. Note that the significance of "child age" may be interpreted as chance significance because it was never significant in all previous schemes. Using this new sample of 745 participants, the study conducts the outcome analysis by subclass, that is, it runs the Cox proportional hazards model for each subclass. The Cox regression controls for demographics of study children, because they are important covariates of the outcome shown by prior studies (Guo et al., 2006). The three dichotomous variables of "caregiver age" were originally included in the study but excluded from the final model due to an empty-cell problem caused by them. Table 6.4 presents results of the outcome analysis. Extracting the coefficients of receipt of substance abuse treatment and the corresponding standard errors from all five subclasses, the study then computes the overall effect by aggregating the five estimated coefficients and computes the associated standard error by aggregating the five estimated standard errors. The purpose of this analysis is to discern whether the treatment effect across all five subclasses is statistically significant. Applying Equations 6.3 and 6.4, the study finds the sample average treatment effect, its standard error, and the p value of the significance test as follows: Table 6.3 Balance Check After Trimming and Subclassification Using Quintiles (Example 6.5.2) 272 273 Source: Data from NSCAW, 2004. 274 Table 6.4 Estimated Cox Regression Models (Estimated Coefficients) by Stratum (Example 6.5.2) Source: Data from NSCAW, 2004. Note: Reference group is shown next to the variable name. * $p < .1$, ** $p < .05$, *** $p < .01$, **** $p < .001$, two-tailed test. 275 Literally, the study finds that by controlling for selection bias, children whose caregivers used substance abuse treatment are more likely to receive a maltreatment rereport, and the hazard of having a rereport for these children is 45.52% higher than that of children whose caregivers did not use substance abuse treatment, although this difference is not statistically significant at the .05 level. Compared to results of the greedy matching of Section 5.8.1 in Chapter 5, the study using subclassification and Cox regression confirms that the hazard rate of having a rereport for caregivers who used substance abuse treatment is higher, but the finding is no longer statistically significant. The Cox regression without controlling for selection bias (i.e., results from the column of "Overall" in Table 6.4) shows that the hazard ratio of treatment is 1.547 (i.e., $\exp(0.436) = 1.547$), close to that estimated by the subclassified Cox regression, but the finding is statistically significant at the .001 level. The statistical significance of the treatment effect from the "overall" model appears influenced by uncontrolled selectivity. 6.5.3 Propensity Score Subclassification in Conjunction With SEM This example employs the 1997 Child Development Supplement (CDS) survey and 1968–1997 PSID data to test the conceptual model presented in Figure 6.1. After deleting cases with missing values on any of the study variables, the study sample comprises 601 children between ages 5 and 12 years in 1997. The data include age-normed standardized scores on four academic achievement variables: letter-word identification, passage comprehension, calculation, and applied problems. Table 6.5 presents sample descriptive statistics and results of the imbalance check. Of five covariates, four are statistically significant in terms of the mean values between the treated (i.e., child who used Aid to Families With Dependent Children [AFDC]) and the comparison (i.e., child never used AFDC) groups, shown by either the Wilcoxon rank sum test or the independent samples t test ($p < .000$). All four covariates have normalized difference values that exceed .25. These results show that the groups differ on four covariates. The two groups lack balance. The results show, in general, that children who ever used AFDC were more 276 likely to be African American and were more likely to have caregivers who had lower income, who had lower levels of education, and who used AFDC over a longer time during their childhood than the caregivers of children who never used AFDC. Therefore, a correction for selection bias is warranted. Table 6.5 Sample Description and Imbalance Check for the Study of the Poverty Impact (Example 6.5.3) Source: Data from Hofferth et al., 2001. Note: AFDC = Aid to Families With Dependent Children. To begin the correction process, a propensity score subclassification is performed on the sample data. The propensity scores were estimated by a binary logistic regression. Using quintiles, the study divides the sample into five strata (i.e., $K = 5$ for $N = 601$) of slightly different sizes (i.e., $n_1 = 121$, $n_2 = 119$, $n_3 = 120$, $n_4 = 121$, $n_5 = 120$). Balance checks on the same set of covariates for each stratum were performed. Results show that except for the last stratum, all covariates are balanced between the treated and untreated groups. For the purpose of illustration, the study did not pursue further data balancing. Using the five strata, we illustrate the basic idea of conducting SEM with multiple-group comparison. Based on exploratory runs, the final SEM specified correlational measurement errors between letter-word identification and passage comprehension, between calculation and applied problems, and between passage comprehension and calculation. The mediating model was tested using a data file that stacks all five strata. In essence, the testing is a group comparison of model structures among the propensity score subclasses. Specifically, the study ran four consecutive models; each of the models had a slightly different constraint on path coefficients. The four models are as follows: 277 Model A—baseline or same-form model that specifies same form but sets all coefficients free among the five subclasses, Model B—same γ_{11} and across the five subclasses, Model C—same β_{21} across the five subclasses, and Model D—same γ_{11} , γ_{21} , and β_{21} across the five subclasses. A series of chi-square difference tests were then implemented to evaluate whether each of the constrained models (i.e., Models B, C, and D) was better than the baseline Model A. Results of these tests are shown in Table 6.6. Given that the study cannot accept Model B of "same γ_{11} and " because the p value from the chi-square difference test is significant, it does not test Model D of "same γ_{11} , γ_{21} , and β_{21} ." Therefore, results from Model C are taken as the final estimates. Table 6.6 Group Comparison in SEM With Propensity Score Subclassification (Example 6.5.3) Source: Data from Hofferth et al., 2001. The fit indices of Model C are as follows: Model $\chi^2(df = 53) = 57.916$, $p = .299$, and root mean square error of approximation (RMSEA) = .03 with a 90% confidence interval of $(.00, .07)$. Applying Equations 6.3, 6.4, and 6.6, the study obtains the following estimates: The direct effect of child use on achievement the direct effect of caregiver's use of AFDC during childhood on child achievement which is not statistically significant; and the indirect effect of caregiver's use of AFDC on child achievement via child use of AFDC These results confirm the mediation model. That is, other things being equal, children who ever used AFDC had an academic achievement score that was 3.375 units lower than those who never used AFDC ($p < .000$); the direct effect of caregiver use of AFDC on child achievement is -1.094 (not significant), and the indirect effect of caregiver use of AFDC on child achievement via the mediator of child use of AFDC is -0.945 ($p < .000$). Of the total effect of caregiver use AFDC on child achievement, 92.0% is direct, and 8.0% is indirect. 6.6 CONCLUSION 278 Propensity score subclassification stems from the statistical tradition of exact subclassification on one or more covariates. As a one-dimensional balancing score, propensity scores provide important efficiencies in subclassification on multiple covariates. Using quintiles of estimated propensity scores, researchers can stratify the study sample into five subclasses. A mean comparison of outcomes between treated and control groups or a multivariate outcome analysis then is conducted within each stratum. The within-stratum treatment effects or model-estimated coefficients are aggregated to form an estimate for the entire sample, and researchers can use the sample statistic in conjunction with an aggregated standard error to perform a hypothesis test to discern whether the overall sample average treatment effect or sample coefficient is statistically significant. Using five subclasses is common and usually removes 90% of bias on observed covariates (Cochran, 1968), although researchers can use more subclasses provided that the sample size is sufficiently large. A postsubclassification balance check should be routinely performed to ensure that subclassification removes selection bias. To address limited overlap of covariates, researchers can discard observations with propensity scores outside the range of $[0.1, 0.9]$ before subclassification or discard observations with propensity scores outside the interval of $[\alpha, 1 - \alpha]$, where α is a robust and efficient cutoff value determined by the formula developed by Crump et al. (2009). Trimming will improve comparability (balance), but it will inevitably result in the loss of cases. This can affect external validity. Using propensity score subclassification, researchers can also perform structural equation modeling to test complex relationships between variables, including the tests of mediating, moderating, and nonrecursive effects. Finally, the subclassification method can be employed with multilevel modeling, such as the stratification/multilevel method developed by Xie et al. (2012). Subclassification on propensity scores is a flexible and efficient means of dealing with selection bias. 279 CHAPTER 7 Propensity Score Weighting Propensity scores may be used without matching or subclassification in a weighting process that reduces the analysis to two steps. This section describes multivariate analysis using propensity scores as sampling weights (i.e., proceeding to Step 2b, shown in Figure 5.1 in Chapter 5). The fundamental characteristic of this method is the use of the inverse probability of treatment assignment as a weight in a multivariate outcome analysis. Unlike making control participants similar to treated participants on propensity scores (i.e., matching) or creating subclasses such that, within a class or stratum, both treated and control participants are homogeneous on propensity scores (i.e., subclassification), propensity score weighting takes a differential amount of information from each participant depending on the participant's conditional probability of receiving treatment. In doing so, the method accomplishes the same goal of balancing data. It makes the estimate of the sample average treatment effect or its inference to the population average treatment effect a weighted average of the difference between observed and potential outcomes. The method is analogous to the weighted analysis researchers typically conduct when analyzing data generated by a complex sampling design, although in the current practice, the goal is to enhance internal validity rather than external validity. There are two advantages of propensity score weighting: It permits most types of multivariate outcome analyses and does not require an outcome variable that is continuous or normally distributed; more attractive, the method usually permits retaining most study participants in the outcome analysis. Weighting does not sacrifice observations as greedy matching or trimming does. Because of this, the method has been widely employed in the social and health sciences. Section 7.1 provides an overview of the weighting method. It shows the statistical principles of weighting and why an approach using the inverse probability of treatment weights has the capacity to correct selection bias. Section 7.2 offers practical guidance for conducting propensity score weighting with empirical data. It describes two types of weighted analyses: One is for estimating the average treatment effect (ATE), and the other is for estimating the average treatment effect for the treated (ATT). Section 7.3 presents three applications: a weighted regression analysis of a continuous outcome, a 280 weighted Cox regression that analyzes time-to-event data where the outcome variable is right and random censored, and a weighted structural equation model that aims to test mediating effects and to partition into direct and indirect effects the total effect of a key exogenous variable on an endogenous variable. Section 7.4 concludes. 7.1 OVERVIEW Propensity score weighting assumes unconfoundedness—that is, using the method, researchers assume that the treatment assignment is independent of the outcomes in both treated and control groups, conditional on the observed covariates, and the probability of assignment is bounded away from 0 and 1. Because of this assumption, weighting can be considered a submodel of those developed by Rosenbaum and Rubin (1983). Rosenbaum (1987) sketches the basic ideas of weighting. Hirano and Imbens (2001) and Hirano, Imbens, and Ridder (2003) describe methodological principles and provide practical guidance. McCaffrey et al. (2004) provide an excellent application that illustrates basic steps in conducting propensity score weighting. They demonstrate strategies for addressing challenges encountered in weighting analyses. According to Imbens and Wooldridge (2009), most propensity score models can be viewed as a weighted analysis of outcomes, because we can generally write these models as different weighting processes that estimate weighted outcome differences between the treated and control groups, with the weights in both groups adding up to one, as Various propensity score models "differ in the way the weights λ_i depend on the full vector of assignments and matrix of covariates (including those of other units). For example, some estimators implicitly allow the weights to be negative for the treated units and positive for control units, whereas others do not. In addition, some depend on essentially all other units whereas others depend only on units with similar covariates values" (Imbens & Wooldridge, 2009, p. 25). The conceptualization of propensity score analysis as an analysis of weighted outcomes provides a pedagogical convenience for understanding the methodological principles behind various propensity score models. Indeed, all corrective approaches discussed in this book, including greedy matching, optimal matching, subclassification, weighting, matching estimators, and kernelbased matching, can be viewed as different approaches to generating λ_i . That being said, it is important to treat the propensity score weighting estimator discussed in this chapter as a special case, a method that is categorically 281 different from other propensity score models. The method directly exploits the inverse of estimated propensity scores as weights in an outcome analysis, and to a large extent, it shares similarities with weighted analyses using unequal sampling weights. Following Imbens and Wooldridge (2009), we now show why an inverse propensity score for observation i can be treated as a weight in estimating population average treatment effects. Using the counterfactual framework depicted in Chapter 2, we can write the population average treatment effect as consider the term $E(Y_{i1})$ first. Because this is the expectation of outcome Y_i for the treated participants, $W_i Y_i = 1 \cdot Y_{i1}$, we have The crucial message conveyed by the above derivation is or for the treated participants, the expectation of outcome $E(Y_{i1})$ equals the expectation of the outcome Y_{i1} multiplied by the observation's inverse probability of . Hence, the derivation receiving treatment or inverse propensity score as a weight for the treated participants (i.e., those shows that we can use whose $W_i = 1$) when estimating ATE. Now consider the second term in calculation as that for $E(Y_{i0})$, we can show that Using a similar This suggests that for the control participants, the expectation of outcome $E(Y_{i0})$ equals the expectation of the outcome Y_{i0} multiplied by the observation's inverse of one minus the probability of receiving treatment, or multiples by . Hence, the as a weight for the control participants derivation shows that we can use (i.e., those whose $W_i = 0$) when estimating ATE. Together, the above derivations show that Therefore, the population ATE can be estimated by a weighting algorithm. The weighted version of tPATE is To express the weights more explicitly, we can formally write the estimator using the inverse probability of treatment weights for ATE as 282 Imbens and Wooldridge (2009) show that the estimator 7.1 is equivalent to a sample average from a random sample and is consistent for tPATE and asymptotically normally distributed. The estimator 7.1 is essentially due to Horvitz and Thompson (1952). Because the Horvitz-Thompson estimator is based on sample averages and is widely employed in weighted analysis to adjust for unequal probabilities employed in stratified sampling, applying the adjustments to propensity score weighting is straightforward. In essence, propensity score weighting is analogous to a weighted analysis that treats propensity score weights as sampling weights. Because analyzing data with a complex sampling design is a routine procedure today and functions of weighted analysis are offered by popular software packages, propensity score weighting is easy to implement, in the sense that the method does not require additional programming efforts. As shown by the examples in Section 7.3, after researchers create propensity score weights for treated and control participants separately, they can perform a weighted outcome analysis—whatever the multivariate outcome model is—simply by specifying the weight variable and calling for a weighting function in the outcome analysis. The advantages of easy implementation, accommodating most types of outcome analyses and retaining most cases from the original sample, are very attractive and are the primary reasons why the method has become popular among researchers from a broad range of disciplines. A few methodological notes are worth acknowledging: 1. Because propensity score weighting has been developed recently, a variety of terms are used in the field. The estimator 7.1 is generally referred to as the inverse probability of treatment weights (IPTW) estimator, and it is synonymous with propensity score weighting for estimating ATE shown by Section 7.2, and the latter name (weighting estimator for ATE) is used, for instance, by Hirano et al. (2003) and McCaffrey et al. (2004). An alternative weighting process that aims to estimate ATT is referred to as weighting by odds (Harder et al., 2010). 2. Users can employ either binary logistic regression or generalized boosted modeling (GBM; see Section 5.3.4) to estimate sample propensity scores. However, regardless of which method is used, the model-predicted probability rather than the logit score should be used to define propensity scores. McCaffrey et al. (2004) used GBM to estimate the sample propensity scores. 283 3. As a routine process, users should always check balance on observed covariates to ensure that weighting helps correct for selection. Recall that propensity score weighting does not use matching and trimming, and therefore, the analysis sample will be the same as the original one. Given this, the method of checking covariate imbalance for optimal matching (see Section 5.5.2) is not applicable, and the bivariate tests and/or normalized differences cannot be readily performed either. Users need to consider employing a different approach to check balance, one that is suitable to the weighted analysis. The approach we recommend is a weighted simple regression or weighted simple logistic regression. Specifically, one runs a weighted regression using a continuous covariate as the dependent variable and a dichotomous treatment variable as the single independent variable. If the covariate being tested is a dichotomous variable, then one runs a weighted logistic regression using the dichotomous covariate as the dependent variable, and again using the dichotomous treatment variable as the single independent variable. If propensity score weighting effectively removes imbalances, then one would hope that the regression (or logistic regression) coefficients from these models are not statistically significant. As usual, if many covariates from the weighted regression analyses remain significant, one then needs to rerun the propensity score logistic regression to generate a new set of estimated propensity scores. 4. Running propensity score weighting, users employ the same procedure as if they were running an outcome analysis with unequal sampling weights. As such, researchers often face the challenge of incorporating sampling weights into a weighted propensity score analysis. In theory, the two types of weights—propensity score weights and sampling weights—are probability-typed quantities, and as such, it's not invalid to incorporate the two types of weights into one by multiplication. Studies using data simulation have shown promising results. For instance, using simulation, DuGoff, Schuler, and Stuart (2014) compared four methods for estimating the treatment effect: a naive estimate ignoring both survey weights and propensity scores, survey weighting alone, propensity score methods (nearest neighbor matching, weighting, and subclassification) alone, and propensity score methods in combination with survey weighting. The results of the simulation led the authors to conclude that combining a propensity score method and survey weighting is necessary to achieve unbiased treatment effect estimates that are generalizable to the original survey target population. Although studies on this topic are emerging, it is worth noting that this again is a rapidly growing area. Social and health sciences researchers should keep their eyes on new developments. Particularly, given that publications on this topic are so new and scarce, we believe that theoretical work supporting the use of combined weights needs further development. 5. As depicted earlier, propensity score weighting can be applied to most 284 types of outcome analyses, and most software packages offer functions to permit a weighted multivariate analysis. In Stata, for instance, most multivariate models can be analyzed by a user-specified weight variable in conjunction with the call of `pweight`. This type of weighted model includes multiple regression, logistic regression, ordered logistic regression, multinomial logit model, Poisson regression, Cox regression, multilevel modeling, and more, although sometimes users need to use the survey module offered by the computing package (e.g., "`svy`," offered by Stata) to implement an analysis. Most software packages for running structural equation modeling (SEM) offer functions for weighted analysis too, and users need to specify the name of the weight variable in conjunction with the call of the weight function. For instance, in Mplus, one needs to specify "weight is varname" to call a weighted SEM, where varname is the name of weight variable. 6. Although weighting with propensity scores is a creative idea and easy to implement, recent studies suggest that it may have limitations. Focusing on the weighting procedure to estimate ATE, Freedman and Berk (2008) conducted a series of data simulations. They found that propensity score weighting was optimal only under three circumstances: (1) Study participants were independent and identically distributed, (2) selection was exogenous, and (3) the selection equation was properly specified (i.e., with correct predictor variables and functional forms). When these conditions were not observed, they found that weighting was likely to increase random error in the estimates. Indeed, weighting appears to bias the estimated standard errors downward, even when selection mechanisms are well understood. Moreover, in some cases, weighting may increase the bias in estimated causal parameters. Given these findings, Freedman and Berk recommend that if investigators have a good causal model, it may be better to fit the model without weights; if the causal model is improperly specified, there can be nontrivial problems in rectifying the situation by weighting, and in such a circumstance, investigators should exercise caution in implementing the weighting procedure. Freedman and Berk warn that it rarely makes sense to use the same set of covariates in the outcome equation and the selection equation that predicts propensity scores. Reflecting on both the developmental nature of the field and the uncertainty surrounding the validity of emerging weighting procedures, Kang and Schafer (2007) showed that the use of inverse probabilities as weights is sensitive to misspecification of the propensity score model when some estimated propensities are small. Caution seems warranted in the use of propensity score weighting in these circumstances. 7.2 WEIGHTING ESTIMATORS In this section, we provide guidelines for using propensity score weighting 285 estimators. First, we show the formulas for estimating two types of weights: One is the set of weights estimating ATE, and the other is the set of weights estimating ATT. Next, we show a variant of the weighting estimator for evaluating ATE when it is necessary to correct for large and influential weights. Finally, we summarize and synthesize the analytic procedure into three steps. 7.2.1 Formulas for Creating Weights to Estimate ATE and ATT 1. For estimating ATE, we define weights as follows: By this definition, when $W = 1$ (i.e., a treated participant), Equation 7.2 becomes when $W = 0$ (i.e., a control), Equation 7.2 becomes This is the same estimator as Equation 7.1 but is expressed by simplified notations. The method is also known as the inverse probability of treatment weights (IPTW) estimator. Propensity scores are estimated by using the sample data, or the observed covariate vector x , and are modelpredicted probabilities of receiving treatment from a binary logistic regression, or GBM, or other methods (e.g., from a probit model). 2. For estimating the ATT, we define weights as follows: By this definition, when $W = 1$ (i.e., a treated participant), Equation 7.3 becomes $\omega(W, x) = 1$; when $W = 0$ (i.e., a control), Equation 7.3 becomes . The method is also known as weighting by odds. In summary, if we denote simply as P , the weight is $1/P$ for a treated participant and $[1/(1 - P)]$ for a control participant when estimating ATE; the weight is 1 for a treated participant and $[P/(1 - P)]$ for a control participant when estimating ATT. After creating these weights, we can simply use them in multivariate analysis. Most software packages allow users to specify the name of a weight variable in procedures of multivariate analysis. The analysis then is analogous to multivariate modeling that incorporates sampling weights. 7.2.2 A Corrected Version of Weights Estimating ATE In practice, researchers sometimes have weights that are very large and thus influential, possibly resulting in biased (or at least very imprecise) estimates of the treatment effect (Robins et al., 2000; Schafer & Kang, 2008). When this happens, it is not recommended that treated individuals with large weights simply be removed because those individuals are generally the best predictors 286 of the outcome under comparison given that a large IPTW weight results from a small propensity score (Harder et al., 2010). To handle influential weights, Robins (1998, 1999b) and Robins et al. (2000) developed a procedure called stabilized stabilization, which multiplies a constant term in Equation 7.2 for the treated participants and the control separately. The constant term equals the expected value of being in the treatment or control groups, respectively. Using stabilization, the weight for a treated participant i ($i = 1, 2, \dots, n_1$) when estimating ATE becomes And the weight for a control participant j ($j = 1, 2, \dots, n_0$) when estimating ATE becomes According to the developers (Robins, 1998, 1999b; Robins et al., 2000), because each of the weight formulas multiplies the original weight for treated and control participants by a constant, respectively, stabilization executed in this way does not affect the point estimate of the treatment effect. However, it does decrease the variance. When the propensity score model is saturated (i.e., the model includes all covariates and possible product terms), the variance of the final outcome analysis will be the same with and without stabilization. Harder et al. (2010) add that this correction does affect the final results when the propensity score model is unsaturated. In their analysis, when they used an unsaturated propensity model, stabilization reduced the variability of the IPTW weights and hence reduced the variance of their ATE estimates. 7.2.3 Steps in Propensity Score Weighting In summary, propensity score weighting consists of

the following three steps: 1. Estimate propensity scores using simple observed covariates x in a logistic regression or similar model. Note that the propensity score may or may not be saturated (i.e., it contains or does not contain the same set of covariates used in the final outcome analysis). 2. Calculate two types of weights: the weight for estimating ATE (i.e., by applying for Equation 7.2) and the weight for estimating ATT (i.e., by applying for Equation 7.3). Note that each type of weight has two versions, depending on whether the participant receives treatment or not. If the study sample contains large and influential weights, users may further correct 287 ATE estimates by multiplying by a constant term, that is, by applying the Equations 7.4a and 7.4b. 3. Specify the weight in an outcome analysis that treats the weight just like a sampling weight. The outcome analysis becomes a propensity score-weighted analysis. 7.3 EXAMPLES To illustrate propensity score weighting, this section presents three examples that demonstrate estimating ATE and ATT in different multivariate outcome analyses—namely, multiple regression, Cox regression, and SEM testing mediational effects. 7.3.1 Propensity Score Weighting With a Multiple Regression Outcome Analysis In this example, we illustrate the analysis of propensity score weighting with a multiple regression outcome analysis. The example employs the same data and research questions as the example using optimal matching (see Section 5.8.2 in Chapter 5). We first employed GBM to estimate the propensity scores for all participants. Next, applying Equations 7.2 and 7.3, we created two weight variables: one for estimating ATE and the other for estimating ATT. Using these weights, we conducted a balance check. Precisely, we ran a weighted regression model for checking the balance of the sample on a continuous covariate (i.e., using a continuous covariate as the dependent variable and using a dichotomous treatment variable as the single independent variable) and a weighted logistic regression model for checking the balance of the sample on a dichotomous covariate (i.e., using the dichotomous covariate as the dependent variable and again using the dichotomous treatment variable as the single independent variable). Table 7.1 presents results of imbalance checking based on this method. The table presents p values associated with the regression coefficients of the treatment variable (i.e., child's use of Aid to Families With Dependent Children [AFDC]) for ATE and ATT weights, respectively. Table 7.1 Covariate Imbalance After Propensity Score Weighting (Example 7.3.1) 288 Source: Data from Hofferth et al., 2001. Note: The balance check used regression for a continuous dependent variable and logistic regression for a dichotomous dependent variable. ATE = average treatment effect where weight is $1/P$ for a treated case and $1/(1 - P)$ for a comparison case; ATT = average treatment effect for the treated where weight is 1 for a treated case and $P/(1 - P)$ for a comparison case. * $p < .05$, ** $p < .01$, *** $p < .001$, two-tailed test. Results were not as good as we had hoped. All covariates for both weights, except for gender, were imbalanced to a statistically significant degree between treated participants and controls. This finding suggests that for this data set, the propensity score model cannot remove covariate imbalance satisfactorily, and therefore, the results of weighted analysis may remain biased. For our demonstration purposes, we ran a propensity score weighting analysis, and these results are presented in Table 7.2. The analysis showed that children who used AFDC had an average score of letter-word identification that was 5.16 points lower than children who never used AFDC ($p < .001$). From a perspective of treatment effects for the treated (i.e., what is the effect if the analyst considers only those individuals assigned to the treatment condition?), we find that children who used AFDC had an average score of letter-word identification that was 4.62 points lower than children who never used AFDC ($p < .01$). Compared with ATE, ATT decreases both in size and in the level of significance. 7.3.2 Propensity Score Weighting With a Cox Proportional Hazards Model This example illustrates propensity score weighting with a Cox proportional hazards model. The example used the same data and research question as those presented in Section 5.8.1, except that the outcome analysis replaced the Kaplan-Meier product limit estimates with Cox regression. Included in this study are 2,758 children from the panel data of the National Survey of Child and Adolescent Well-Being (NSCAW). After a listwise deletion of missing data, the study sample comprises 2,723 children. Recall that the study focused on whether the use of substance abuse services by caregivers was predictive of the 289 timing of a maltreatment report in the 18-month period after the baseline interview. Study participants who did not have a report at the end of the 18-month window were defined as censored. Of the 2,723 children included in the study, 294 (10.8%) had female caregivers who received substance abuse treatment, and the remaining 2,429 (89.2%) had female caregivers who did not receive treatment services. Table 7.2 Regression Analysis of Letter-Word Identification Score in 1997 With Propensity Score Weighting (Example 7.3.1) Source: Data from Hofferth et al., 2001. Note: AFDC = Aid to Families With Dependent Children; ATE = average treatment effect where the weight is $1/P$ for a treated case and $1/(1 - P)$ for a comparison case; ATT = average treatment effect for the treated where the weight is 1 for a treated case and $P/(1 - P)$ for a comparison case. * $p < .05$, ** $p < .01$, *** $p < .001$, two-tailed test. The sample descriptive statistics, the coefficients of the logistic regression predicting the propensity scores, and balance check prior to propensity score weighting are shown in Table 6.2 of Section 6.5.2 in Chapter 6 (i.e., the section illustrating propensity score subclassification). The balance check for the propensity score-weighted analyses was conducted by running weighted logistic regression, because all covariates were operationalized as dichotomous variables. In the weighted analysis, a covariate was used as the dependent variable, and the treatment indicator was used as a predictor. The ATE and ATT weights were employed in each logistic regression model separately. Results show that propensity score weighting in general balances data. Only two covariates in the weighted analysis using ATE weights—"caregiver arrest" and "caregiver prior receipt of AOD treatment"—were statistically significant; all 290 other covariates were not statistically significant. None of the covariates was significant using ATT weights. Table 7.3 presents the results of the weighted Cox proportional hazards models. For comparison purposes, the table also shows the results from the unweighted model (i.e., the Cox regression without correcting for selection). Note that a weighted analysis in Stata, as in other software packages, is estimated automatically by an algorithm that estimates robust standard errors. To be consistent, we also estimated robust standard errors for the unweighted model. The ATE-weighted analysis shows that the hazard rate of having a maltreatment report for children whose caregivers used the substance abuse treatment is, on average, 127.1% higher than that for children whose caregivers did not use the treatment ($p < .001$), which is much higher than that estimated by the unweighted model that shows a hazard ratio of 1.547 ($p < .01$). The unweighted model indicates that the hazard rate of having a maltreatment report for children whose caregivers used substance abuse treatment is some 54.7% higher than that for children whose caregivers did not use the treatment. The ATT-weighted analysis shows that the average treatment effect for the treated children is 1.496 ($p < .05$), or the hazard rate of having a maltreatment report for the treated children is 49.6% higher than that for children whose caregivers did not use the treatment. This is slightly lower than the hazard rate difference estimated by the unweighted model. Taken together, the propensity score weighting analysis produces different results than the unweighted model, and the information obtained from the analysis is revealing for child welfare researchers, policy makers, and advocates of children who receive child protective services. 7.3.3 Propensity Score Weighting With an SEM This example illustrates propensity score weighting with an SEM. The example uses the same data and research question as those presented in Section 6.5.3. The same SEM conceptualized in Figure 6.1 is tested by a propensity score weighting approach. Specifically, this example employs GBM to estimate propensity scores for the sample participants. After obtaining estimated propensity scores, the analysis applies Equations 7.2 and 7.3 to calculate propensity score weights for estimating ATE and ATT. Finally, the analysis treats the propensity score weights as sampling weights when running the mediation model. The weighted SEM is implemented by using the software package Mplus. Figure 6.1 depicts the conceptual model. Recall that the research question focuses on whether the use of AFDC by a caretaker during his or her own childhood has both direct and indirect effects on the child's academic achievement via the mediator of the child's use of AFDC. This study used data collected from the time of the child's birth to 1997, which was the time point at 291 which academic achievement data were available. The variable for child use of AFDC was a dichotomous variable coded as never used AFDC versus ever used AFDC. It was considered the treatment variable of the study. The variable caregiver's use of AFDC during childhood was the number of years a caregiver used AFDC during the time the caregiver was a child between ages 6 and 12 years, and the variable for child academic achievement was a latent variable with four indicators. By this specification, β_{21} reflects the direct effect of child use of AFDC on achievement, γ_{21} reflects the direct effect of caregiver use of AFDC in childhood on child achievement, and $\gamma_{11}\beta_{21}$ reflects the indirect effect of caregiver use on child achievement. After a listwise deletion of cases with missing data, the study had a sample of 601 children who were between the ages of 5 and 12 years in 1997. Sample descriptive statistics and results of an imbalance check are shown in Table 6.5. Results suggest that the treatment and comparison groups differed on four covariates and the sample was imbalanced. All four covariates had a normalized difference that exceeded .25. Therefore, a correction of selection bias using propensity score weighting is warranted. Table 7.3 Estimated Cox Proportional Hazard Models (Example 7.3.2) 292 293 Source: Data from NSACW, 2004. Note: Reference group is shown next to the variable name. * $p < .05$, ** $p < .01$, *** $p < .001$, two-tailed test. Using these covariates, the study estimated propensity scores using GBM. Propensity score weights for estimating ATE based on Equation 7.2 and weights for estimating ATT based on Equation 7.3 were then created. The mediation model was fitted and estimated by the software package Mplus. With weights and continuous indicator variables, the study chose the MLR estimator (i.e., an Mplus option for maximum likelihood estimation with robust standard errors; L. K. Muthén & Muthén, 2010). Based on exploratory runs, for the ATE-weighted model, the analysis specified correlational measurement errors between letterword identification and passage comprehension, between calculation and applied problems, and between passage comprehension and calculation. For the ATT-weighted model, the analysis specified correlational measurement errors between letter-word identification and calculation, as well as between calculation and applied problems. Results using ATE weights show the model has acceptable fit to the data: Model $\chi^2(df = 5) = 2.948$, $p = .708$ and root mean square error of approximation (RMSEA = .00) with a 90% confidence interval of [.00, .04]. The estimated direct impact of child use of AFDC on academic achievement = -10.163 (SE = 1.413), the estimated impact of caregiver use of AFDC during childhood on and the estimated impact of caregiver's use of AFDC on AFDC during childhood on academic achievement (SE = .401). The indirect effect of caregiver's use of AFDC during childhood on achievement via By applying Equation 6.6, the study child use of AFDC obtained the standard error of the indirect effect as . Performing a z test by using $Z^* = 294$ the study finds that the indirect effect is statistically significant ($p < .000$). These results confirm the conceptual model of the mediating role of child use of AFDC: Other things being equal, children who ever used AFDC had an academic achievement score that was 10.163 units lower than those who never used AFDC ($p < .000$); the direct effect of caregiver's use of AFDC as a child on child achievement is $-.867$ ($p < .05$), and the indirect effect of caregiver's use of AFDC on child achievement via the mediator of child use of AFDC is $-.701$ ($p < .000$). Of the total effect of caregiver's use on child achievement, 55.3% is direct and 44.7% is indirect. Results using ATT weights are similar to those using ATE weights. The ATT-weighted model also has acceptable fit: Model χ^2 (df = 6) = 4.981, $p = .546$, and RMSEA = .00 with a 90% confidence interval of (.00, .05). Results showed that other things being equal, children who ever used AFDC had an academic achievement score that was 9.744 units lower than those who never used AFDC ($p < .000$). The direct effect of caregiver's use of AFDC on child achievement is $-.809$ ($p < .05$), and the indirect effect of caregiver's use of AFDC on child achievement via the mediator of child use of AFDC is $-.585$ ($p < .000$). Of the total effect of caregiver's use of AFDC on child achievement, 58.0% is direct and 42.0% is indirect. 7.3.4 Comparison of Models and Conclusions of the Study of the Impact of Poverty on Child Academic Achievement We have demonstrated different methods to evaluate the impact of poverty (i.e., welfare dependence) on child academic achievement. Our examples represent a common practice in propensity score analysis: That is, instead of using a single method in a specific study, we use multiple methods, conduct follow-up comparisons and sensitivity analyses, and attempt to draw conclusions based on a thorough investigation of convergences and divergences across models. So what do the results of different models tell us about the impact of poverty on academic achievement? Table 7.4 compares results from the following models: independent samples t test (Table 5.8), unadjusted ordinary least squares (OLS) regression (Table 5.8), optimal matching (fully) with the Hodges-Lehmann aligned rank test (Section 5.8.3), regressing difference scores of outcome on difference scores of covariates after optimal pair matching (Section 5.8.4), and multiple regression analysis with propensity score weighting (Section 7.3.1). We exclude results of SEM analysis using propensity score subclassification (Section 6.5.3) and the SEM analysis using propensity score weighting (Section 7.3.3) to achieve the highest level of comparability across models, because the two SEMs aimed to test a slightly different conceptual model: direct and indirect effects of caregiver use of AFDC in childhood on child achievement. On the basis of Table 7.4, we can draw the following conclusions. 295 Comparison of Findings Across Models Estimating the Impact of Poverty on Children's Academic Achievement (Example 7.3.4) Table 7.4 Source: Data from Hofferth et al., 2001. * $p < .05$, ** $p < .01$, *** $p < .001$, one-tailed test. First, all models estimated a significant treatment effect. However, it is worth mentioning that this may not always be the case when performing propensity score analysis. We know that a simple covariance control, such as regression or regression-type analysis, ignores selection bias and, therefore, tends to produce biased and inconsistent estimates about treatment effects. After correcting for endogeneity by using propensity score models, estimates may be different. In this case, the effect of AFDC participation on academic achievement is large, and therefore, effects remained significant even after we introduced sophisticated controls. Second, among all estimates, which are more accurate or acceptable? Of the six models, t test and OLS regression by design did not control for selection bias; the propensity score weighting for ATE and ATT attempted to control it but failed. Therefore, the only acceptable estimates are those offered by optimal full matching using the Hodges-Lehmann aligned rank test and the difference score regression after pair matching. The estimate is -1.97 ($p < .05$) from optimal full matching and -3.17 ($p < .05$) from pair matching. On the basis of these findings and in the context of the observed covariates, we may conclude that poverty, on average, causes a reduction in letter-word identification score by a range of 2 to 3 points. In contrast, both the independent t test and OLS regression exaggerated the impact: The t test exaggerated the impact by 398% or 210% and OLS regression exaggerated the impact by 140% 296 or 49%. Note that not only were the estimated effects exaggerated by the t test and OLS regression, but also the estimated significance level was exaggerated (i.e., both models show a p value less than .001). Finally, the examples underscore the importance of conducting propensity score analyses. Researchers should explore the use of these types of models simply because they appear—when compared with traditional approaches—to provide more precise estimates of treatment effects in observational studies where measures of potential selection artifacts are collected. 7.4 CONCLUSION Propensity score weighting is widely used. Its uptake is probably influenced by the familiarity of many researchers with the concept of weighting in survey research methods. Despite its widespread adoption, some researchers have concerns about the validity of applying the method in certain contexts. Notably the concerns drawn by Freedman and Berk (2008) from their theoretical work and data simulation are worth consideration and warrant a further investigation. Nonetheless, weighting has many promising features, particularly when outcome variables are not continuous and nonnormally distributed, such as a time-to-event variable with censoring. It is useful also when conceptual models hypothesize complex relationships, including mediation. NOTE 1. Note that we are using the term causes. Indeed, it took a long time before we could finally use the term! And our use here must still be conditioned. It is causal only in the context of observed heterogeneity, that is, selection for which we have adequate measurement. 297 CHAPTER 8 Matching Estimators This chapter focuses on a collection of matching estimators, including the simple matching estimator, the bias-corrected matching estimator, the variance estimator assuming a constant treatment effect and homoscedasticity, and the variance estimator allowing for heteroscedasticity (Abadie et al., 2004; Abadie & Imbens, 2002, 2006). Matching is a shared characteristic among these estimators and those described in Chapters 5 and 9. However, the matching estimators presented in this chapter do not use logistic regression to predict propensity scores. They require fewer decisions, are easy to implement, and do not involve nonparametric estimation of unknown functions. Because of these advantages, these matching estimators are an attractive approach for solving many problems encountered in program evaluation. Section 8.1 provides an overview of the matching estimators. We compare these methods with those discussed in other chapters and give particular attention to similarities and differences in the methodological features. Our intent in this section is to provide contextual information sufficient to differentiate matching estimators from other methods of propensity score analysis. Section 8.2 is the core of the chapter and describes the methodology of matching estimators. Although the focus of this review is on the simple matching estimator and the bias-corrected matching estimator, we review Abadie and Imbens's (2006) study on the large sample properties of matching estimators. Section 8.3 summarizes key features of the Stata nmatch program, which can be used to run matching estimators. Section 8.4 gives detailed examples. Because matching estimators estimate average treatment effects for the treated, Section 8.4 also shows how to use matching estimators to conduct efficacy subset analyses (ESAs) that test hypotheses pertaining to levels of treatment exposure (i.e., dose analyses). Section 8.5 presents the conclusion of the chapter. 8.1 OVERVIEW As discussed in Chapter 2, a seminal development in the conceptualization of program evaluation was the Neyman-Rubin counterfactual framework. The key assumption of the framework is that individuals selected into treatment and 298 nontreatment groups have potential outcomes in both states: the one in which the outcomes are observed and the one in which the outcomes are not observed. Thus, for the treated group, in addition to an observed mean outcome under the condition of treatment $E(Y_1|W = 1)$, the framework assumes that an unobserved mean outcome exists under the condition of nontreatment $E(Y_0|W = 1)$. Similarly, participants in the control group have both an observed mean $E(Y_0|W = 0)$ and an unobserved mean $E(Y_1|W = 0)$. The unobserved potential outcomes under either condition are missing data. Based on the counterfactual framework, the matching estimators directly impute the missing data at the unit level by using a vector norm. That is, at the unit level, a matching estimator imputes potential outcomes for each study participant. Specifically, it estimates the value of $Y_i(0)|W_i = 1$ (i.e., potential outcome under the condition of control for a treatment participant) and the value of $Y_i(1)|W_i = 0$ (i.e., potential outcome under the condition of treatment for a control participant). After imputing the missing data, matching estimators can be used to estimate various average treatment effects, including the sample average treatment effect (SATE), the population average treatment effect (PATE), the sample average treatment effect for the treated (SATT), the population average treatment effect for the treated (PATT), the sample average treatment effect for the controls (SATC), and the population average treatment effect for the controls (PATC). As previously mentioned in Section 5.2, a central challenge inherent to observational study is the dimensionality of covariates or matching variables. In essence, as the number of matching variables increases, so does the difficulty of using exact matching to find a match for a given treated participant. The methods described in Chapters 5 and 9 use logistic regression to predict propensity scores that, in turn, are used to reduce multiple matching variables to a single score. They solve the dimensionality problem. As such, treated and control participants who have the same propensity score values are deemed to have the same distributions on the observed covariates. The matching estimators do not use logistic regression to predict propensity scores. Instead, these methods use a vector norm to calculate distances on the observed covariates between a treated case and each of its potential control cases. The vector norm is used to choose the outcome of a control case whose distance on covariates is the shortest vis-à-vis other control cases. This outcome serves as the counterfactual for the treated case. Similarly, the matching estimators can be used to choose the outcome of a treated case whose distance on covariates is the shortest vis-à-vis other treated cases. This outcome serves as the counterfactual for the control case. To calculate a vector norm, the matching estimators choose one of two kinds of variance matrices: the inverse of the sample variance matrix or the inverse of the sample variance-covariance matrix. When choosing the inverse of the 299 sample variance-covariance matrix to calculate the vector norm, the matching estimator calculates Mahalanobis metric distances. Thus, as described in Section 5.4.1, this method becomes Mahalanobis metric matching, which was developed prior to the invention of propensity score matching (Cochran & Rubin, 1973; Rubin, 1976, 1979, 1980a). Notably, even though these methods share features of conventional approaches, the matching estimators developed by Abadie and Imbens (2002, 2006) expand the Mahalanobis metric matching in several important ways. They estimate (a) average treatment effects for the controls, (b) treatment effects that include both sample and population estimates, (c) variances and standard errors for statistical significance tests, and (d) a bias correction for finite samples (when the Mahalanobis metric matching is not exact). Two assumptions are embedded in the matching estimators: (1) the assignment to treatment is independent of the outcomes, conditional on the covariates, and (2) the probability of assignment is bounded away from 0 and 1, which is also known as an overlap assumption requiring sufficient overlap in the distributions of the observed covariates (Abadie et al., 2004). The first assumption is the same as the strongly ignorable treatment assignment assumption (Rosenbaum & Rubin, 1983), the fundamental assumption that we have examined in several discussions in this book. This is a restrictive and strong assumption, and, in many cases, it may not be satisfied. However, the assumption is a conceptual link that connects observational studies with nonrandomized designs to theories and principles developed for randomized designs. At some point and in one way or another, most evaluation analyses are conditioned on the ignorable treatment assignment assumption (Abadie et al., 2004; Imbens, 2004). The overlap assumption requires some overlapping of the estimated propensity scores in the treatment and control conditions. It implies that the treated and control groups share a common support region of propensity scores in the sample data. If this assumption is violated, it is not appropriate to use the matching estimators. Under such a condition, researchers might want to consider using optimal matching (see Section 5.4.2), which is more robust against violations of overlap. Alternatively, a trimming procedure developed by Crump and colleagues (2009) can be applied to address the limited overlap problem (see Section 6.2). As discussed earlier, matching estimators share a key characteristic with all other methods introduced in Chapters 5 and 9, that is, matching estimators match a treated case to a control (or vice versa) based on observed covariates. However, the matching estimators use a simpler mechanism for matching (i.e., a vector norm), which makes the approach much easier to implement than other methods. Unlike other methods in which matching is based on propensity scores (e.g., nearest neighbor [Section 5.4.1] or optimal matching using the network flow theory [i.e., Section 5.4.2]), the matching estimators do not require 300 postmatching analysis such as survival analysis, hierarchical linear modeling, the Hodges-Lehmann aligned rank test, or regression adjustment based on difference scores. Omitting the postmatching analysis is an advantage because reducing the number of analytic procedures involved in matching also reduces the number of subjective decisions a researcher has to make. This is usually desirable. The matching estimators allow evaluators to estimate effects for both the sample and the population. The difference between the SATE and the PATE is based on whether the effect can be inferred beyond the study sample. As such, SATE and PATE are useful for answering different questions. Abadie et al. (2004) used a job-training program to illustrate the differences. SATE is useful in evaluating whether the job-training program (i.e., the sample data at hand) was successful. In contrast, if the evaluator wants to know whether the same program would be successful in a second sample from the population, PATE is useful. Although SATE and PATE produce the same coefficients, they estimate standard errors differently. Indeed, the estimated standard error of the population effect is, in general, larger than that of the sample effect and, therefore, may lead to different conclusions about significance. This variation makes sense because a successful treatment may exist in one sample and not in another sample from the same population. The relationship between SATE and PATE is similar to the relationship between SATT and PATT and to the relationship between SATC and PATC. SATC and PATC are estimable only by the matching estimators. In the case of SATC, the average treatment effect for the control group is based on this question: What would the sample treatment effect for the treated group look like if the controls received the treatment condition and the treatment cases received the control condition? Similarly, for PATC, the average treatment effect for the control condition indicates what the population treatment effect for the treated group would look like if the controls received the treatment condition and the treatment cases received the control condition. If the variables used in matching accounted for all cases of selection bias and the evaluation data met all the assumptions of the matching estimators, SATT and SATC would appear to have values of similar magnitude. Thus, the difference between the two coefficients holds the potential to indicate both the level of hidden selection bias and the departure of data from model assumptions. However, although this sounds quite useful, it has a serious limitation. In practice, both problems may work together to exert joint and entangled effects. Nonetheless, a difference between SATT and SATC can be an indicator that alternative analyses with different assumptions should be tried. In Chapter 2, we reviewed the stable unit treatment value assumption, or SUTVA (Rubin, 1986), which holds that the potential outcomes for any unit do not vary with the treatments assigned to any other units, and there are no different versions of the treatment. As described in Chapter 2, SUTVA is an 301 assumption that facilitates the investigation or estimation of counterfactuals as well as a conceptual perspective that underscores the importance of using appropriate estimators when analyzing differential treatment effects. The SUTVA assumption imposes exclusion restrictions on outcome differences. Based on these restrictions, economists underscore the importance of analyzing average treatment effects for the subpopulation of treated units because that analysis is frequently more important than the effect on the population as a whole. Analysis for the subpopulation is especially a concern when evaluating the importance of a narrowly targeted intervention, such as a labor market program (Heckman, Ichimura, & Todd, 1997, 1998; Imbens, 2004). Matching estimators offer an easy tool for this task. What statisticians and econometricians have called evaluating average treatment effects for the treated (also treatment of the treated, TOT) is similar to the efficacy subset analysis (ESA) framework that is found in the intervention research literature (Lochman, Boxmeyer, Powell, Roth, & Windle, 2006). Because the matching estimators evaluate a potential outcome for each study unit, it is not cumbersome to estimate the average treatment effects for user-defined subsets of units and then test research hypotheses related to differential treatment exposure. Following the tradition of the econometric literature, but in contrast to some of the statistical literature, the matching estimators focus on matching with replacement (Abadie & Imbens, 2006). That is, the matching estimators allow individual observations to be used as a match more than once. Matching with replacement makes matching estimators different from all matching methods described in Chapter 5. Denoting $KM(i)$ as the number of times unit i is used as a match—given that M matches per unit are used—the matching estimators described in this chapter allow $KM(i) > 1$, whereas the matching methods described in Chapter 5 require $KM(i) \leq 1$. According to Abadie and Imbens (2006), Matching with replacement produces matches of higher quality than matching without replacement by increasing the set of possible matches. In addition, matching with replacement has the advantage that it allows us to consider estimators that match all units, treated as well as controls, so that the estimand is identical to the population average treatment effect. (p. 240) When matching with replacement, the distribution of $KM(i)$ plays an important role in the calculation of the variance of the estimator and, therefore, $KM(i)$ is a key factor in examining the large sample properties of matching estimators (Abadie & Imbens, 2006). The development of variances for various matching estimators is especially attractive, because it makes possible significance testing of treatment effects 302 estimated by matching estimators. Unlike matching with nonparametric regression that relies on bootstrapping to draw statistical inferences (i.e., described in Chapter 9), the matching estimators offer a consistent estimator for variance estimation and thus allow for a more rigorous significance test than those methods that use bootstrapping. Research to date suggests that "bootstrapping methods for estimating the variance of matching estimators do not necessarily give correct results" (Abadie et al., 2004, p. 300). Although the matching estimators offer several advantages, they also share a common limitation with other matching methods. That is, most matching estimators contain a conditional bias term whose stochastic order increases with the number of continuous variables. As a result, matching estimators are consistent. Abadie and Imbens (2006) have shown that the variance of not matching estimators remains relatively high. Consequently, matching with a fixed number of matches does not lead to an efficient estimator; specifically, it does not achieve the semi-parametric efficiency bound calculated by Hahn (1998). These problems are discussed in Section 8.2. From our own simulation studies, we have found that the matching estimators are more sensitive to the violation of the strongly ignorable assumption than other methods, which means that, under certain conditions (e.g., high selection on observables), matching estimators will produce larger bias in effect estimates, when compared to other methods. We discuss this issue in Chapter 11 when comparing the sensitivity of different estimators to selection bias. 8.2 METHODS OF MATCHING ESTIMATORS In this section, we review the basic methodological features of the matching estimators. Primarily, our review follows the work of Abadie et al. (2004) and Abadie and Imbens (2006). We focus on two issues: (1) the point estimates of various treatment effects (i.e., the coefficients of SATE, PATE, SATT, PATT, SATC, and PATC) using either a simple matching estimator or a bias-corrected matching estimator and (2) the estimates of the variances of various treatment effects (i.e., the standard errors of SATE, PATE, SATT, PATT, SATC, and PATC) assuming homoscedasticity or allowing for heteroscedasticity. In practice, researchers will use a combination of one method for estimating the coefficient and one method for estimating the variance (equivalently the standard error). A typical evaluation uses bias-corrected matching with a variance estimation allowing for heteroscedasticity. All methods described in this section are labeled jointly as the collection of matching estimators. 8.2.1 Simple Matching Estimator For each observation i , the unit-level treatment effect is $\tau_i = Y_i(1) - Y_i(0)$. As discussed earlier, one of the two outcomes is always missing (i.e., either $Y_i(0)$ or $Y_i(1)$ is missing, depending on whether the unit's treatment condition is $W_i = 1$ or $W_i = 0$). Under the exogeneity and overlap assumptions, the simple matching estimator imputes the missing potential outcome by using the average outcome for individuals with "similar" values on observed covariates. We first consider matching with one observed covariate. This is the most basic case of simple matching, under which condition the estimator simply takes the outcome value of the controlled case that is the closest match on the observed covariate. Under the condition of tie, the estimator takes the mean value of outcomes of the tied cases. Similarly, the estimator takes the outcome value of the treated case that is the closest match on the observed covariate for a control case. Let $JM(i)$ denote the set of indices for the matches for unit i that are at least as close as the M th match and $\# JM(i)$ denote the number of elements of $JM(i)$; the simple matching estimator can be expressed by the following equations: Table 8.1 illustrates how the simple matching estimator works when there is a single covariate x_1 . In this example, three units are controls ($W = 0$), four units are treated ($W = 1$), and one covariate x_1 is observed. The imputed values of potential outcomes are highlighted. For the first unit (a control case with $x_1 = 2$), the exact match is the fifth unit (because x_1 for $i = 5$ is also equal to 2), so the imputed value $y(1)$ (i.e., the potential outcome under the condition of treatment for $i = 1$) is 8 by taking the value of y for $i = 5$. For the second unit (a control case with $x_1 = 4$), treated units 4 and 6 are equally close (both with $x_1 = 3$, a closest value to 4), so the imputed potential outcome $y(1)$ is the average of the outcomes for units 4 and 6, namely, $(9 + 6)/2 = 7.5$. Note that for $i = 2$, $\#JM(i) = 2$, because two matches ($i = 4$ and $i = 6$) are found for this participant. We next consider matching with more than one covariate. Under this condition, the simple matching estimator uses the vector norm (i.e., $\|x\|_v = \sqrt{x'v}$ with positive definite matrix v ; see Note 1 at the end of the chapter) to calculate distances between one treated case and each of its multiple possible controls and chooses the outcome of the control case whose distance is the shortest among all as the potential outcome for the treated case. In a similar fashion, the simple matching estimator imputes the missing outcome under the condition of treatment for a control case. Specifically, we define $\|z - x\|_v$ as the distance between vectors x and z , where x represents the vector of the observed 304 covariate values for observation i , and z represents the vector of the covariate values of a potential match for observation i . There are two choices for v : the inverse of the sample variance matrix (i.e., a diagonal matrix with all off-diagonal elements constrained to be zero) or the inverse of the sample variance-covariance matrix (i.e., a nondiagonal matrix with all off-diagonal elements to be nonzero covariances). When the inverse of the sample variance-covariance matrix is used, $\|z - x\|_v$ becomes a Mahalanobis metric distance. As noted in Chapter 5, some researchers (e.g., D'Agostino, 1998) defined the variance-covariance matrix differently than the matching estimators described in this chapter for the Mahalanobis metric matching, where they use the inverse of the variance-covariance matrix of the control observations, rather than that of all sample observations (i.e., both the treated and control observations) to define v . Table 8.1 An Example of Simple Matching With One Observed Covariate For Seven Observations Let $dM(i)$ denote the distance from the covariates for unit i , X_i , to the M th nearest match with the opposite treatment condition. Allowing for the possibility of ties, at this distance, fewer than M units are closer to unit i than $dM(i)$ and at least M units are as close as $dM(i)$. With multiple covariates, $JM(i)$ still represents the set of indices of the matches for unit i that are at least as close as the M th match, but M matches are chosen using a vector norm that meets the condition of nearest distances as follows: If there are no ties, the number of elements in $JM(i)$ is M but may be larger. Previously we mentioned that $KM(i)$ is the number of times unit i is used as a match given that M matches per unit are used. With the new notation introduced, we can now define $KM(i)$ more precisely as the number of times i is used as a match for all observations l of the opposite treatment condition, each time 305 weighted by the total number of matches for observation l : where $\mathbb{1}\{\cdot\}$ is the indicator function, which is equal to 1 if the expression in brackets is true but equal to 0 otherwise. If we denote the total sample size as N , the number of treated cases as N_1 , and the number of controls as N_0 , then we can express these numbers by using the notation we just introduced: Under the condition of more than one observed covariate, the missing potential outcome for each unit i is imputed by using the same equations as Equation 8.1. Unlike the case of one observed covariate, the imputation of potential outcomes now uses the vector norm. Using hypothetical data, we provide an illustration of simple matching using the vector norm when matching uses multiple covariates. Table 8.2 shows an example of simple matching using three observed covariates for seven observations. The table illustrates minimum distance for each unit that was calculated by the vector norm using the inverse of a sample variance matrix. Details of calculating the minimum distances are explained in the following. Table 8.2 An Example of Simple Matching With Three Observed Covariates For Seven Observations With Minimum Distance Determined by Vector Norm Using the Inverse of a Sample Variance Matrix First, an x vector is formed by the values of x_1 , x_2 , and x_3 for the unit being considered for imputation of a potential outcome. Thus, the x vector for $i = 1$ is (2 4 3). This unit is a control case ($W = 0$). Because there are four treated units in the sample data (i.e., $i = 4, 5, 6, 7$), there are four distances for unit $i = 1$. 306 For each of the treated units, a z vector is formed by the values of x_1 , x_2 , and x_3 ; that is, z for $i = 4$ is (3 4 6), z for $i = 5$ is (2 3 4), z for $i = 6$ is (3 3 2), and z for $i = 7$ is (1 1 3). With seven observations, the sample variances are $\text{Var}(x_1) = 1.810$, $\text{Var}(x_2) = 1.810$, and $\text{Var}(x_3) = 1.952$. Hence, the inverse of the sample variance matrix is Using vector norm $\|z - x\|_v$, where v is the inverse of the sample variance matrix, we calculate the four distances for $i = 1$ as follows: Distance between the pair of $i = 1$ and $i = 4$: Distance between the pair of $i = 1$ and $i = 5$: Distance between the pair of $i = 1$ and $i = 6$: Distance between the pair of $i = 1$ and $i = 7$: 307 The preceding calculation verifies two values shown in Table 8.2: for $i = 1$ (i.e., the first row), the minimum distance equals 1.0648, which is the distance between $i = 1$ and $i = 5$ because a minimum of 1.0648 is the smallest of the four distances we just calculated, and the imputed value $y^*(1) = 8$, which is the observed outcome for $i = 5$. We chose the outcome of $i = 5$ as $y^*(1) = 8$ because $i = 5$ has the minimum distance on observed covariates from $i = 1$. Potential outcomes for all other units can be obtained by replicating the above process. Minimum distance can also be determined by the Mahalanobis metric distance that defines v as the inverse of the sample variance-covariance matrix. Table 8.3 illustrates the calculation of minimum distance for each unit that is based on the inverse of a sample variance-covariance matrix. The distances so calculated are exactly the same as Mahalanobis metric distances. We explain the main features of this calculation on the following pages. Table 8.3 An Example of Simple Matching With Three Observed Covariates For Seven

Observations With Minimum Distance Determined by Vector Norm Using the Inverse of a Sample Variance-Covariance Matrix First, the x vector for $i = 1$ is (2 4 3), and the z vector for $i = 4$ is (3 4 6), for $i = 5$ is (2 3 4), for $i = 6$ is (3 3 2), and for $i = 7$ is (1 1 3). The inverse of the sample variance-covariance matrix for this data set of seven observations is Using vector norm $\|z - x\|_w$, we can calculate the four distances for $i = 1$ as follows: Distance between the pair of $i = 1$ and $i = 4$: 308 Distance between the pair of $i = 1$ and $i = 5$: 5 Distance between the pair of $i = 1$ and $i = 6$: 2.5518 Distance between the pair of $i = 1$ and $i = 7$: 6. The preceding calculation verifies two values shown in Table 8.3: for $i = 1$ (i.e., the first row), the minimum distance = 2.5518, which is the distance between $i = 1$ and $i = 6$, because 2.5518 is the smallest of the four distances, and the imputed value = 6, which is the observed outcome for $i = 6$. Potential outcomes for all other units can be obtained by replicating the above process. Finally, we consider the calculation of various treatment effects. After imputing the missing potential outcomes, we now have two outcomes for each study unit. One is an observed outcome, and the other is an imputed potential outcome (or the counterfactual). Taking average values in a varying fashion, we obtain point estimates of various treatment effects as follows: Sample average treatment effect (SATE): 309 Sample average treatment effect for the treated (SATT): Sample average treatment effect for the controls (SATC): As noted earlier, a population effect is exactly the same as the point estimate of its corresponding sample effect. Thus, PATE = SATE, which can be obtained by applying Equation 8.2; PATT = SATT, which can be obtained by applying Equation 8.3; and PATC = SATC, which can be obtained by applying Equation 8.4. When running matching estimators, the number of matches for each unit must be considered. In the previous examples, we used all units in the opposite treatment condition as potential matches and selected a single unit based on the minimum distance. Because matching estimators use matching with replacement, in theory, all controls can be selected as matches for a treated unit, and all treated units can be selected as matches for a control unit. Going to the other extreme, the researcher can choose only one match for each unit using the nearest neighbor. The drawback of using one match is that the process uses too little information in matching. Like all smoothing parameters, the final inference of matching estimators can depend on the choice of the number of matches. To deal with this issue, Abadie et al. (2004) recommend using four matches for each unit, "because it offers the benefit of not relying on too little information without incorporating observations that are not sufficiently similar" (p. 298). 8.2.2 Bias-Corrected Matching Estimator Abadie and Imbens (2002) found that when the matching is not exact, the simple estimator will be biased in finite samples. Specifically, with k continuous covariates, the estimator will have a bias term that corresponds to the matching discrepancies (i.e., the differences in covariates between matched units and their matches). To remove some of the bias that remains after matching, Abadie and Imbens (2002) developed a bias-corrected matching estimator. The adjustment uses a least squares regression to adjust the difference within the matches for the differences in their covariate values. The regression adjustment is made in four steps: 1. Suppose that we are estimating the SATE. In this case, we run regressions using only the data in the matched sample. Define $\mu(w) = E\{Y(w)|X = x\}$ 310 for $w = 0$ (control condition) or $w = 1$ (treatment condition). Using the data of the matched sample, we run two separate regression models: One uses the data of $w = 0$, and the other uses the data of $w = 1$. Each regression model uses $Y(w)$ as a dependent variable, and all covariates are used as independent variables. 2. At this stage, we have obtained two sets of regression coefficients, one for $w = 0$ and one for $w = 1$. Let the intercept of the regression function be and the slope vector be W we choose and that minimize the weighted sum of squared residuals using $KM(i)$ as a weight. Precisely, the adjustment term for $w = 0, 1$, is a predicted value based on the following equation: where In other words, we choose squared residuals and that minimize the weighted sum of 3. After obtaining the adjustment term for $w = 0$ and $w = 1$, we can use the term to correct the bias embedded in the simple matching estimator. The bias-corrected estimator then uses the following equations to impute the missing potential outcomes: 4. The above steps illustrating the bias correction process use the point estimate SATE as an example (equivalently PATE, because the two coefficients are exactly the same). If we are interested in estimating SATT or PATT, we then need only estimate the regression function for the controls. . If we are interested in estimating SATC or PATC, we then need only estimate the regression function for the treated. The bias-corrected matching estimator can always be used to replace the simple matching estimator. It is especially useful when matching on several covariates of which at least one is a continuous variable. This correction is needed because exact matching is seldom exact when matching uses continuous covariates. For an application of the bias-corrected matching estimator, see Hirano and Imbens (2001). On balance, the most important function of biascorrected matching is to adjust for poor matches, and the method does not really 311 correct for bias except in some special and unrealistic instances. For instance, the method cannot (and is not designed to) correct bias generated by the omission of important covariates or bias due to hidden selection. 8.2.3 Variance Estimator Assuming Homoscedasticity Abadie and Imbens (2002) developed a variance estimator for various treatment effects. They first considered a variance estimator under two assumptions: (1) The unit-level treatment effect $\tau_i = Y_i(1) - Y_i(0)$ is constant, and (2) the conditional variance of $Y_i(W)$ given X_i does not vary with either the covariates x or the treatment w , which is known as an assumption of homoscedasticity. Under these assumptions, we can obtain variances for various effects by applying the following formulas. Variance of SATE: Variance of SATT: Variance of SATC: Variance of PATE: where Variance of PATT: 312 Variance of PATC: Taking the square root of each variance term, we obtain a standard error of the point estimate. We can then use the standard error either to calculate a 95% confidence interval of the point estimate or to perform a significance test at a given level of statistical significance. The ratio of the point estimate over the standard error follows a standard normal distribution, which allows users to perform a z test. 8.2.4 Variance Estimator Allowing for Heteroscedasticity The assumptions about a constant treatment effect and homoscedasticity may not be valid for certain types of evaluation data. In practice, evaluators may have data in which a term of conditional error variance in Equations 8.6 to 8.11 varies by treatment condition w and covariate vector x . To deal with this problem, Abadie and Imbens (2002) developed a robust variance estimator that allows for heteroscedasticity. The crucial feature of this robust estimator is its utility for estimating a conditional error variance for all sample points. The algorithm includes a second matching procedure such that it matches treated units to treated units and control units to controls. When running a computing software package, such as `nnmatch` in Stata, users must specify the number of matches used in the second matching stage across observations of the same treatment condition. This number need not be the same as the number of matches used in estimating the treatment effect itself. For details of the robust estimator, see Abadie et al. (2004, p. 303). 8.2.5 Large Sample Properties and Correction Although matching estimators, including those developed by other authors such as Cochran and Rubin (1973), Rosenbaum and Rubin (1983), and Heckman and Robb (1985), have great intuitive appeal and are widely used in practice, their formal large sample properties were not established until recently. This delay was due in part to the fact that matching with a fixed number of matches is a highly nonsmooth function of the distribution of the data, which is not amenable to standard asymptotic methods for smooth functionals. Recently, however, 313 Abadie and Imbens (2006) developed an analytic approach to show the large sample properties of matching estimators. In the following, we briefly highlight their findings. Abadie and Imbens's (2006) results indicated that some of the formal large sample properties of matching estimators are not very attractive. First, they demonstrated that matching estimators include a conditional bias term whose stochastic order increases with the number of continuous matching variables, and therefore, matching The order of this bias term may be greater than estimators are not consistent. Second, in general, the simple matching consistent. However, the simple matching estimator does not estimator is achieve the semi-parametric efficiency bound as calculated by Hahn (1998). For cases when only a single continuous covariate is used to match, Abadie and Imbens have shown that the efficiency loss can be made arbitrarily close to zero by allowing a sufficiently large number of matches. Third, despite these poor formal properties, matching estimators are extremely easy to implement and do not require consistent nonparametric estimation of unknown functions. As such, matching estimators have several attractive features that may account for their popularity. Fourth, Abadie and Imbens proposed an estimator of the conditional variance of the simple matching estimator that does not require consistent nonparametric estimation of unknown functions. This conditional variance is essentially estimated by the variance estimator that allows for heteroscedasticity and involves a second matching procedure (see Section 8.2.4). The crucial idea is that instead of matching treated units to controls, the estimator of the conditional variance matches treated units to treated units and control units to controls in the second stage. Finally, on the basis of results of the large sample properties of matching estimators, Abadie and Imbens concluded that bootstrapping is not valid for matching estimators. 8.3 OVERVIEW OF THE STATA PROGRAM NNMATCH Software for implementing the matching estimators is available in Stata, MATLAB, and R. In this section, we review a user-developed program in Stata called `nnmatch`. It processes all the estimators described in this chapter. As a user-developed program, `nnmatch` is not included in the regular Stata package. To search the Internet for this software, users can use the `findit` command, followed by `nnmatch` (i.e., `findit nnmatch`), and then follow the online instructions to download and install the program. After installation, users should check the help file to obtain basic instructions for running the program. The work of Abadie et al. (2004), published by The Stata Journal, was written specifically to address how to use `nnmatch` for evaluating various treatment effects discussed in this chapter; users may find this reference helpful. The `nnmatch` program can be initiated using the following basic syntax: 314 `nnmatch depvar treatvar varlist, c(att) m(#) metric(maha) biasadj(bias) /// robust(#) population keep(filename)` In this command, `depvar` is the outcome variable on which users want to assess the difference between treated and control groups, that is, the outcome variable showing treatment effect; `treatvar` is the binary treatment membership variable that indicates the intervention condition; and `varlist` specifies the covariates to be used in the matching. These statements are required and must be specified. The rest of the statements are optional, and omission of their specifications calls for default specifications. The term `c` specifies the type of treatment effects to be evaluated, and three values may be specified in the parentheses: `ate`, `att`, and `atc`, which stand for average treatment effect, average treatment effect for the treated, and average treatment effect for the controls, respectively. By default, `nnmatch` estimates `ate`. The `m` specifies the number of matches that are made per observation. Users replace `#` in the syntax with a specific number and include that number in the parentheses. The term `metric` specifies the metric for measuring the distance between two vectors of covariates, or the type of variance matrix users selected to use in the vector norm. By default, `nnmatch` uses the inverse of sample variance matrix; `metric(maha)` causes `nnmatch` to use the inverse of the sample variance-covariance matrix in the vector norm and to evaluate Mahalanobis metric distances. The term `biasadj(bias)` specifies that the bias-corrected matching estimator be used. By default, `nnmatch` uses the simple matching estimator. If the user specifies `biasadj(bias)`, `nnmatch` uses the same set of matching covariates, `varlist`, to estimate the linear regression function in the bias adjustment. Alternatively, the user can specify a new list of variables in the parentheses, which causes `nnmatch` to enter a different set of covariates in the regression adjustment. The term `robust(#)` specifies that `nnmatch` estimate heteroscedasticity-consistent standard errors using `#` matches in the second matching stage. A specific number is used to replace `#` in the preceding syntax, and users should include that number in the parentheses. The number need not be the same as that specified in `m(#)`. By default, `nnmatch` uses the homoscedastic/constant variance estimator. The term `population` causes `nnmatch` to estimate a population treatment effect (i.e., `ate`, `att`, or `atc`). By default, `nnmatch` estimates a sample treatment effect. Last, `keep(filename)` saves a temporary matching data set in the file `filename.dta`. A set of new variables is created and saved in the new Stata data file that may be used for follow-up analysis. Table 8.4 exhibits the syntax and output of six `nnmatch` examples: estimation of SATE, PATE, SATT, PATT, SATC, and PATC using four matches per observation, as well as bias-corrected and robust variance estimators that allow 315 for heteroscedasticity. Substantive findings of these examples are discussed in the next section. 8.4 EXAMPLES In this section, we use two examples to illustrate the application of matching estimators in program evaluation. The first example shows the evaluation of six treatment effects using a bias-corrected estimator with variance estimator allowing for heteroscedasticity. The second example illustrates the application of matching estimators to an ESA for which doses of treatment becomes a central concern. The second example also runs all analyses following missing data imputation based on 50 imputed files. As such, the second example mimics study conditions that are likely to be found in real program evaluation. 8.4.1 Matching With Bias-Corrected and Robust Variance Estimators This section presents an example showing the application of matching with biascorrected and robust variance estimators. Using a sample and matching variables similar to those in the example given in Section 5.8.2, we analyze a subsample of 606 children from the data of the 1997 Child Development Supplement (CDS) to the Panel Study of Income Dynamics (PSID) and the core PSID annual data from 1968 to 1997 (Hofferth et al., 2001). The primary study objective was to test a research hypothesis regarding the causal effect of childhood poverty on developmental outcomes, specifically academic achievement. The study tested the effect of participation in a welfare program—an indicator of poverty—on test performance. We refer readers to Section 2.8.5 for a review of the conceptual framework and substantive details of the study. Table 8.4 Exhibit of Stata `nnmatch` Syntax and Output Running Bias-Corrected Matching Estimators With Robust Standard Errors 316 317 318 Source: Data from Hofferth et al., 2001. For this application, we report findings from the examination of a single domain of academic achievement: the age-normed passage comprehension score of the Woodcock-Johnson Revised Tests of Achievement (Hofferth et al., 2001). Higher scores on this measure are considered an indication of higher academic achievement. The passage comprehension score is used as the outcome variable in this study. Readers should note that this study analyzed data for 606 children rather than the 1,003 children included in the analysis presented in Section 5.8.2. Because 397 cases were missing passage comprehension scores, those cases were removed from the analytic sample. On the basis of the research question, we classified the study participants into two groups: children who ever used AFDC (Aid to Families With Dependent Children) from birth to current age in 1997 and those who never used AFDC during the same period. Thus, this dichotomous variable indicated the treatment condition in the study: those who ever used AFDC versus controls who never used AFDC. Of the 606 study children, 188 had used AFDC and were considered the treated group, and 418 participants had never used AFDC and were considered the control group. To assess the treatment effect (i.e., participation in the AFDC program as an indicator of poverty) on academic achievement, we considered the following covariates or matching variables: (a) current income or poverty status, which was measured as the ratio of family income to poverty threshold in 1996; (b) caregiver's education in 1997, which was measured as years of schooling; (c) caregiver's history of using welfare, which was measured as the number of years (i.e., a continuous variable) the caregiver used AFDC during the caregiver's childhood (i.e., ages 6 to 12 years); (d) child's race, which was 319 measured as African American versus non-African American; (e) child's age in 1997; and (f) child's gender, which was measured as male versus female. It is worth noting that of the six matching variables, four were continuous variables and only child's race and gender were categorical variables. Given this condition, it is impossible to conduct exact matching for this data set, and it is therefore important to use the bias-corrected matching estimator to correct for bias corresponding to the matching discrepancies between matched units and their matches on the four continuous covariates. In our example, we used the same set of matching variables as the independent variables for the regression adjustment in the bias correction process. Following the recommendation of Abadie et al. (2004), we chose four matches per observation in the analysis. The choice of a variance estimator deserves some explanation. Note that the homoscedastic variance estimator assumes that the unit-level treatment effect is constant and that the conditional variance of $Y_i(w)$ given X_i does not vary with either covariates or the treatment. To test whether our data met the homoscedastic assumption, we first ran a regression of the passage comprehension scores on the six matching variables plus the binary treatment variable. We then performed the Breusch-Pagan and Cook-Weisberg tests of heteroscedasticity for each of the seven independent variables. Our results from the Breusch-Pagan and Cook-Weisberg tests showed that child's age was statistically significant ($p < .000$) and indicated that the conditional variance of the outcome variable was not constant across levels of child's age. On the basis of this finding, we decided to use the robust variance estimator that allows for heteroscedasticity. We used the same number of matches (i.e., four matches) in the second matching stage to run the robust variance estimator. Table 8.5 presents results of our analysis of the study data. The interpretation and findings of the study may be summarized as follows. First, as noted earlier, a specific sample effect is the same as its corresponding population effect in magnitude (e.g., both SATE and PATE are equal to -4.70). The two effects differ from each other only on the standard error (e.g., the standard error for SATE was 1.76969, whereas the standard error for PATE was 1.765187). Second, our results suggested that childhood poverty strongly affected the children's academic achievement. On average, children who used AFDC in childhood had a passage comprehension score 4.7 units lower than that of children who had never used AFDC in childhood. This finding held true after we took selection bias into consideration for six observed covariates. With regard to the subpopulation of treated participants, the treatment effect was even larger: -5.23 , or 0.53 units larger than the sample (or population) average treatment effect. Third, had all controls (i.e., children who never used AFDC) used AFDC and all treated children not used AFDC, then on average, the control children would have a passage comprehension score 4.47 units lower than their counterparts. Note that in this study, SATT equaled -5.23 and SATC 320 equaled -4.47 , or a difference of 0.76 units. This difference is attributable either to additional selection bias that was not accounted for in the study or to study data that violated assumptions of matching estimators, which suggests the need for further scrutiny. Fourth, a population effect indicates whether the tested intervention will be effective in a second sample taken from the same population. Taking SATT ($p = .003$) and PATT ($p = .002$) as examples, the study indicated that the treatment effect for the treated group was statistically significant in the sample at a level of .01. If we take a second sample from the population, we are likely to observe the same level of treatment effect for the treated, and the effect should remain statistically significant at a level of .01. Finally, our results showed that all six treatment effects were statistically significant ($p < .05$), and all 95% confidence intervals did not contain a zero. Thus, we concluded that the study data could not reject a null hypothesis of a zero treatment effect, and represented by participation in the AFDC program and conditioned on the available data, childhood poverty appears to be an important factor causing children's poor achievement in passage comprehension. Table 8.5 Estimated Treatment Effects (Effects of Child's Use of AFDC) on Passage Comprehension Standard Score in 1997 Using Bias-Corrected Matching With Robust Variance Estimators (Example 8.4.1) Source: Data from Hofferth et al., 2001. Note: 95% CI = 95% confidence interval. 8.4.2 Efficacy Subset Analysis With Matching Estimators In this example, we illustrate how to use matching estimators to conduct an ESA, that is, to estimate treatment effects by dosage or exposure level. In Chapter 10, we describe a method of dose analysis that uses ordered logistic 321 regression or multinomial logistic regression or a generalized propensity score estimator to predict propensity scores, and then performs a nonbipartite matching or propensity score weighting analysis or dose-response function analysis. Because the matching estimators directly evaluate sample and population average treatment effects for the treated, they permit a direct ESA to test hypotheses regarding treatment dosage. Therefore, we selected the matching estimators to conduct the dosage analysis. This example uses a sample and matching variables similar to those of the example described in Section 4.4.2. The findings reported here represent preliminary findings from an evaluation of the "Social and Character Development" program that was implemented in North Carolina. The study sample comprised more than 400 study participants. The North Carolina intervention included a skills-training curriculum, Making Choices, which was designed for elementary school students. The primary goals of the Making Choices curriculum were to increase the social competence and to reduce the aggressive behavior of students. The treatment group received a multiemeent intervention, which included 29 Making Choices classroom lessons delivered over the course of the third-grade year and eight follow-up or "booster shot" classroom lessons delivered in each of the fourth and fifth-grade years. Students assigned to the control group received regular character development and health education instruction and did not receive any lessons from the Making Choices curriculum. Several valid and reliable instruments were used to measure student social competence and aggressive behavior, and outcome data were collected for both the treated students and controls at the beginning and end points of each academic year for the third-, fourth-, and fifth-grade years. Therefore, six waves of panel data were available for the evaluation. The example used only the outcome data collected during the fourth and fifth grades; that is, the data measuring behavior change 1 or 2 years after the intervention. For most students, outcome data were collected by different teachers at different grades, and these data were likely to reflect some rater effects. To remove the rater effects, we analyzed change scores within a grade (i.e., using an outcome variable at the end point of a grade minus the outcome at the beginning point of the same grade). Typically, scores within a grade were made by the same teachers. As mentioned in Section 4.4.2, the Making Choices intervention used group randomization with 14 schools (Cohort 1 = 10 schools; Cohort 2 = 4 schools). However, despite the cluster randomized design, a preliminary analysis showed that the sample data were not balanced on many observed covariates. This indicated that the group randomization had not worked as planned. In some school districts, as few as four schools met the study criteria and were eligible for participation. As a result of the smaller than anticipated sample, the two intervention schools differed systematically on covariates from the two control schools. When the investigators compared data from these schools, they found 322 that the intervention schools differed from the control schools in several significant ways: The intervention schools had (a) lower academic achievement scores on statewide tests (Adequate Yearly Progress), (b) a higher percentage of students of color, (c) a higher percentage of students receiving free or reduced-price lunches, and (d) lower mean scores on behavioral composite scales at baseline. Using bivariate tests and logistic regression models, the researchers found that these differences were statistically significant at the .05 level. The researchers were confronted with the failure of randomization. Had these selection effects not been taken into consideration, the evaluation of the program effectiveness would be biased. Hence, it is important to use propensity score approaches, including matching estimators, to correct for the selection bias in evaluation. Like most evaluations, the data set contained missing values for many study variables. Before performing the evaluation analysis, we conducted a missing data imputation using the multiple imputation method (R. A. Little & Rubin, 2002; Schafer, 1997). With this method, we generated 50 imputed data files for each outcome variable. Results from our analysis showed that with 50 data sets, the imputation achieved a relative efficiency of 99%. Following the convention of practice in imputing missing data, we imputed missing values for all cases, but cases that had missing data on the outcome variable were deleted. As such, the sample size for final analysis varies by outcome variable. With multiply imputed files (i.e., 50 distinct data files in this example), we first ran `nnmatch` for each file and then used Rubin's rule (R. A. Little & Rubin, 2002; Schafer, 1997) to aggregate the point estimates and standard errors to generate one set of statistics for the significance test for each outcome variable. We refer readers to this text's companion webpage, where we provide the syntax for running `nnmatch` 50 times for each outcome and for the aggregation using Rubin's rule. To analyze the outcome changes that occurred in the fourth- and fifth-grade years, we first analyzed the change scores for the sample as a whole using three methods: (1) optimal pair matching, which was followed by regression adjustment; (2) optimal full matching, which was followed by the Hodges-Lehmann aligned rank test; and (3) matching estimators. We provide results of these analyses in Table 8.6. Table 8.6 Estimated Treatment Effects Measured as Change Scores in the Fourth and Fifth Grades by Three Estimators (Example 8.4.2) 323 Source: Data from SACD, 2008. Note: $+p < .1$, $*p < .05$, $**p < .01$, $***p < .001$, two tailed. Results from the first two methods (i.e., optimal pair matching and optimal full matching) were not promising. On the basis of the design of the intervention program, we expected positive findings (i.e., the intervention would be effective in changing behavioral outcomes); however, none of the results from the optimal pair matching with regression adjustment was statistically significant. The situation improved slightly with the results of the optimal full matching with the Hodges-Lehmann test, in which some outcomes showed a statistical trend ($p < .10$), and two variables (i.e., social competence and prosocial) showed statistical significance ($p < .05$). When faced with such situations, researchers need to seek a plausible explanation. We thought that there were at least two plausible explanations for the nonsignificant findings: One is that the intervention was not effective, and the other is that our evaluation data violated assumptions embedded in the evaluation methods we had used and, therefore, the results reflect methodological artifacts. In practice, there is no definitive way to find out which explanation is true. However, the results of the third analytic method, matching estimators, showed more significant findings than the previous two analyses. It is worth noting that even when using matching estimators, we are still likely to find that the intervention was not effective. Moreover, we might also find that results from the matching estimators are methodologically erroneous, but in this case, the error goes in the other direction: The violation of model assumptions might have produced "overly optimistic" findings. The discussion of which of the three methods is most suitable to our study and which set of results is more trustworthy is beyond the scope of this chapter. However, we emphasize that 324 using multiple approaches in evaluation research is important, particularly for programs whose effects may be marginal, or whose effect sizes fall into the "small" category defined by Cohen (1988). As underscored by Cohen, detecting a small effect size is often an important objective for a new inquiry in social and health sciences research. This point was also emphasized by Sosin (2002), whose study used varying methods, including sample selection, conventional control variable, instrumental variable, and propensity score matching, to examine a common data set. Sosin found that the various methods provided widely divergent estimates. In light of this finding, he suggested that researchers regularly compare estimates across multiple methods. Returning to our example, we first considered findings for the whole sample and then divided the sample into subsets based on treatment dosage. We then conducted ESA on each of the subsets. In this study, we defined dosage as the number of minutes a student received Making Choices classroom lessons. Based on this definition (which relied on teacher reports of implementation and fidelity), the dosage for all controls was zero. Furthermore, we defined three subsets on the basis of grade-level booster shot dosage: (1) adequate (and recommended) exposure to intervention, that is, students who received Making Choices for more than 240 minutes and less than 379 minutes; (2) high exposure to intervention, that is, students who received unusually high program exposure of more than 380 minutes of Making Choices lessons; and (3) the control group, who received zero minutes of Making Choices. Table 8.7 shows the sample size and distribution of the three subsets by grade. With regard to ESA, the subsets defined above allowed us to make two comparisons for each grade: the adequate exposure group versus the comparison group, and the high exposure group versus the comparison group. Our general hypothesis was that the Making Choices intervention produces desirable behavioral changes for the treatment group, and students who have adequate or more exposure to the intervention will show higher levels of changed behavior. Table 8.7 Sample Size and Distribution of Exposure Time to Program Intervention ("Dosage") by Grade (Example 8.4.2) 325 Source: Data from SACD, 2008. Note: Adequate = adequate exposure to intervention (240 to 379 minutes); High = high exposure to intervention (380 or more minutes); Comp. = comparison group (0 minutes). The matching we conducted for our analysis had the following characteristics: we chose bias-corrected matching that used the same set of matching variables in the regression model as for the bias correction, we used four matches per observation, our matching strategy used the robust variance estimator to allow for heteroscedasticity and used four matches per observation in the second matching stage, and our matching estimated SATT. We present the results of the ESA in Table 8.8. On balance, the ESA confirmed the research hypotheses for two core outcomes for the fourth-grade year (i.e., aggression and relational aggression). That is, the treatment effects for the treated on these outcomes were not only statistically significant but also in the same direction of the hypothetical sign. Furthermore, the high-exposure group exhibited a larger effect than the adequate-exposure group. Second, for the social competence measures, including prosocial behavior and emotional regulation, the direction and size of the findings are relatively consistent over the fourth and fifth grades. This pattern is consistent also with the hypotheses. Greater variation is observed in the high-exposure groups where sample sizes are smaller. Finally, the sample treatment effects for the treated group were statistically significant for most outcomes in the analyses of the adequate-exposure group in both the fourth- and fifth-grade years. In sum, the ESA suggested that the recommended exposure level of 240 minutes of program content produces positive effects. At that dose level, the Making Choices program appears to promote prosocial behavior and reduce aggressive behavior. Program exposure at a higher level may have some gain for fourth-grade children, but it appears to have a negligible effect for fifth-grade children. On balance, this conclusion is consistent with the theory, objective, and design of the Making Choices intervention. 326 Table 8.8 Efficacy Subset Analysis Using Matching Estimators: Estimated Average Treatment Effects for the Treated (SATT) by Dosage (Example 8.4.2) Source: Data from SACD, 2008. Note: $+p < .1$, $*p < .05$, $**p < .01$, $***p < .001$, two tailed. 8.5 CONCLUSION This chapter described a method that is widely used in observational studies: matching. The discussion focuses on the collection of matching estimators developed by Abadie and Imbens (2002). These include the vector norm, bias correction using a linear regression function, and robust variance estimation involving second-stage matching. Recently, formal study has examined the large sample properties of various matching estimators, including those developed by researchers other than Abadie and Imbens. In general, matching estimators are consistent, and therefore, the results of large sample properties are not so attractive. However, using a correction that matches treated units to treated units and control units to controls, the robust estimator of asymptotic variance has proven to be promising. Matching estimators are intuitive and appealing, but caution is warranted in at least two areas of application. First, matching with continuous covariates poses ongoing challenges. When matching variables are continuous, users need to be cautious and make adjustments, such as using a bias-corrected matching estimator, for bias. Note that the presence of continuous matching variables does not appear to be a severe problem in propensity score matching, where 327 continuous as well as categorical variables are treated as independent variables in the logistic regression predicting the propensity scores. Second, bootstrapping to estimate variances is problematic, and the direct estimation procedure developed by Abadie and Imbens (2002) appears preferable. Note that the bootstrapping method is used in statistical inference for matching with nonparametric regression, which is a topic we tackle in the next chapter. The matching estimators discussed in this chapter handle heteroscedasticity, but they do not correct for inefficiency induced by clustering. In social and health sciences research, intracluster correlation, or clustering, is frequently encountered in program evaluation (Guo, 2005). However, the current version of matching estimators does not take the issue of clustering into consideration. The developers of matching estimators are aware of this limitation and are working toward making improvements to matching estimators that will provide adjustments for clustering (G. Imbens, personal communication, October 10, 2007). Despite limitations, the collection of matching estimators described in this chapter offers a clear and promising approach to balancing data when treatment assignment is nonignorable. The method is easy to implement and requires of the researcher few subjective decisions. NOTE 1. The vector norm originally defined by Abadie et al. (2004) was $\|x\|_w = (\sum V_i x_i^2)^{1/2}$, that is, a square root of the quantity $\sum V_i x_i^2$. We verified results from the Stata program `nnmatch` developed by Abadie et al. We found that the vector norm used by `nnmatch` did not take the square root. Therefore, in our presentation of the vector norm, we removed the square root and defined the vector norm as $\|x\|_w = \sum V_i x_i^2$. This modification did not change the nature of minimum distance because, if a is a minimum value among remains a minimum value among a_1, a_2, \dots, a_n . 328 CHAPTER 9 Propensity Score Analysis With Nonparametric Regression In this chapter, we review propensity score analysis with nonparametric regression. This method was developed by Heckman, Ichimura, and Todd (1997, 1998). The 1997 article from Heckman et al. shows the application of propensity score analysis with nonparametric regression to evaluations of jobtraining programs, whereas the 1998 study presents a rigorous distribution theory for the method. A central feature of this method is the application of nonparametric regression (i.e., local linear regression with a tricube kernel, also known as lowess) to smooth unknown and possibly complicated functions. The method allows estimation of treatment effects for the treated by using information from all possible controls within a predetermined span. Because of this feature, the method is sometimes called kernel-based matching (Heckman et al., 1998). The model is sometimes called a difference-in-differences approach (Heckman et al., 1997), when it is applied to two-time-point data (i.e., analyzing pre- and posttreatment data) to show change triggered by an intervention in a dynamic fashion. The three terms—propensity score analysis with nonparametric regression, kernel-based matching, and the difference-in-differences method—are used interchangeably in this chapter. Although the asymptotic properties of lowess have been established, it is technically complicated to program and calculate standard errors based on these properties. Therefore, to implement the estimator, bootstrapping is used to draw statistical inferences. In general, this method uses propensity scores derived from multiple matches to calculate a weighted mean that is used as a counterfactual. As such, kernel-based matching is a robust estimator. Section 9.1 provides an overview of propensity score matching with nonparametric regression. Section 9.2 describes the approach by focusing on three topics: the kernel-based matching estimators and their applications to two-time-point data, a heuristic review of lowess, and a review of issues pertaining to the asymptotic and finite-sample properties of kernel-based matching. Section 9.3 summarizes key features of two computing programs in Stata (i.e., `psmatch2` and `bootstrap`) that can be used to run all the models described in this chapter. Section 9.4 presents an application with two-time-point data. Because local 329 linear regression can be applied to postintervention outcomes that do not constitute a difference-in-differences, we also show how to use it in evaluations with one-time-point data. Section 9.5 presents the conclusion to the chapter. 9.1 OVERVIEW In contrast to kernel-based matching, most matching algorithms described in previous chapters are 1-to-1 or 1-to- n (where n is a fixed number) matches. That is, 1-to-1 and 1-to- n methods are designed to find one control or a fixed number of controls that best match a treated case on a propensity score or on observed covariates X . In practice, this type of matching is not very efficient because we may find controls that sum to more than n for each treated case within a predetermined caliper. Often, the number of controls close to a treated case varies within a caliper, but information on the relative closeness of controls is ignored. Kernel-based matching constructs matches using all individuals in the potential control sample in such a way that it takes more information from those who are closer matches and downweights more distal observations. By doing so, kernel-based matching uses comparatively more information than other matching algorithms. Both kernel-based matching and optimal matching use a varying number of matches for each treated case; however, the two methods employ different approaches. Kernel-based matching uses nonparametric regression, whereas optimal matching uses network flow theory from operations research. Optimal matching aims to minimize the total distance and uses differential weights to take information from control cases. But it does so by optimizing a "cost" defined by a network flow system. As described in Section 5.4.2 and Equation 5.6, the total distance an optimal matching algorithm aims to minimize is a weighted average, which is similar to kernel-based matching. However, optimal matching uses three methods to choose a weight function, all of which depend on the proportion of the number of treated cases (or the proportion of the number of controls) in a matched set to the total number of treated cases (or controls), or the proportion of both treated and control cases falling in set s among the sample total. It is the choice of the weighting function that makes kernel-based matching fundamentally different from optimal matching. In choosing a weighting function, kernel-based matching uses lowess, a nonparametric method for smoothing unknown and complicated functions. As discussed in Chapter 2, Heckman (2005) sharply contrasted the econometric model of causality to the statistical model of counterfactuals. This is reflected in the development of kernel-based matching. Heckman and his colleagues argued that two assumptions embedded in Rosenbaum and Rubin's (1983) framework for propensity score matching (i.e., the strongly ignorable treatment assignment and overlap assumptions) were too strong and restrictive. Under these conditions, conceptually different parameters (e.g., the 330 mean effect of treatment on the treated, the mean effect of treatment on the controls, and the mean effect of

randomly assigning persons to treatment) become the same (Heckman et al., 1998). Heckman and his colleagues thought that these three effects should be explicitly distinguished, and for the evaluation of narrowly targeted programs such as a job-training program, they argued that the mean effect of treatment on the treated is most important. Kernel-based matching serves this purpose, and it is a method for estimating the average effect of treatment on the treated (ATT). For an elaboration of the ATT perspective, see Heckman (1997) and Heckman and Smith (1998). To overcome what they saw as the limitations of Rosenbaum and Rubin's framework and to identify the treatment effect on the treated, Heckman and colleagues developed a framework that contained the following key elements. First, instead of assuming strongly ignorable treatment assignment, or $(Y_0, Y_1) \perp W|X$, they imposed a weaker condition assuming $Y_0 \perp W|X$. Under this assumption, only the outcome under the control condition is required to be independent of the treatment assignment, conditional on observed covariates. Second, instead of assuming full independence, the Heckman team imposed mean independence, or $E(Y_0|W = 1, X) = E(Y_1|W = 0, X)$. That is, conditional on covariates, only the mean outcome under the control condition for the treated cases is required to be equal to the mean outcome under the treated condition for the controls. Third, their framework included two crucial elements: separability and exclusion restrictions. Separability divides the variables that determine outcomes into two categories: observables and unobservables. This separation permits the definition of parameters that do not depend on unobservables. Exclusion restrictions isolate different variables that determine outcomes and program participation. Specifically, the exclusion restriction partitions the covariate X into two sets of variables (T, Z) , where the T variables determine outcomes, and the Z variables determine program participation.² Putting these elements together, Heckman and colleagues developed a framework suitable for the evaluation of the treatment effect for the treated, or $E(Y_1 - Y_0|W = 1, X)$, rather than the average treatment effect $E(Y_1|W = 1) - E(Y_0|W = 0)$. This framework extended Rosenbaum and Rubin's framework to a more general and feasible case by considering $U_0 \perp W|X$, where U_0 is an unobservable determining outcome Y_0 . Under the exclusion restrictions, Heckman and colleagues further assumed that the propensity score based on X (i.e., $P(X)$) equals the propensity score based on program participation (i.e., $P(Z)$), or $P(X) = P(Z)$, which leads to $U_0 \perp W|P(Z)$. By these definitions and assumptions, Heckman et al. argued that their framework no longer implied that the average unobservables under the control condition, conditional on the propensity score of program participation, equaled zero. Similarly, the Heckman group argued that their framework no longer 331 implied that the average unobservables under the treatment condition, conditional on the propensity score of program participation, equaled zero. More formally, there is no need under this framework—they contended—to assume $E(U_0|P(Z)) = 0$ or $E(U_1|P(Z)) = 0$, where U_1 is an unobservable determining outcome Y_1 . The only required assumption under the framework is that the distributions of the unobservables are the same in the populations of $W = 1$ and $W = 0$, once data are conditioned on $P(Z)$. Thus, Heckman and colleagues devised an innovative general framework for the evaluation of program effects, and they addressed constraints in Rosenbaum and Rubin's framework, which rely on stronger and more restrictive assumptions. With a fixed set of observed covariates X , should an analyst develop propensity score $P(X)$ using X and then match on $P(X)$, or should an analyst match directly on X ? Heckman and colleagues (1998) compared the efficiency of the two estimators (i.e., an estimator that matches on propensity score $P(X)$ and an estimator that matches on X directly), and concluded, "There is no unambiguous answer to this question" (p. 264). They found that neither estimator was necessarily more efficient than the other and that neither estimator was practical because both assume that the conditional mean function and $P(X)$ are known values, whereas in most evaluations, these values must be estimated. When the treatment effect is constant, as in the conventional econometric evaluation models, they reported an advantage for matching on X rather than on $P(X)$. However, when the outcome Y_1 depends on X only through $P(X)$, there is no advantage to matching on X over matching on $P(X)$. Finally, when it is necessary to estimate $P(X)$ or $E(Y_0|W = 0, X)$, the dimensionality of the X is indeed a major drawback to the practical application of the matching method. Both methods (i.e., matching on X or matching on $P(X)$) are "data-hungry" statistical procedures. On the basis of their study, Heckman and his colleagues recognized that matching on $P(X)$ avoids the dimensionality problem by estimating the mean function conditional on a one-dimensional score. The advantage of using the propensity score is simplicity in estimation. Therefore, in the kernel-based matching, Heckman et al. developed a two-stage procedure to first estimate $P(X)$ and then use nonparametric regression to match on $P(X)$. It is important to note that there exists some skepticism about the weaker assumptions developed by Heckman and colleagues. The question is whether the so-called weaker assumptions are so much weaker that in practice they would be any easier to evaluate than the stronger assumptions associated with Rosenbaum and Rubin's framework. For instance, Imbens (2004) noted that although the mean-independence assumption is unquestionably weaker, in practice it is rare that a convincing case is 332 made for the weaker assumption without the case being equally strong for the stronger version. The reason is that the weaker assumption is intrinsically tied to functional-form assumptions, and as a result one cannot identify average effects on transformations of the original outcome (such as logarithms) without the stronger assumption. (p. 8) 9.2 METHODS OF PROPENSITY SCORE ANALYSIS WITH NONPARAMETRIC REGRESSION We will start the discussion of kernel-based matching by highlighting a few of its key features. Specifically, we note how kernel-based matching constructs a weighted average of counterfactuals for each treated case and then how this method calculates the sample average treatment effect for all treated cases. In this case, we assume that the weighting procedure is given; that is, we ignore the computational details of nonparametric regression. Notwithstanding, applying nonparametric regression to propensity score analysis is a creative and significant contribution made by Heckman and his colleagues. The development of nonparametric regression stems from empirical problems in smoothing complicated mathematical functions, and these have general objectives beyond the field of observational studies. To more fully explain the logic of nonparametric regression, we provide a heuristic review of lowess. Specifically, we focus on the main features of local linear regression and the determination of weights by using various functions. We conclude this section by summarizing the main findings of studies that have examined the asymptotic and finite-sample properties of lowess. 9.2.1 The Kernel-Based Matching Estimators The kernel and local linear matching algorithms were developed from nonparametric regression methods for curve smoothing (Heckman et al., 1997, 1998; J. A. Smith & Todd, 2005). These approaches enable the user to perform one-to-many matching by calculating the weighted average of the outcome variable for all nontreated cases and then comparing that weighted average with the outcome of the treated case. The difference between the two terms yields an estimate of the treatment effect for the treated. A sample average for all treated cases (denoted as ATT in Equation 9.1) is an estimation of the sample average treatment effect for the treated group. Hence, the one-to-many matching method combines matching and analysis (i.e., comparison of mean difference on outcome measure) into one procedure. Denote I_0 and I_1 as the set of indices for controls and program participants, respectively, and Y_0 and Y_1 as the outcomes of control cases and treated cases, respectively. To estimate a treatment effect for each treated case $i \in I_1$, 333 outcome Y_{1i} is compared with an average of the outcomes Y_{0j} for matched case $j \in I_0$ in the untreated sample. Matches are constructed on the basis of propensity scores $P(X)$ that are estimated using the logistic regression on covariates X . Precisely, when the estimated propensity score of an untreated control is closer to the treated case $i \in I_1$, the untreated case gets a higher weight when constructing the weighted average of the outcome. Denoting the average treatment effect for the treated as $ATT(i, j)$, the method can be expressed by the following equation: where n_1 is the number of treated cases, and the term $W(i, j)$ measures the weighted average of the outcome for all nontreated cases that match to participant i on the propensity score differentially. It is worth noting that in $W(i, j)Y_{0j}$ sum over all controls $j \in I_0 \cap Sp$. This feature is Equation 9.1, a crucial element of kernel-based matching because it implies that each treated case matches on all controls falling into the common-support region rather than 1-to-1 or 1-to- n . Furthermore, this element implies that the estimator forms a weighted average by weighting the propensity scores differentially or using different weights of $W(i, j)$. In Equation 9.1, $W(i, j)$ is the weight or distance on propensity score between i and j . We explain how to determine $W(i, j)$ in the next subsection. Heckman, Ichimura, and Todd (1997, 1998) used a difference-in-differences method, which is a special version of the estimated ATT effect. In this situation, Heckman and his colleagues used longitudinal data (i.e., the outcomes before and after intervention) to calculate the differences between the outcome of the treated cases and the weighted average differences in outcomes for the nontreated cases. Replacing Y_{1i} with $(Y_{1it} - Y_{1it'})$, and Y_{0j} with $(Y_{0jt} - Y_{0jt'})$, where t denotes a time point after treatment and t' a time point before treatment, we obtain the difference-in-differences (DID) estimator: Note that each treated participant has a difference $(Y_{1it} - Y_{1it'})$ and his or her multiple matches have an average difference. The difference between the two values yields the difference-in-differences that measures the average change in outcome that is the result of treatment for a treated case $i \in I_1$. Taking the average over all treated cases of the sample—that is, taking the —analyst obtains the difference-in-average of differences estimate of sample treatment effect for the treated cases. 334 In Chapter 5, we discussed the common-support region problem that is frequently encountered in propensity score matching. As illustrated in Figure 5.2, propensity score matching typically excludes participants from the study when they have no matches. Cases cannot be matched because treated cases fall outside the lower end of the common-support region (i.e., cases with low logit) and nontreated cases fall outside the upper end of the common-support region (i.e., cases with high logit). Even for matched cases, the potential for matches at the two ends of the region may be sparse, which means the estimation of treatment effects for the treated is not efficient. To deal with this problem, Heckman, Ichimura, and Todd (1997) recommended a trimming strategy: The analyst should discard the nonparametric regression results in regions where the propensity scores for the nontreated cases are sparse and typically use different trimming specifications to discard 2%, 5%, or 10% of study participants at the two ends.3 Researchers may treat these trimming specifications as sensitivity analyses. Under different trimming specifications, findings indicate the sensitivity of treatment effects for the treated to the distributional properties of propensity scores. 9.2.2 Review of the Basic Concepts of Local Linear Regression (lowess) Nonparametric regression methods are used to determine $W(i, j)$ of Equations 9.1 and 9.2. In this subsection, we describe the general ideas of the kernel estimator and the local linear regression estimator that use a tricube kernel function, where the second estimator is known as lowess. Readers are referred to Hardle (1990) as a comprehensive reference that describes the general nonparametric regression approach. In addition, Fox (2000) is a helpful reference for kernel-based matching. This subsection is primarily based on Fox (2000), and we include an example originally developed by Fox to illustrate the basic concepts of nonparametric regression.4 As previously mentioned, nonparametric regression is a curve-smoothing approach (often called scatterplot smoothing), but what is curve smoothing? In typical applications, the nonparametric regression method passes a smooth curve through the points in a scatterplot of y against x . Consider the scatterplot in Figure 9.1 that shows the relationship between the life expectancy of females (the y variable) and gross domestic product (GDP) per capita (the x variable) for 190 countries (J. Fox, personal communication via e-mail, October 1, 2004). The dashed line was produced by a linear least squares regression. The linear regression line is also known as a parametric regression line because it was where and are the two produced by the parametric regression parameters of interest. In Figure 9.1, the linear regression line does not very accurately reflect the relationship between life expectancy (equivalently the mortality rate) and economic development. To improve the fit of the line, a researcher could 335 analyze the relationship between y and x by seeking a mathematical function and, then using that function, drawing a smooth curve that better conforms to the data. Unfortunately, the relationship between these two variables may be too complicated to be developed analytically. Now look at the solid line in Figure 9.1. Although it is imperfect, it is a more accurate representation of the relationship between life expectancy and economic development. This line is produced by nonparametric regression. We now turn to the central question, "How do we draw a smooth curve using nonparametric regression?" Figure 9.1 Illustration of the Need for a Better Curve Smoothing Using Nonparametric Regression Source: Data and syntax from Fox, 2004. Before we can draw a smooth curve using nonparametric regression, we first need to do local averaging. Local averaging means that for any focal point x_0 , the magnitude of its y value is a local weighted average determined by all y values neighboring x_0 . Consider Figure 9.2, where we defined the 120th ordered x value as our focal point or x_{120} . (That is, you sort the data by x in an ascending order first and then choose any one value of x as your focal point. By doing so, the neighboring observations close to the focal point, x_{120} in the current instance, are neighbors in terms of the focal point.) The chosen country x_{120} in this example is the Caribbean nation of Saint Lucia, whose observed GDP per capita was \$3,183 and life expectancy for females was 74.8 years. We then 336 define a window, called a span, that contains $0.5N = 95$ observations of the data using the focal point as the center. The span is bounded in the figure by the dashed lines. As the scatter shows, within the span, some countries had female life expectancies greater than that of Saint Lucia (i.e., > 74.8), whereas some countries had female life expectancies less than that of Saint Lucia (i.e., < 74.8). , should be intuitively, the local average of the y value for x_{120} , denoted as z close to 74.8, but it should not be exactly 74.8. We take the y values of all neighboring points in the span into consideration, so that the smoothed curve represents the relationship between the x and y variables. Thus, the local average is a weighted mean of all y values falling into the span such that it gives greater weight to the focal point x_{120} and its closest neighbors and less weight to distant points when constructing the weighted mean. The method of constructing the weighted mean for a focal point (i.e.,) using various kernel functions is called the kernel estimator. Figure 9.2 The Task: Determining the y Value for a Focal Point x_{120} Source: Data and syntax from Fox, 2004. Let $z_i = (x_i - x_0)/h$ denote the scaled, signed distance between the x value for the i th observation and the focal x_0 , where the scale factor h is determined by the kernel function. The fraction that is used to determine the number of observations that fall into a span is called bandwidth. In our example, we defined a span containing 50% of the total observations centering on the focal 337 point; thus, the value of bandwidth is 0.5. The kernel function, denoted as $G(z_i)$, . Having is a function of z_i and is the actual weight to form the fitted value of calculated all $G(z_i)$ s within a bandwidth for a focal point, the researcher can obtain a fitted value at x_0 by computing a weighted local average of the y s as Several methods have been developed to estimate a kernel function. Common kernel functions include the following: 1. The tricube kernel: For this kernel function, h is the number of observations falling into a span centered at the focal x_0 when calculating $z_i = (x_i - x_0)/h$. 2. The normal kernel (also known as the Gaussian kernel) is simply the standard normal density function: For this kernel function, h is the standard deviation of a normal distribution centered at x_0 when calculating $z_i = (x_i - x_0)/h$. Other kernel functions include (a) the rectangular kernel (also known as the uniform kernel), which gives equal weight to each observation in a span centered at x_0 and, therefore, produces an unweighted local average, and (b) the Epanechnikov kernel, which has a parabolic shape with support $[-1, 1]$ and the kernel is not differentiable at $z = \pm 1$. The tricube kernel is a common choice for the kernel-based matching. As shown in Figure 9.3, the weights determined by a tricube kernel function follow a normal distribution within the span. In the following, we use the data from the example to show the calculation of for the focal point x_{120} . Figure 9.4 shows results of the the weighted mean calculation. In our calculation, the focal point x_{120} is Saint Lucia, whose GDP per capita is \$3,183 (i.e., $x_{120} = 3,183$). For South Africa (the nearest neighbor country below Saint Lucia on the data sheet in Figure 9.4), the z and $GT(z)$ can be determined as follows: 338 Figure 9.3 Weights Within the Span Can Be Determined by the Tricube Kernel Function Source: Data and syntax from Fox, 2004. because $|.4947| < 1$, we obtain $GT(z)$ by taking a tricube function using Equation 9.4: $GT(z) = (1 - |z|)^3 = (1 - |.4947|)^3 = .68$. Taking another neighboring country (see Figure 9.4), Slovakia, as an example, the z and $GT(z)$ can be determined as because $|.8737| < 1$, we obtain $GT(z)$ by taking a tricube function using Equation 9.4: $GT(z) = (1 - |z|)^3 = (1 - |.8737|)^3 = .04$. Taking Venezuela as an example, the z and $GT(z)$ can be determined as because $|3.2947| > 1$, we obtain $GT(z) = 0$ using Equation 9.4. Note that with respect to the closeness of x_i to x_0 , Venezuela is viewed as a distal country relative to Saint Lucia, because Venezuela's $|z|$ is greater than 1. As such, Venezuela's weight is $GT(z) = 0$, which means that Venezuela contributes . Using Equation 9.3, we nothing to the calculation of the weighted mean obtain the weighted mean for the focal point x_{120} as 339 Figure 9.4 The y Value at the Focal Point x_{120} Is a Weighted Mean Source: Data and syntax from Fox, 2004. Calculation of the Weighted Mean for x_{120} : The y Value at the Focal Point x_{120} Is a Weighted Mean Source: Data and syntax from Fox, 2004. In the illustration, some 94 countries fell into the span, but we used only five indeed, of the five, only three proximately neighboring countries to calculate qualify as having an absolute value of z that is less than 1 under 340 All other countries within the span would have a zero weight $GT(z)$ like that for Poland and Venezuela. They would contribute nothing to the calculation of . The preceding calculation is for one focal country, Saint Lucia. We now replicate the above procedure for the remaining 189 countries to obtain 190 weighted means. We obtain a smooth curve by connecting all 190 weighted means, such as that shown in Figure 9.5. The procedure to produce this smooth curve is called kernel smoothing, and the method is known as a kernel estimator. Figure 9.5 The Nonparametric Regression Line Connects All 190 Average Values Source: Data and syntax from Fox , 2004. In contrast to the kernel estimator, local linear regression (also called local polynomial regression or lowess) uses a more sophisticated method to calculate the fitted y values. Instead of constructing a weighted average, local linear regression aims to construct a smooth local linear regression with estimated β_0 and β_1 that minimizes where $G((x_i - x_0)/h)$ is a tricube kernel. Note that 341 which can be determined by the same kernel estimator previously described. is a predicted value With a local linear regression, the fitted y value or falling onto a regression line, where the regression line is typically not parallel to the x -axis. Figure 9.6 shows how lowess predicts a fitted y value locally. Connecting all 190 fitted y values produced by the local linear regression, the researcher obtains a smoothed curve of lowess that should look similar to Figure 9.5. Figure 9.6 The Local Average Now Is Predicted by a Regression Line. Instead of a Line Parallel to the x -Axis Source: Data and syntax from Fox, 2004. As described previously, bandwidth is the fraction that is used to define the span (in the above example, it was equal to 0.5). Therefore, bandwidth determines the value of h , which is the number of observations falling into the span (in the previous example, $h = 95$). The choice of bandwidth affects the level of smoothness of the fitted curve, and it is an important specification that affects the results of kernel-based matching. We reexamine this issue in our discussion of the finite-sample properties of lowess. We have reviewed two methods of nonparametric regression: the kernel estimator and local linear regression. The primary purpose for this review is to describe the determination of $W(i, j)$ that is used in Equations 9.1 and 9.2. Now, thinking of the x -axis of a scatterplot as a propensity score and the y -axis as a 342 outcome variable on which we want to estimate average treatment effect for the treated, $W(i, j)$ becomes the weight derived from the distance of propensity score between a treated case $i \in I_1$ and each nontreated case $j \in I_0$. For each treated case i , there are n_0 weights or $n_0|W(i, j)|$ s, where n_0 stands for the number of nontreated cases. This condition exists because each treated case i would have distances on propensity scores from all nontreated cases. By using either the kernel estimator or lowess, kernel-based matching calculates $W(i, j)$ s for each i in such a way that it gives a large value of $W(i, j)$ to a that has a short distance (is more proximal) on the propensity score from i , and a small value of $W(i, j)$ to a that has a greater distance (is more distal) on the propensity score from i . Precisely, kernel matching employs the kernel estimator to determine $W(i, j)$ by using the following equation: where $G(\cdot)$ is a kernel function; h is the number of observations falling into the bandwidth; P_i, P_j , and P_k are estimated propensity scores; and P_i is a focal point within the bandwidth. At present, the tricube kernel is the most common and is the recommended choice for $G(\cdot)$. Unlike the notation we used for the previous review, P_i is the focal point or a propensity score for a treated case for which we want to establish the weighted average of counterfactuals, and P_j and P_k are the propensity scores of the j th and k th nontreated cases falling into the span, that is, $j \in I_0$ and $k \in I_0$. Local linear matching employs local linear regression or lowess with a tricube function to determine $W(i, j)$ by using the following equation: where $G(\cdot)$ is a tricube kernel function and $G_{ij} = ((P_j - P_i)/h)$. In evaluations, local linear matching based on Equation 9.7 appears somewhat more common than kernel matching based on Equation 9.6. In choosing one or the other, Smith and Todd (2005) advise, Kernel matching can be thought of as a weighted regression of Y_0 on an intercept with weights given by the kernel weights, $W(i, j)$, that vary with the point of evaluation. The weights depend on the distance between each comparison group observation and the participant observation for which the counterfactual is being constructed. The estimated intercept provides 343 the estimate of the counterfactual mean. Local linear matching differs from kernel matching in that it includes in addition to the intercept a linear term in P_i . Inclusion of the linear term is helpful whenever comparison group observations are distributed asymmetrically around the participant observations, as would be the case at a boundary point of P or at any point where there are gaps in the distribution of P . (pp. 316–317) 9.2.3 Asymptotic and Finite-Sample Properties of Kernel and Local Linear Matching Several studies have examined the asymptotic properties of kernel and local linear matching methods (Fan, 1992, 1993; Hahn, 1998; Heckman et al., 1998). Between the two methods, local linear regression appears to have more promising sampling properties and a higher minimax efficiency (Fan, 1993). This may explain, in part, the predominance of local linear matching in practice applications. Heckman et al. (1998) presented an asymptotic distribution theory for kernel-based matching estimators. Beyond the scope of this book, this theory involves proofs of the asymptotic properties of the kernel-based estimators. Heckman et al. argued that the proofs justify the use of estimated propensity scores (i.e., conducting kernel-based matching using estimated propensity score $P(X)$ rather than matching on X directly) under general conditions about the distribution of X . Notwithstanding, in practice the implications of the asymptotic properties of kernel-based matching remain largely unknown. When applying nonparametric regression analysis of propensity scores to finite samples, especially small samples, the extent to which asymptotic properties apply or make sense is far from clear. Under such a context, researchers should probably exercise caution in making statistical inferences, particularly when sample sizes are small. Frölich (2004) examined the finite-sample properties for the kernel and local linear matching. Two implications are perhaps especially noteworthy: (1) It is important to seek the best bandwidth value through cross-validation of the nonparametric regression estimator, and (2) trimming (i.e., discarding nonparametric regression results in regions where the propensity scores for the nontreated cases are sparse) seems not to be the best response to the variance problems associated with local linear matching. On the basis of Frölich (2004), we tested various specifications of bandwidth and trimming strategies for simulated data and empirical data. We found that for empirical applications, methods need to be developed to seek the best bandwidth and to handle the variance problem induced by the common-support region. It appears important to test various bandwidth values and trimming schedules. As we mentioned earlier, it is prudent at this point to treat different specifications of bandwidth values and trimming schedules as sensitivity 344 analyses. Caution would be warranted when estimated treatment effects for the treated vary by bandwidth specifications (i.e., when the study findings are sensitive to bandwidth values and trimming schedules). The results of kernel and local linear estimations involve weighted average outcomes of the nontreated cases. Because the asymptotic distribution of weighted averages is relatively complicated to program, no software packages are currently available that offer parametric tests to discern whether a group difference is statistically significant. As a common practice, researchers use bootstrapping to estimate the standard error of the sample mean difference between treated and nontreated cases. However, Abadie et al. (2004) and Abadie and Imbens (2006) have warned that bootstrapping methods for estimating the variance of matching estimators do not necessarily give correct results. Thus, in practice, conducting a significance test of a treatment effect for the treated derived from kernel-based matching may be problematic, and researchers should be cautious when interpreting findings. 9.3 OVERVIEW OF THE STATA PROGRAMS PSMATCH 2 AND BOOTSTRAP To implement propensity score analysis with nonparametric regression, users can run a user-developed program psmatch2 (Leuven & Sianesi, 2003) and the regular program bootstrap in Stata. Section 5.7 of this book provides the information for downloading and installing psmatch2. The basic syntax to run psmatch2 for estimating the treatment effect for the treated with a kernel matching is as follows: psmatch2 depvar, kernelname(varlist) kerneltpe(kernel_type) /// pscore(varname) bwidth(real) common trim(integer) In this command, depvar is the variable indicating treatment status, with depvar = 1 for the treated and depvar = 0 for the control observations. The keyword kernel is used to request a kernel matching. The term outcome(varlist) specifies on which outcome variable(s) users want to assess the treatment effect for the treated; users specify the name of the outcome variable(s) in the parentheses. To run difference-in-differences or two-time-point data, users may create a change-score variable (i.e., a difference in the outcome variable between Time 2 and Time 1) before running psmatch2 and then specify the change-score variable in the parentheses. The term kerneltpe(kernel_type) specifies the type of kernel, and one of the following five values may be specified in the parentheses as a kernel type: epan—the Epanechnikov kernel, which is the default with the kernel matching; tricube—the tricube kernel, which is the default with the local linear regression matching; normal—the normal 345 (Gaussian) kernel; uniform—the rectangular kernel; and bweight—the bweight kernel. The term pscore(varname) specifies the variable to be used as propensity score; typically, this is the saved variable of propensity score users created by using logistic before running psmatch2. The term bwidth(real) specifies a real number indicating the bandwidth for kernel matching or local linear regression matching; the default bwidth value is 0.06 for the kernel matching. The term common imposes a common-support region that causes the program to drop treatment observations with a propensity score higher than a maximum or less than a minimum propensity score of the controls. The term trim(integer) causes the program to drop "integer %" of the treatment observations at which the propensity score density of the control observations is the lowest. That is, suppose the user attempts to trim 5% of the treated cases; then the specification is trim(5). The basic syntax to run psmatch2 for estimating a treatment effect for the treated with a local linear regression matching looks similar to the above syntax specifying kernel matching. The only change the user must make is to replace the keyword kernel with the keyword llr. Thus, the syntax for the local linear regression matching is as follows: psmatch2 depvar, llr outcome(varlist) kerneltpe(kernel_type) /// pscore(varname) bwidth(real) common trim(integer) Note that for local linear regression matching or llr, the default kernel type is tricube. The Stata program bootstrap can be used to run bootstrap sampling to obtain an estimation of the standard error and a 95% confidence interval for the estimated average treatment effect for the treated. The bootstrap program can also be invoked by an abbreviation bs. The only required syntax of bootstrap consists of the following two elements: the previous command, which runs psmatch2 for the kernel-based matching, and r(att), which indicates that we want to do a bootstrap sampling on the estimated average treatment effect for the treated att. Each of the preceding two elements should be enclosed in double quotes. The following syntax shows the running of psmatch2 to obtain an estimated treatment effect for the treated using local linear regression matching and then the running of bs to obtain the bootstrap estimation of standard error: psmatch2 aodserv, outcome(extern) pscore(logit) llr bs "psmatch2 aodserv, outcome(extern) pscore(logit) llr" r(att) Similar to running psmatch2 for nearest neighbor matching or Mahalanobis matching (see Section 5.7), it is important to create a random variable and then sort data by this variable before invoking psmatch2. To guarantee that the same results are obtained from session to session, users control for seed number by 346 using a set seed command. Table 9.1 exhibits the syntax and output for a typical analysis estimating a treatment effect for the treated with a local linear regression matching. The analysis consists of the following steps: (a) We first run a logistic regression to obtain the predicted probabilities for all observations by using logistic; (b) we then create a logit score and define the logit as the propensity score, create a difference score that is a difference of the outcome variable between two time points (the difference score will be specified as the outcome variable in the subsequent analysis, and by doing so, we are conducting a difference-in-differences analysis), and sort the sample data in a random order and set up a seed number to ensure that we will obtain the same results from session to session; (c) we run psmatch2 using the keyword llr to request a local linear regression matching (note that without specifying kerneltpe and bwidth, we use the default kernel of "tricube kernel" and a default bandwidth value of 0.8) and then run bs to obtain bootstrap estimation of the standard error and a 95% confidence interval; and (d) we run a similar analysis using a different bandwidth by specifying bw(.01) and a similar analysis to trim 5% of treated cases by specifying trim(5). 9.4 EXAMPLES We include two examples in this section to illustrate the application of kernel-based matching in program evaluation. The first example shows the application of local linear regression matching to the analysis of two-time-point data that uses a difference-in-differences estimator. In this example, we also illustrate how to specify different values of bandwidth and how to trim by using a varying schedule of trimming; then, combining these specifications together, we show how to conduct a sensitivity analysis. The second example illustrates the application of local linear regression matching to one-time-point data. We also compare the results of kernel-based matching with those produced by matching estimators (i.e., Abadie et al., 2004). Table 9.1 Exhibit of Stata psmatch2 and bs Syntax and Output Running Matching With Nonparametric Regression 347 348 349 350 Source: Data from NSCAW, 2004. 9.4.1 Analysis of Difference-in-Differences Here we use again the sample and matching variables from Section 4.4.1. This example shows the analysis of a subsample of data for 1,407 children obtained from the longitudinal data set of the National Survey of Child and Adolescent Well-Being (NSCAW). Of these study participants, 112 were children of caregivers who had received substance abuse treatment services (i.e., the treatment group), and the remaining 1,295 participants were children of caregivers who did not receive substance abuse treatment services (i.e., the comparison group). The research examined two study questions: At 18 months after their involvement with child protective services, how were the children of caregivers who received substance abuse treatment services faring? Did these children have more severe behavioral problems than their counterparts whose caregivers did not receive substance abuse treatment services? In Section 4.4.1, we used the Heckit treatment effect model to analyze onetime-point data. The psychological status of children at the 18th month after caregivers became involved with child protective services. In the current analysis, we use two-time-point data: The psychological status of children measured at the NSCAW baseline and the same variables measured at 18 months after the baseline. As such, this analysis permits a longitudinal inquiry of 351 a difference-in-differences, that is, a difference in the psychological status of children given a difference in the participation of caregivers in substance abuse treatment services. In addition to externalizing and internalizing function scores (measures of psychosocial status), the current analysis included an additional outcome variable: the total score of the Achenbach Children's Behavioral Checklist (CBCL/4–18; Achenbach, 1991). Based on caregiver ratings, the study uses three measures of child behavior as outcomes: externalizing scores, internalizing scores, and total scores. A high score on each of these measures indicates more severe behavioral problems. Based on the design of kernelbased matching, the current analysis is a one-to-many matching, and it is a comprehensive investigation of the treatment effect for the treated. In this case, treatment comprises child protective supervision and the participation of caregivers in a drug abuse intervention program. No specific services are provided to remediate child behavioral problems. The difference-in-differences estimator in the current analysis uses local linear regression to calculate the weighted average difference for the nontreatment group using a tricube kernel and a default bandwidth. On the basis of the literature on the finite-sample properties of local linear matching, we tested the sensitivity of findings to different specifications on bandwidth and trimming. While holding trimming constant, three bandwidth values were used: 0.01, 0.05, and 0.8. We also tested the sensitivity of findings to variations in the trimming level. We applied the following three trimming schedules (i.e., imposed a common support by dropping treated observations whose propensity scores are higher than the maximum or less than the minimum propensity score of the nontreated observations) while fixing the bandwidth at the default level: 2%, 5%, and 10%. Standard errors for the difference-in-differences estimates were obtained through bootstrapping. The standard error of estimated difference-in-differences was used further to estimate a 95% bootstrap confidence interval for the average treatment effect for the treated. We report the 95% confidence interval using the bias-correction method, so that a meaningful effect is reasonably unlikely to occur by chance as indicated by a 95% confidence interval that does not include a zero. Table 9.2 Estimated Average Treatment Effects for the Treated on CBCL Change: Difference-in-Differences Estimation by Local Linear Regression (Example 9.4.1) 352 Source: Data from NSCAW, 2004. Note: CBCL = Child Behavior Checklist; DID = differences-in-differences. a. The T tests show that two unadjusted mean differences are not statistically different. *The 95% confidence interval does not include a zero, or $p < .05$ for a two-tailed test. Table 9.2 shows the estimated average treatment effects for the treated group. Taking the externalizing score as an example, the data indicate that the mean externalizing score for the treatment group increased from baseline to the 18th month by 0.15 units, and the mean score for the nontreatment group decreased from baseline to the 18th month by 1.82 units. The unadjusted mean difference between groups is 1.97, meaning that the average change for the externalizing score for the treatment group is 1.97 units higher (or worse) than that for the nontreatment group. The difference-in-differences estimation further adjusts for the heterogeneity of service participation by taking into consideration the distance on propensity scores between a treated case and its nontreated matches in the calculation of the treatment effects for the treated. The point estimate of the difference-in-differences on externalizing is 3.37, which falls into a 95% bootstrap confidence interval bounded by 0.27 and 5.43. That is, we are 95% confident that a nonzero difference on externalizing between treated and 353 nontreated groups falls into this interval. The next significant difference on the adjusted mean is the total score. The point estimate of the difference-in-differences is 2.76, which falls into a 95% bootstrap confidence interval bounded by 0.96 and 5.12. The 95% bootstrap confidence interval of the difference-in-differences for the internalizing score contains a zero, and therefore we are uncertain whether such difference is statistically significant at a significance level of 0.05. Sensitivity analyses of different bandwidth specifications and different trimming strategies tend to confirm the results. That is, for the externalizing and total scores, all analyses (except the total CBCL score associated with the large bandwidth) show a 95% bootstrap confidence interval bounded by nonzero difference-in-differences estimates. Similarly, for the internalizing score, all analyses show a 95% bootstrap confidence interval that includes a zero difference-in-differences estimate. The study underscores the importance of analyzing change of behavioral problems between service and nonservice groups using a corrective procedure such as propensity score analysis with nonparametric regression. The unadjusted mean differences based on the observational data underestimate the differences of behavior problems between the two groups, and the underestimation is $3.37 - 1.97 = 1.4$ units for the externalizing score and $2.76 - 1.03 = 1.73$ for the total score. It is worth noting that the analysis used bootstrapping for significance testing—which may be problematic—and thus is a limitation of the study. The results of the study should be interpreted with caution. Conditionally, the kernel-based matching analysis of NSCAW observational data suggests that the combination of protective supervision and substance abuse treatment for caretakers involved in the child welfare system should not alone be expected to produce developmental benefits for children. Additional confirmatory analyses and discussion of these findings are available elsewhere (Barth, Gibbons, & Guo, 2006). 9.4.2 Application of Kernel-Based Matching to One-Point Data In this application, we use the same sample and variables presented in Section 8.4.1. The only difference is that Section 8.4.1 used the matching estimators, whereas the current analysis uses kernel-based matching. We have included the current illustration for two reasons: (1) We want to show that kernel-based matching can also be applied to one-point data for which the treatment effect for the treated is not a difference-in-differences, and (2) we want to compare kernel-based matching with the matching estimators. With regard to the first purpose, we show in our syntax (available on the companion webpage of this book) that the analysis of one-point data is, perhaps, more intuitive than the analysis of difference-in-differences. That is, with onepoint data, you can specify the outcome variable directly in the psmatch2 354 statement, whereas with two-point data, you must first create a change-score variable using data from two time points and then specify the change-score variable as the outcome in the psmatch2 statement. Recall from the example presented in Section 8.4.1 that the research objective for this study was to test a hypothesis pertaining to the causal effect of childhood poverty (i.e., children's use of the Aid to Families With Dependent Children [AFDC] welfare

program) on children's academic achievement. The study examined one domain of academic achievement: the age-normed "passage comprehension" score in 1997 (i.e., one time point) of the Woodcock-Johnson Revised Tests of Achievement (Hofferth et al., 2001). A higher score on this academic achievement measure indicated higher achievement. The study sample consisted of 606 children, of whom 188 had used AFDC at some time in their lives (i.e., ever used group) and who were considered the treated cases. The remaining 418 children had never received AFDC (i.e., never used group) and were considered the comparison cases. This study used the following six covariates: (1) current income or poverty status, measured as the ratio of family income to poverty threshold in 1996; (2) caregiver's education in 1997, which was measured as years of schooling; (3) caregiver's history of using welfare, which was measured as the number of years (i.e., a continuous variable) a caregiver participated in the AFDC program during his or her childhood ages of 6 to 12 years old; (4) child's race, which was measured as African American versus non-African American; (5) child's age in 1997; and (6) child's gender, which was measured as male versus female. Note that in Section 8.4.1, these covariates were used as matching variables in a vector-norm approach, but in the current analysis, they are used in the estimation of a propensity score for each study child. In other words, the six covariates were included in a logistic regression, and then the propensity score was matched by the local linear regression. Table 9.3 presents results of the study. Kernel-based matching estimated an average treatment effect for the treated of -4.85 , meaning that children who used the AFDC welfare program during childhood score 4.85 units lower on the passage comprehension measure than those who did not use AFDC during childhood, after controlling for observed covariates. This effect was statistically significant at a .05 level. Note that the matching estimator produced a similar finding: The estimated treatment effect for the treated group was -5.23 , and the effect was statistically significant at a .05 level. Although the matching estimator's estimate of the effect was slightly larger, both estimators find the effect statistically significant and thus lead to a consistent conclusion with regard to testing the research hypothesis. This study implies that (a) at least for this data set, both matching and kernel-matching estimators produce the same substantive findings and appear to be just about equally useful and (b) because the same issue was examined with different statistical methods and the findings converge, the substantive conclusion of the study is more convincing than that 355 produced by either method alone. The application underscores an important point for observational studies: Researchers should compare estimates across multiple methods. Table 9.3 Estimated Treatment Effect for the Treated (Child's Use of AFDC) on Passage Comprehension Standard Score in 1997: Comparing the Propensity Score Analysis With Nonparametric Regression With Bias-Corrected Matching and Robust Variance Estimator (Example 9.4.2) Source: Data from Hofferth et al., 2001. Note: 95% CI = 95% confidence interval; NC = not comparable due to bootstrap; SATT = sample average treatment effect for the treated. 9.5 CONCLUSION This chapter described the application of local linear regression matching with a tricube kernel to the evaluation of average treatment effects for the treated. Kernel-based matching was developed to overcome perceived limitations within the Rosenbaum and Rubin (1983) counterfactual framework. Heckman and colleagues (1997, 1998) made important contributions to the field: (a) Unlike traditional matching, kernel-based matching uses propensity scores differentially to calculate a weighted mean of counterfactuals, which is a creative way to use information from all controls; (b) applying kernel-based matching to two-time-point data, the difference-in-differences estimator permits analysis of treatment effects for the treated in a pretest/posttest trial fashion; and (c) by doing so, the estimator is more robust in terms of handling measurement errors. It eliminates temporarily invariant sources of bias that may arise, when program participants and nonparticipants are geographically mismatched or respond in systematically biased ways to survey questionnaires. Kernel-based matching was developed also on the basis of a rigorously proven distribution theory. However, as far as we know, the distributional theory developed by Heckman and his colleagues is not incorporated into 356 computing programs for estimating the standard errors of the estimated treatment effect for the treated. The use of bootstrapping in statistical inference is a limitation, and results based on bootstrapping should be interpreted with caution. NOTES 1. In this chapter Heckman et al. (1998) refers to Heckman, Ichimura, and Todd (1998), and Heckman et al. (1997) refers to Heckman, Ichimura, and Todd (1997). 2. In practice, T and Z may or may not be composed of the same variables. 3. Smith and Todd (2005) describe a more sophisticated procedure to determine the proportion of participants to be trimmed. They suggest a method for determining the density cutoff trimming level. 4. We are grateful to John Fox for providing this example and the R code to produce the figures for the example. 357 CHAPTER 10 Propensity Score Analysis of Categorical or Continuous Treatments: Dosage Analyses This chapter describes propensity score procedures for estimating effects when treatment has multiple categories or is continuous. The development of these approaches stems from the same propensity score framework developed by Rosenbaum and Rubin (1983), but these procedures extend consideration from a binary treatment condition to multiple treatments, where levels of an intervention or program may be categorical or even continuous in nature. Sometimes known as modeling treatment dosage, this topic is widely studied by medical and health researchers in the context of determining the effects of differential amounts of treatment on outcomes. In the social sciences, there is a growing interest in studying treatments that take on a continuum of values. Indeed, many research questions can be conceptualized as efforts to discern an optimal dose of a treatment, rather than whether the treatment in a full presentation is effective. As in observational studies of a binary treatment, researchers need to consider a corrective procedure such as a propensity score analysis, if and only if they suspect there is a selection process involved in determining who uses what amount of treatment. That is, a corrective procedure is needed when treatment exposure is not random. In a typical drug trial, randomization guarantees that assigning participants to treatment levels is not confounded with participant characteristics. Hence a dosage effect can be straightforwardly modeled by a continuous independent variable, or a set of dichotomous variables, via a multivariate outcome model, such as a regression. Applying the potential outcome framework (i.e., the counterfactual framework) to the multilevel-treatments case, the correction of endogeneity becomes more complicated, because in this setting, there is more than one potential outcome—if treatment has five levels, for each given treatment, researchers have an observed outcome linking to the received treatment plus four other potential outcomes. Hence, identification of treatment effects has to rely on additional assumptions, and the analytical procedure becomes more complicated. Section 10.1 is an overview of the development of modeling categorical or 358 continuous treatments using propensity scores. Sections 10.2 to 10.4 describe three different methods. Of the three methods, the generalized propensity score (GPS) method, depicted by Section 10.4, is highly recommended, because with the available computing software program, implementing the method is straightforward. Section 10.5 is a technical description of running the Stata gpscore program. Section 10.6 shows examples of empirical applications. Section 10.7 presents a summary and a conclusion. 10.1 OVERVIEW All methods described thus far concern a binary treatment condition, that is, a treated group compared with a control group. However, in practice, more than two conditions may be compared. This often happens when we want to estimate the impact of treatment dosage. For example, in clinical trials, one may be interested in estimating the dose-response function where the drug dose takes on a continuum of values (Efron & Feldman, 1991). In a school-based study, a researcher may have accurately recorded the number of minutes of student exposure to an intervention, and therefore, in addition to receiving or not receiving treatment (i.e., a zero minute of dosage), it may be informative to test the hypothesis that an increase in exposure to treatment results in improved outcomes (e.g., Section 8.4.2 in Chapter 8). In economics, researchers are interested in the effect of schooling on wages, where schooling is measured as years of education (Card, 1995b; Imai & Van Dyk, 2004); the effect of aid to firms, where aid is measured on a continuum of monetary values (Bia & Mattei, 2007); or the effect of the amount of a lottery prize on subsequent labor earnings (Imbens, Rubin, & Sacerdote, 2001). In political science, one may be interested in the combined effects of different voter mobilization strategies, such as phone calls and door-to-door visits (Gerber & Green, 2000). Other examples of treatment dosage include the length of exposure to intervention-related advertisements, such as the frequencies of seeing antidrug commercials on TV, hearing them on radio, or reading antidrug ads in newspapers or magazines (Lu, Zanotto, Hornik, & Rosenbaum, 2001); the effect of smoking—precisely the frequency and duration of smoking—on medical expenditures (Imai & Van Dyk, 2004); the effect of differing food stamp subsidies on food insecurity (GibsonDavis & Foster, 2006); the number of times using services or visiting doctors, or the number of treatment sessions attended (Foster, 2003; Kluge, Schneider, Uhlendorff, & Zhao, 2012); and the like. The analysis of treatment dosage with propensity scores may be generalized in two directions. In the first direction, one estimates a single scalar propensity score using ordered logistic regression, matches on the scalar propensity score, and proceeds as one might in a two-treatment-group situation (Joffe & Rosenbaum, 1999). In the second direction, one estimates propensity scores for each level of treatment dosage (i.e., if there are five treatment conditions 359 defined by differential doses, one estimates five propensity scores for each participant). Propensity scores derived in this fashion are called generalized propensity scores (GPS). This method, originally developed by Imbens (2000), uses the inverse of a particular estimated propensity score as a sampling weight to conduct a multivariate analysis of outcomes. It shares common features with propensity score weighting. Hirano and Imbens (2004) extended the GPS approach in three steps: estimating GPS using a maximum likelihood regression, estimating the conditional expectation of the outcome given the treatment and GPS, and estimating the dose-response function to discern treatment effects as well as their 95% confidence bands. The single scalar and two GPS procedures are described in the following. 10.2 MODELING DOSES WITH A SINGLE SCALAR BALANCING SCORE ESTIMATED BY AN ORDERED LOGISTIC REGRESSION The first dose-analysis method applies propensity score matching to a continuous-treatment scenario. This constitutes an extension of propensity score matching under a binary condition. The method was originally proposed by Joffe and Rosenbaum (1999). Lu et al. (2001) review details of the method with illustrations. When moving from binary to multiple treatments, matching essentially requires the creation of matched pairs in such a way that high- and low-dose groups have similar or balanced distributions of observed covariates. However, balancing covariates with propensity scores under the condition of multiple doses raises three considerations or, perhaps, complications (Lu et al., 2001). First, under the multiple-doses condition, the original definition of a propensity score (i.e., it is a conditional probability of receiving treatment, or a single scalar score, given observed covariates) is no longer applicable. Because there are multiple doses, each participant now can have multiple propensity scores, and each score compares one dose with another. Indeed, the second method of modeling multiple doses defines propensity scores in this way and estimates multiple propensity scores. Joffe and Rosenbaum's (1999) method uses a single scalar balancing score and shows that such a score exists only for certain models. These include McCullagh's (1980) ordered logistic regression and a Gaussian multiple linear regression model with errors of constant variance. The key application issue is the choice of a statistical model that estimates the propensity score, and the recommended model is the ordered logistic regression (Joffe & Rosenbaum, 1999; Lu et al., 2001). Using an ordinary least squares (OLS) regression to estimate the propensity score is problematic, because such a model assumes errors of constant variance, but in practice, error variances often vary by levels of treatment dosage, and 360 heteroscedasticity is likely to be present. Second, under the multiple-doses condition, one needs to redefine the distance between a treated case and a control case in optimization of matching. In this situation, the goal is to identify pairs that are similar in terms of observed covariates but very different in terms of dosage. Hence, the distance must measure both the similarity in terms of covariates and the differences in terms of doses. And finally, under the multiple-doses condition, the matching algorithm employed is also different from that employed in the binary condition. In the network flow literature of operations research, matching one group to another disjoint group (i.e., under the binary condition) is called a bipartite matching problem. But matching under the condition of multiple doses is considered matching within a single group and is called nonbipartite matching. Because of this difference, the optimization algorithms one uses in the binary condition, such as the algorithms of optimal matching described in Section 5.4.2 in Chapter 5, are no longer applicable. These three complexities have led to the development of new features of matching with multiple doses. The matching procedure using a single scalar score is summarized next. Step 1: Develop a single scalar score based on an ordered logistic regression. We first run an ordered logistic regression in the form of McCullagh (1980). The distribution of doses Z_k for a sample of K participants ($k = 1, 2, \dots, K$), given observed covariates x_k , is modeled as assuming there are five treatment doses being modeled. This model compares the probability of a response greater than or equal to a given category ($d = 2, \dots, 5$) to the probability of a response less than this category, and the model is composed of $d - 1$ parallel linear equations. In the linear part of the model, θ_d is called a "cutoff" value to be used in calculating predicted probabilities of falling into each of the five responses, with the probability of the omitted response equalling 1 minus the cumulative probability of falling into all of the other four categories. Note that there are four θ_d values for five ordered responses and five model-based predicted probabilities for each participant k . None of these quantities is a single scalar. The crucial feature of Joffe and Rosenbaum's (1999) model is that it defines as the estimated propensity score, or $e(x_k) =$, because the distribution of doses given covariates depends on the observed covariates only through x_k and the observed covariates x and the doses Z are conditionally independent given the scalar U under this setup, 361 $e(x_k) =$ is a single balancing score, and the maximum likelihood estimate is used after running the ordered logistic regression in the matching. Step 2: Calculate distance between participants k and k' , where $k \neq k'$. Recall that matching is an optimization problem in which we wish to minimize the sample total distances between treated and control participants. This goal remains the same for the multiple-doses condition, but the distance formula is revised. Lu et al. (2001) provide the following equation to calculate the distance under the multiple-doses condition: where and are the estimated propensity scores, and Z_k and $Z_{k'}$ are dose values ($= 1, 2, \dots, d$, if there are d doses) for k and k' , respectively. The main aspect of Equation 10.1 is e , which is a vanishingly small but strictly positive number ($\epsilon > 0$). The constant e is a formal device signifying how perfect matches on covariates or doses will be handled. It is e that makes the calculation of distances for the multiple-doses condition differ from that for the binary condition. Thus, ϵ serves two functions. It specifies that (1) if participants k and k' have the same dose, $Z_k = Z_{k'}$, then the distance between them is ∞ even if they have identical observed covariates $x_k = x_{k'}$, and the distance between them is 0, and (2) when two participants have identical observed covariates $x_k = x_{k'}$ and the distance between them is 0, the dose will be smaller as the difference in dose ($Z_k - Z_{k'}$) increases. distance Step 3: Conduct nonbipartite pair matching using the distances so defined. For a sample of K participants, each participant k has a distance from each of the remaining participants in the sample on the estimated propensity scores. The researcher then conducts an optimal pair matching in such a way that the total distance associated with all matched pairs is minimized. Each of the resultant pairs then contains one high-dose participant and one low-dose participant, which is forbidden by Equation 10.1. It is worth because noting that the optimal matching under the current context is the so-called nonbipartite matching, which is different from the bipartite optimal matching described in Section 5.4.2 in Chapter 5. Therefore, one has to use special software programs to conduct the matching. Specifically, the R program optmatch developed by Hansen (2007) conducts bipartite matching and, therefore, should not be used in the current context. Lu et al. (2001) used a revised algorithm based on Derigs's (1988) program. An alternative strategy is to use SAS Proc Assign described in Ming and Rosenbaum (2001). 362 Step 4: Check covariate balance after matching. Having obtained matched pairs, the next step involves checking covariate balance between high- and low-dose participants to see how well the propensity score matching performed, that is, whether high- and low-dose participants are comparable in terms of observed covariates. The balance check is straightforward; that is, you may calculate mean differences and conduct independent sample t tests between the high- and low-dose participants on each observed covariate. One hopes that at this stage, all t tests would show nonsignificant differences between the high- and low-dose participants. If significant differences remain, you may return to the ordered logistic regression and matching steps to change specifications and rerun the previous analyses. Step 5: Evaluate the impact of treatment doses on the outcome. At this final stage, the impact of treatment doses on outcome differences is estimated. Because multiple doses are modeled, it is possible not only to evaluate treatment effectiveness but also to assess the degree to which dosage plays a role in affecting outcome differences. The outcome evaluation pools together all pairs showing the same contrasts of high and low doses, calculates mean difference on the outcome variable between the high- and low-dose participants, and conducts a Wilcoxon signed rank test to evaluate whether the difference is statistically significant. For an illustration, see Lu et al. (2001). 10.3 MODELING DOSES WITH MULTIPLE BALANCING SCORES ESTIMATED BY A MULTINOMIAL LOGIT MODEL Imbens (2000) proposed estimating multiple balancing scores by using a multinomial logit model and then conducting an outcome analysis that employs the inverse of a specific propensity score as a sampling weight. Compared to single-scalar balancing, this method requires fewer assumptions and is easier to implement. Moreover, Imbens's method can be used with several unordered treatments. This approach has two steps. Step 1: Estimate generalized propensity scores (GPS) by using a multinomial logit model. Imbens (2000) first defines the conditional probability of receiving a particular dose of treatment given the observed covariates as the GPS, which can be estimated by the multinomial logit model. Suppose there are d treatment doses; then each participant will have d generalized propensity scores, and in this context, propensity scores are no longer a scalar function. Step 2: Conduct outcome analyses by following the process of propensity score weighting. Next, the researcher calculates the inverse of a specific GPS 363 and defines the inverted propensity score as a sampling weight to be used in outcome analysis (i.e., analysis with propensity score weighting). Denoting $e(x_k, d) = P(D = d | X = x)$ as the generalized propensity score of receiving treatment dose d for participant k with observed covariates x , then the inverse of the GPS (i.e., $1/e(x_k, d)$) is defined as a sampling weight for participant k . Note that even though each participant has multiple propensity scores obtained from the multinomial logit model, only one such score is used and defined in the propensity score weighting analysis: It is the predicted probability for participant k to fall into the d dose category that is used, and the inverse of this score is defined as the weight in the outcome analysis. After creating the weight, we simply use it in multivariate analyses evaluating outcome differences. Most software packages allow users to specify the name of a weight variable in procedures of multivariate analysis. The analysis then is analogous to multivariate modeling that incorporates sampling weights, as described in Chapter 7. In the outcome analysis, a set of $d - 1$ dummy variables is created, with one dose category omitted as a reference group. These dummy variables are specified as predictor variables in the outcome analysis. They signify the impact of doses on the outcome variable. The p value associated with the coefficient for each dummy variable indicates the dosewise statistical significance and can be used in hypothesis testing. An example illustrating Imbens's method is presented in Section 10.6.1. For estimating propensity scores using generalized boosted models (GBM) in the setting of multiple treatments, readers are referred to McCaffrey et al. (2013). 10.4 THE GENERALIZED PROPENSITY SCORE ESTIMATOR Building on Imbens's (2000) estimator using multinomial logit model, Hirano and Imbens (2004) developed a generalization of the binary treatment propensity score and labeled the method a GPS estimator. Bia and Mattei (2008) developed a software program in Stata called gpscore (i.e., the programs of gpscore.ado, doseresponse_model.ado, and doseresponse.ado) to implement the GPS estimator. Following Hirano and Imbens (2004) and Bia and Mattei (2008), we describe the basic ideas of the GPS estimator below. We have a random sample of size N , indexed by $i = 1, 2, \dots, N$. For each unit i , we observe a $p \times 1$ vector of pretreatment covariates, X_i ; the treatment received, T_i ; and the value of the outcome variable associated with this treatment, Y_i . Using the counterfactual framework, unit i has a set of potential $Y_i = 1, \dots, N$, where Γ is a continuous set of outcomes, defined as potential treatment values. Under this definition, is the unit-level dose364 response function, and a dosage analysis is interested in the average doseAccording to Hirano and Imbens (2004), Y_i response function T_i , and X_i , $i = 1, \dots, N$, are defined on a common probability space; T_i is continuously distributed with respect to the Lebesgue measure on Γ and $Y_i = Y_i(T_i)$ is a well-defined random variable. To simplify the notation, we drop the i subscript in the following description. With the preceding notation, the GPS is defined as the conditional density of the treatment T given the covariates, or the GPS is $R = r(T, X)$, where $r(T, X) = P(T = t | X = x)$. The GPS has a balancing property similar to the propensity score under the setting of binary treatment. Hirano and Imbens (2004) proved two theorems with respect to the balancing property of GPS: (1) weak unconfoundedness given the GPS—suppose that assignment to the treatment is weakly unconfounded, given X (i.e., $Y(1) \perp T | X$, for all $T \in \Gamma$); then, for every t , the theorem implies that the GPS can be used to eliminate any bias associated with differences in the observed covariates X . (2) Bias removal with GPS—suppose that assignment to the treatment is weakly and unconfounded, given X ; then, this last equation shows the formula for evaluating the average outcome at the treatment level t , given observed covariates X . Hirano and Imbens (2004) refer to their assumption as weak unconfoundedness, because they do not require joint independence for all t that is, instead, they require conditional independence to potential outcomes, hold for each value of the treatment. With the above development, researchers can execute the GPS estimator by following three basic steps depicted next. Step 1: Modeling the conditional distribution of the treatment given covariates. Practically, this is the step in which researchers estimate the GPS at a given level of treatment and observed covariates X and then perform balance check. The balance check used by Hirano and Imbens (2004) is a special type—that is, it employs the idea of subclassification. However, in this circumstance, it evaluates the covariate difference, one at a time for each element in the vector X , between two user-defined groups within a subclass, rather than estimating the outcome difference between treated and control groups as one normally does in estimating treatment effects with subclassification. Because this procedure is important and complicated, we present the algorithm in a technical fashion first, and then we explain the procedure following closely the Hirano and Imbens (2004) example. Hirano and Imbens (2004) used a flexible parametric approach to estimate the GPS. That is, they assume a normal distribution for the treatment, given the covariates: 365 where $g(T_i)$ is a suitable transformation of the treatment variable. While the parametric model assumes a normal distribution, the actual distribution of treatment dosage T_i in the sample may not be normally distributed. To correct for nonnormality, one can do a transformation of T_i or by applying other transformations. Using a maximum likelihood regression using $g(T_i)$ as the dependent variable and the covariates X_i as the independent variables. This is the baseline model for estimating GPS. The formula for estimating GPS for each observation based on the estimated regression model is Next, users need to test the balancing property and show whether the GPS balances observed covariates. The balance check involves six steps. We use the notation defined by the developers of the software program Stata gpscore (Bia & Mattei, 2008, pp. 357–358). 1a. Divide the set of potential treatment values, Γ into K intervals based on a user-specified rule. Typically one uses the sample distribution of the treatment variable to decide how many intervals to use and what cutoff values should be. One may choose a K that makes each interval of G_1, \dots, G_K to have approximately equal size or based on another rule that makes sense. 1b. Within each treatment interval, using $k = 1, \dots, K$, compute the GPS at a userspecified representative point such as a median of the treatment variable. We denote this treatment level as tG_k . In other words, we compute the GPS ($r(T_k, X_i)$ for unit i at $tG_k \in G_k$. 1c. For each k , $k = 1, \dots, K$, block on the scores ($r(TG_k, X_i)$ using m intervals. This is similar to the process of subclassification, but the purpose here is to calculate balance (i.e., mean difference or other statistic) on each covariate of X . Quintiles of $r(TG_k, X_i)$ may be used to divide the sample denote the m GPS into five groups, or $m = 5$ intervals. Let interval for the k th treatment interval, G_k . 1d. Within each interval of calculate the mean difference of each and covariate between units that belong to the treatment interval, but belong to another units that are in the same GPS interval, treatment interval. 366 1e. Combine the m differences in means, calculated in Step 1d, by using a weighted average, with weights given by the number of observations in this is analogous to the calculation of an each GPS interval aggregated sample statistic in subclassification, and the formula is similar to Equation 6.1, except that one replaces the mean outcome variable Y with the mean of a covariate X , and the two groups being compared in the and current setting are those belonging to the treatment interval those belonging to another treatment interval 1f. For each G_k , $k = 1, \dots, K$, use test statistics (e.g., the Student's t statistics or the Bayes factors) to check the balance. One hopes to have a t statistic below 1.645 or a Bayes factor below 1.00 from the balance check, under which condition the test statistics indicate very slight evidence against the balancing property. The preceding process combines both the estimation of GPS and the balance check into one step, playing a crucial role in the application of the Hirano and Imbens estimator. To illustrate these steps, we now use the example originally presented by Hirano and Imbens (2004). The study uses data from the survey of Massachusetts lottery winners. In this study, researchers were interested in estimating the effect of the prize amount (i.e., the treatment dosage) on subsequent labor earnings (i.e., the outcome variable). A propensity score approach is well suited to this study, because the study sample involved selection. The study sample comprises 237 individuals who won a major prize in the lottery and who provided nonmissing data in the survey. The original sample was much larger, and many individuals (about 50% of the original sample) did not respond to the survey. The selection of nonresponses was obviously nonrandom, and thus, the analysis of the 237 individuals with nonmissing data warranted a correction. The study aimed to evaluate the treatment levels (i.e., the amount of lottery prize) on subsequent labor force earnings. The study used 13 covariates. Table 10.1 presents these covariates and results of the balance check. To conduct a GPS analysis, we first check the normality of the treatment dosage, that is, check the conditional distribution of the lottery prize variable given covariates. The distribution of the prize, as expected, was highly skewed. Following Hirano and Imbens (2004), we choose a logarithm transformation of the treatment dosage, or specify the g function of as a natural logarithm. Next, we run the maximum likelihood regression for all sample observations. In this regression, $\ln(\text{prize})$ is the dependent variable, and all 13 covariates are the independent variables. The estimated regression coefficients from this model will be used to estimate GPS. Table 10.1 Hirano and Imbens's Example: Balance Given the 367 Generalized Propensity Score—1 Statistics for Equality of Means Source: Hirano and Imbens, 2004, Table 7.2, p. 81. Next, we need to estimate GPS $r(t, x)$. A crucial concern here is at which t , or which treatment level, we want to estimate the GPS. Note that in the continuous-treatment circumstance, there are many levels of t or $\ln(\text{prize})$; hence, there are possibly many GPSS, each associated with a specific treatment value. Using the potential-outcome framework, each individual has an observed outcome associated with $\ln(\text{prize})$ given covariates X , plus potential outcomes related to all other $\ln(\text{prize})$ values in the sample. The actual number of GPSSs this estimator evaluates, as a matter of fact, is a few GPS scores associated with user-specified treatment levels, or a few sets of scores associated with userspecified $\ln(T)$. Hirano and Imbens (2004) suggested that one first divides the entire sample into K subclasses, following Step 1a. After taking the logarithm, the $\ln(\text{prize})$ or $\ln(T)$ ranges from 0 to 485. Hirano and Imbens chose $K = 3$ and divided the entire sample into three treatment intervals as $[0, 23]$, $[23, 80]$, and $[80, 485]$. As such, the first group has 79 observations, the second has 106, and the third has 52. There is no specific guidance on how to choose cutoff points and the number of K in this process. In practice, it probably makes sense to divide the entire sample into a few equally sized groups. After creating these three groups, we then choose the median $\ln(T)$ for each group to estimate three sets of GPSSs. For instance, the median $\ln(T)$ for the first group of $[0, 23]$ is 14. We then estimate the GPS at a treatment value of 14 by applying the modelestimated coefficients and Equation 10.2 for all 237 individuals. Repeating this 368 process for the other two groups (i.e., for the groups of $[23, 80]$ and $[80, 485]$) based on the median values of $\ln(\text{prize})$ for these two groups, we obtain another two sets of estimated GPSSs for all 237 individuals. Now we can use these three sets of GPSs to test whether the GPS estimator balances covariates. Taking the first group $[0, 23]$ and the covariate age as an example, we now illustrate the process of balance checking by following Steps 1c through 1f. For this test, we use the estimated GPS at the treatment level of 14 (i.e., $r(14, X_i)$) for all 237 observations. Using a modified subclassification strategy, we use quintiles of the estimated propensity scores to divide the entire sample into five groups—that is, we choose $m = 5$, as indicated by Step 1c. The five groups are bounded by the estimated GPS $r(14, X_i)$ as follows: $[.06, .21]$, $[.21, .28]$, $[.28, .34]$, $[.34, .39]$, and $[.39, .45]$. The number of observations in each group is 84, 39, 53, 36, and 25, respectively. Within the first group $[.06, .21]$, 16 belong to the "corrected group" $[0, 23]$, or the group from which we used its median of 14 to calculate the GPS, and 68 belong to the other two groups (i.e., belong to $[23, 80]$ and $[80, 485]$). We then calculate the mean value of age for the 16 observations that belong to $[0, 23]$ and fall into the interval $[.06, .21]$, as well as the mean value of age for the 68 observations that do not belong to $[0, 23]$ but fall into the same interval $[.06, .21]$; we compute the difference of the two means on age and its associated standard error (SE). Repeating this process for all other four groups (i.e., for groups $[.21, .28]$, $[.28, .34]$, $[.34, .39]$, and $[.39, .45]$), we finally obtain five mean differences of age and their associated SEs. Using the proportion of each group in the entire sample as a weight, we aggregate the five mean differences and their SEs to obtain a sample mean difference and SE (i.e., using a procedure similar to subclassification described in Chapter 6 and formulas similar to Equations 6.1 and 6.2), we obtain a t statistic on age for group $[0, 23]$, adjusted for GPS, as 0.1 (i.e., the highlighted number in Table 10.1). The t statistic on age for unadjusted group $[0, 23]$, as shown in Table 10.1, is -1.7 . This value is obtained simply by taking an independent samples t test based on the mean difference of age between the 79 observations in group $[0, 23]$ and the 237 $- 79 = 158$ observations in groups $[23, 80]$ and $[80, 485]$. Comparing the value of the t statistic adjusted for GPS ($t = 0.1$) with the t statistic unadjusted for GPS ($t = -1.7$), we see that the GPS algorithm greatly improves balance on age for the group $[0, 23]$. Repeating the preceding process for all other covariates and all three groups, we obtain all t statistics of Table 10.1. Results of the balance check are good, indicating that the GPS algorithm improves covariate balances. Of the 39 t statistics (i.e., 13 covariates \times 3 groups), after adjustment of GPS, only 2 are larger than 1.96 (i.e., $t = 2.1$ for "Earnings Year-5" and "Earnings Year-6" for the group $[80, 485]$), compared to 16 prior to adjustment. The t statistics adjusted for GPS are automatically reported by the Stata 369 program gpscore. Alternatively, users can request Bayes factors to check balance. The cutoff values indicating levels of balance are summarized by Bia and Mattei (2008) and shown in Table 10.2. Step 2: Estimating the conditional expectation of the outcome given the treatment and GPS. In this step, the conditional expectation of the outcome, Y_i , given T_i and R_i , is modeled as a flexible function of its two arguments. In practice, one can use a quadratic approximation or polynomial approximations of order not higher than three. The quadratic approximation is According to Hirano and Imbens (2004), there is no direct meaning to the estimated coefficient in this model, except that testing whether all coefficients involving the GPS are equal to zero can be interpreted as a test of whether the covariates introduce any bias. Table 10.2 "Order of Magnitude" Interpretations of the Test Statistics Source: Bia and Mattei, The Stata Journal, 2008, p. 363. a. The order of magnitude interpretations of the Bayes factor we applied were proposed by Jeffreys (1961). Using the same lottery prize data of 237 observations, Hirano and Imbens (2004) regressed the outcome, earnings 6 years after winning the lottery, on the prize T_i and the logarithm of the score R_i , using all second-order moments of prize and log score. The estimated coefficients are presented below, with the SE for each coefficient shown in the parentheses: The F statistic for the overall model is significant at the .05 level, and therefore, the model indicates that the covariates introduce bias. 370 Step 3: Estimating the dose-response function to discern treatment effects as well as their 95% confidence bands. Given the estimated parameters in Step 2, researchers now can estimate the average potential outcome at treatment level t , which is also known as the dose-response function. This is the final statistic that shows the outcome differences associated with treatment dosage. The doseresponse function is given by At this stage, one can estimate the dose-response functions for user-selected treatment values and use bootstrapping to form standard errors and confidence intervals. The Stata doseresponse program can draw two types of plots to show the final results: One plot shows the dose-response function, and the other plot shows the estimated treatment effect function (also known as estimated derivatives). The final results of the lottery prize example are shown in Figure 10.1. Note that in plotting the dose-response function or treatment effects, the treatment values shown on the x -axis use the original scale of the treatment variable—that is, they represent $\$10,000$, $\$20,000$, . . . of the lottery prize, rather than values on the scale of $\ln(\text{prize})$. Taking the dose-response function as an example, the results show that there is a sharp decline in earnings 6 years after winning the lottery, when the prize ranges from $\$10,000$ to $\$20,000$; the earnings tend to be constant when the prize ranges from $\$20,000$ to $\$50,000$; after the prize level of $\$50,000$, the earnings decrease as prize levels increase. The treatment effect function shows estimated derivatives, which in economic terminology show the marginal propensity to earn out of unearned income. According to Hirano and Imbens (2004), the yearly prize money is viewed as unearned income, and the derivative of average labor income with respect to this is the marginal propensity to earn out of unearned income. For a precise interpretation of the treatment effect function, we refer to Hirano and Imbens (2004). Together, the study finds "that those with low (lottery) earnings are much more sensitive to income changes than those with higher (lottery) earnings" (Hirano & Imbens, 2004, p. 83). 10.5 OVERVIEW OF THE STATA GPSCORE PROGRAM The Stata package developed by Bia and Mattei (2008) consists of three programs: gpscore.ado, doseresponse_model.ado, and doseresponse.ado. Use the key word gpscore to search from the Internet and then download the package (i.e., in Stata, one may use findit gpscore and then follow the online instructions to download and install the program). When running the program, users first need to identify a transformation of the treatment variable to satisfy the normality assumption by using the program gpscore.ado and then provide 371 exactly this transformation and specification of the GPS model in the input to the p r o g r a m doseresponse.ado. The commands and options used in doseresponse.ado also work for the other two programs. Table 10.3 exhibits the syntax and output running doseresponse using the example of a lottery prize. We now explain important options and specifications of running doseresponse. Users first specify cutoff values that divide the sample into userdefined groups (Step 1a). This is operationalized by creating a new variable cut, and it should be specified in the syntax as cutpoints(cut). The transformation of the treatment variable to make it close to a normal distribution is specified by $t_transf(\ln)$, which, in this example, requests a natural logarithm transformation. Users need to specify which statistic of the treatment variable they choose to estimate GPS scores (Step 1b), and they do so in index(p50), which, in the current example, requests the median or the 50th percentile of the treatment variable as the point at which one estimates GPS for all units. For balance checks, users need to inform the program how many groups to use (Step 1c) by specifying $nq_gps(5)$, in this case using five groups. To specify the conditional expectation of the outcome given the treatment and GPS, users request polynomial approximations of order not higher than three— $reg_type_t(\text{quadratic})$ $reg_type_gps(\text{quadratic})$ $interaction(1)$ are the specifications to request the quadratic approximation. bootstrap(yses) boot_reps(100) requests bootstrap for the calculation of 95% confidence bands and uses 100 bootstrap replications. To draw the graph showing the doseresponse function, users first define a vector of the treatment levels using matrix define and then specify the vector name tp in the points option, as points(tp). Figure 10.1 Estimated Dose-Response Function, Estimated Derivative, and 95% Confidence Bands 372 Source: Data from Bia and Mattei, 2008. The output first presents the summary statistics of the transformed treatment variable. The regression model using a maximum likelihood estimator is shown next. This is the model the algorithm uses to estimate the GPS subsequently. Results of the test of normality of the transformed treatment variable are presented next; by default, the program reports the test using the Kolmogorov-Smirnov method. The output then reports the results of the balance check on all covariates for each of the user-defined groups—these are the results shown in Table 10.1. The output then presents the regression model estimating the conditional expectation of the outcome given the treatment and GPS. Finally, the program produces two graphs showing the dose-response function and treatment effect function with 95% confidence bands. Table 10.3 Exhibit of Stata doesresponse Syntax and Output Running GPS Estimator 373 374 375 376 377 Source: Data from Bia and Mattei, 2008. 10.6 EXAMPLES 10.6.1 Modeling Doses of Treatment With Multiple Balancing Scores Estimated by a Multinomial Logit Model In this example, we illustrate the analysis modeling multiple doses of treatment with

multiple balancing scores where the propensity scores are estimated using a multinomial logit model. With one exception, the illustration employs the same data and research questions as those for Section 5.8.2 in Chapter 5. In the analysis of Section 5.8.2, the treatment "child's use of AFDC" is binary—378 whether or not a study child ever used Aid to Families With Dependent Children (AFDC) from birth to current age in 1997. The authors of this study also examined each year's Panel Study of Income Dynamics (PSID) data for all study children from their birth to 1997 and obtained, for each child, the percentage of time using AFDC from birth to their current age in 1997. This example, then, assesses the impact of using different levels of AFDC participation on academic achievement. We hypothesize that an increase in levels of using AFDC decreases academic achievement in a linear fashion. After examining the distribution of percentages of time using AFDC from birth to current age, we create three dose groups: never used, 1% to 33.9% of the time using AFDC from birth to current age in 1997, and 34% to 100% of the time using AFDC from birth to current age in 1997. The distribution of these three doses is shown in Table 10.4. To investigate the impact of doses on academic achievement, we employ Imbens's (2000) method. A multinomial logit model was first estimated. Variables showing statistically significant differences from bivariate analyses were entered into the multinomial model as predictors. Results of the multinomial model are shown in Table 10.5. In Stata, we used the predict command following the estimation of the multinomial model to predict the generalized propensity scores for all study children. Doing so, we obtain three scores for each child, and each score indicates the generalized propensity of never using AFDC, the generalized propensity of using AFDC for 1% to 33.9% of the time, and the generalized propensity of using AFDC for 34% to 100% of the time. We then define the inverse of the propensity score predicting the study child's probability of using the actual dose as a sampling weight. We use the letter-word identification score in 1997 as the outcome variable to measure academic achievement and run a weighted OLS regression that also controls for clustering effects in the outcome analysis. The results of the outcome analysis are shown in Table 10.6. Table 10.4 Distribution of Dose Categories (Example 10.6.1) Source: Data from Hofferth et al., 2001. Table 10.5 Multinomial Logit Model Predicting Generalized Propensity (Example 10.6.1) 379 Source: Data from Hofferth et al., 2001. Note: Likelihood ratio $\chi^2(10) = 486.78$, $p < .000$, pseudo $R^2 = .313$. The multinomial logit model employs Dose Category 3 "34% to 100% of time" as a reference (omitted) category. AFDC = Aid to Families With Dependent Children. $**p < .01$, $***p < .001$, two-tailed test. Table 10.6 Regression Analysis of the Impact of Dose of Aid to Families With Dependent Children on the Letter-Word Identification Score in 1997 With and Without Propensity Score Adjustment (Example 10.6.1) 380 Source: Data from Hofferth et al., 2001. Note: AFDC = Aid to Families With Dependent Children. $*p < .1$, $**p < .05$, $***p < .01$. Assuming that $(Y1i, Y2i)$ have a joint normal distribution, with means (μ_1, μ_2) and covariance matrix define 395 The condition $Y1 > Y2$ implies $u < Z$. The mean income of hunters is given by where $\sigma_{2u} = \text{Cov}(u_2, u)$, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density function and standard normal distribution function, respectively. The mean income of fishermen is given by where $\sigma_{2u} = \text{Cov}(u_2, u)$. Because we have $\sigma_{2u} - \sigma_{1u} > 0$. Given the preceding definitions, we now can consider different cases. Case 1: $\sigma_{1u} < 0$, $\sigma_{2u} > 0$. In this case, the mean income of hunters is greater than μ_1 and the mean income of fishermen is greater than μ_2 . Under this condition, those who choose hunting have incomes better than the average income of hunters, and those who choose fishing have incomes better than the average income of fishermen. Case 2: $\sigma_{1u} < 0$, $\sigma_{2u} < 0$. In this case, the mean income of hunters is greater than μ_1 , and the mean income of fishermen is less than μ_2 . Under this condition, those who choose hunting are better than average in both hunting and fishing, but they are better in hunting than in fishing. Those who choose fishing are below average in both hunting and fishing, but they are better in fishing than in hunting. Case 3: $\sigma_{1u} > 0$, $\sigma_{2u} > 0$. This is the reverse of Case 2. Case 4: $\sigma_{1u} > 0$, $\sigma_{2u} < 0$. This is not possible, given the definitions of σ_{1u} and σ_{2u} . The above model has a significant impact on all econometric discussions of self-selection and the development of models correcting for self-selection. The key features of the Roy model are (a) that it was based on rational-choice theory, specifically the idea that individuals make a selection by optimizing an outcome, and (b) that the process cannot be assumed to be random, and factors determining the structure of selection should be explicitly specified and modeled. Maddala (1983) showed that Gronau (1974), H. G. Lewis (1974), and 396 Heckman (1974) followed a Roy model when they began their studies examining women in the labor force. In these studies, the observed distribution of wages became a truncated distribution, and the self-selectivity problem became a problem of incidental truncation. It was this pioneering work that instigated discussion of the consequences of self-selectivity and that motivated the development of econometric models to correct for self-selectivity. Although these models differ in methodology, a feature that is shared among all the models is that the impact of selection bias is neither relegated away nor assumed to be random. Rather, extending the choice perspective, it is explicitly used and estimated in the correction model. The more important models are (a) Heckman's two-stage sample selection model and (b) the variant of Heckman's two-stage model—Maddala's treatment effect model using a maximum likelihood estimator. In contrast, models that correct for selection bias by following the statistical tradition make a crucial assumption. They assume that selection follows a random process. Manski (2007) elaborated on this distinction, saying, "Whereas economists have often sought to model treatment selection in nonexperimental settings as conscious choice behavior, statisticians have typically assumed that treatment selection is random conditional on specified covariates. See, for example, Rubin (1974) and Rosenbaum and Rubin (1983)" (p. 151). Assuming random selection offers numerous advantages in the development of correction procedures. However, in practice, researchers need to consider whether such assumptions are realistic and choose a valid approach for correction. Although self-selection might be assumed to be random, other sources of selection (i.e., researcher selection, measurement selection such as rater effects, administrative selection, and attrition-induced selection) cannot be assumed to be random. Corrective methods for these sources of selection bias need to be developed. Despite these different perspectives (i.e., "selection is a rational choice and should be explicitly modeled" vs. "selection is random conditional on specified covariates"), the two traditions converge on a key feature: Both emphasize the importance of controlling with measured covariates that affect selection. Shortly after Heckman (1978, 1979) developed his two-stage estimator that used the predicted probability of receiving treatment, Rosenbaum and Rubin (1983) developed their propensity score matching estimator. Both models share a common feature of using the conditional probability of receiving treatment in correction. Another difference between the econometric tradition and the statistical tradition lies in the level of restrictiveness of assumptions. Because observational studies analyze potential outcomes or counterfactuals, the developer must impose assumptions on the model to make model parameters identifiable. Thus, a concern or topic of debate is the extent to which model assumptions are realistic. Based on the preceding discussion, four key issues in the development of 397 strategies to correct for selection bias emerge: (1) The econometric approach emphasizes the structure of selection and therefore underscores a direct modeling of selection bias, (2) the statistical approach assumes that selection is random conditional on covariates, (3) both approaches emphasize direct control of observed selection by using a conditional probability of receiving treatment, and (4) the two approaches are based on different assumptions and differ on the restrictiveness of assumptions. It is clear then that assumptions play a crucial role in correction. Users of correction models should be aware of the assumptions embedded in each model, should be willing to check the tenability of assumptions in their data, should choose a correction model that is best suited to the nature of the data and research questions, and should interpret findings with caution. In Chapters 4 through 10, we reviewed the assumptions embedded in each of the seven correction models—the Heckman sample selection model, propensity score matching, propensity score subclassification, propensity score weighting, matching estimators, kernel-based matching, and propensity score analysis of categorical and continuous treatments. A summary of the key assumptions for each model is provided in Table 11.1. Note that these correction models vary by the types of treatment effects estimated. Some models can be used to estimate both the average treatment effect and the average treatment effect for the treated, whereas other models can estimate only one of the two effects. We encourage users of corrective models to consider the assumptions listed in Table 11.1 and condition study findings on the degree to which assumptions are satisfied. 11.2 A MONTE CARLO STUDY COMPARING CORRECTIVE MODELS We conducted a Monte Carlo study to show the importance of checking the tenability of assumptions related to the corrective methods and to compare models under different scenarios of data generation (i.e., different types of selection bias). A Monte Carlo study is a simulation exercise designed to shed light on the small-sample properties of competing estimators for a given estimating problem (Kennedy, 2003). The same objectives can certainly be accomplished by other means, such as using statistical theories to derive the results analytically. That type of analysis underpins the work of the original developers of the corrective methods, and the main findings of those analyses have been highlighted in the chapters of this book. We chose to use a Monte Carlo study to compare models because such a simulation approach allows us to examine comparative results in a way that is more intuitive and less technical. Many Monte Carlo studies have been conducted on the properties of corrective approaches, but most of these studies have been comparisons of properties under different settings within one of the seven corrective methods. 398 For instance, Stolzenberg and Relles (1990), Hartman (1991), and Zuehlke and Zeman (1991) conducted Monte Carlo studies that examined various aspects of the Heckman or Heckit models. Kennedy (2003) reviewed these studies and summarized the main findings: Relative to subsample ordinary least squares (OLS) and on a mean square error criterion, the Heckman procedure does not perform well when the errors do not have a normal distribution, the sample size is small, the amount of censoring is small, the correlation between the errors of the regression and selection equations is small, or the degree of collinearity between the explanatory variables in the regression and selection equations is high. Kennedy warned that the Heckman model does poorly—and even does more harm than good—in the presence of high collinearity. Zhao (2004) conducted a Monte Carlo study to compare propensity score matching with covariate matching estimators under different conditions. Zhao found that selection bias due only to observables was a strong assumption; however, with a proper data set and if the selection-only-on-observables assumption was justifiable, matching estimators were useful for estimating treatment effects. Furthermore, Zhao found no clear winner among different matching estimators and that the propensity score matching estimators rely on the balancing property. Table 11.1 Key Assumptions and Effects by Correction Model 399 Freedman and Berk (2008) conducted data simulations to examine the properties of propensity score weighting. As mentioned earlier, they found that the weighting approach requires correctly specifying a causal model. In contrast to these studies, few Monte Carlo simulations have been conducted to compare the properties of corrective methods under a fixed setting; this is the objective of our Monte Carlo study. In our study, we compared six models (i.e., the OLS regression, the Heckit treatment effect model using maximum likelihood estimation, the propensity score one-to-one matching in conjunction with a postmatching regression analysis [PSM], the matching estimators, the propensity score weighting, and the propensity score subclassification) under a given setting. We ruled out kernel-based matching because it estimates only the average treatment effect for the treated, and the six 400 models listed above were not all designed to estimate this effect. Thus, our Monte Carlo study focused on the average treatment effect. We simulated two data generation settings (i.e., selection on observables and selection on unobservables) and compared performance across the six models within each setting. Our aim for the Monte Carlo study was to address the following four research questions: (1) Within each setting of selection bias, which model performs the best, and how are the six models ranked in terms of bias and mean square error criteria? (2) Within each setting of selection bias and modifying model specifications to simulate a pragmatic, real-world application, which model performs the best, and how are the six models ranked in terms of bias and mean square error criteria? (3) What conclusions, in terms of the sensitivity of model performance to the mechanisms of data generation, can we draw from the simulation study? Last, (4) comparing the data generation process defined by a given setting with assumptions embedded in a given model, how important and restrictive are the assumptions? We emphasize that the Monte Carlo study simulates very limited settings of data generation. There are certainly many other settings—actually an unlimited number of settings. 1 Therefore, the conclusions of our study cannot be generalized to other settings. We did not attempt to compare all six methods in general settings to determine which model was the best. The main purpose of the Monte Carlo study is to show the importance of checking data assumptions, that is, to demonstrate that models have different assumptions and the performances of the models under a common setting of data generation will vary. 11.2.1 Design of the Monte Carlo Study In this subsection, we first present a statistical framework that outlines the process of treatment assignment, and then we show specifications for the two settings of selection bias simulated by the Monte Carlo study. Last, we show the model specifications for each of the six models under these settings and the evaluation criteria we used. The statistical framework we adopted for this study was drawn from Heckman and Robb (1985, 1986, 1988), which aimed to model the assignment mechanism that generated treatment and control groups. Let $Y1i$ and $Y0i$ denote potential outcomes for observation i under the conditions of treatment and control, respectively. The potential outcomes $Y1i$ and $Y0i$ can be expressed as deviations from their means: Combining these two expressions with the observation rule given by the definition of the treatment assignment dummy variable W_i (i.e., Equation 2.1 in 401 Chapter 2, or $Y_i = W_iY1i + (1 - W_i)Y0i$), the equation for any Y_i is where $u_i = u0 + W_i(u1 - u0)$. Equation 11.1 is known as the structural equation. For a consistent estimate of the true average treatment effect, W_i and u_i must be uncorrelated. Consider a supplemental equation, known as the assignment or selection equation, as a latent continuous variable, we equation, that determines W_i . Denoting where Z_i is a row vector of values on various exogenous observed variables that affect the assignment process, α is a vector of parameters that typically needs to be estimated, and v_i is an error term that captures unobserved factors that affect assignment. The treatment dummy variable W_i is determined by the following rule: $W_i = 1$ if $u_i > c$, and $W_i = 0$ if $u_i < c$ (c is a cutoff value). Additional covariates X_i may be included in Equation 11.1, and X_i and Z_i may be in common. We can distinguish between two settings in which W_i and the error term u_i in Equation 11.1 can be correlated. Setting 1: Selection on the observables. When Z_i and u_i are correlated, but u_i and v_i are uncorrelated, we have the condition known as selection on the observables. This condition is alternatively known as the strongly ignorable treatment assignment assumption (Rosenbaum & Rubin, 1983). Under this condition, a statistical control such as an OLS regression of Equation 11.1 will be unbiased as long as Z_i is sufficiently included in the equation. Setting 2: Selection on the unobservables. When Z_i and u_i are uncorrelated, but u_i and v_i are correlated, the condition is known as selection on the unobservables. This setting simulates hidden selection bias and the violation of the regression assumption about uncorrelated error terms with an independent variable. Under this setting, results of OLS estimation are expected to be biased and inconsistent. In practice, controlling for selection bias due to unobservables is difficult. The data generation process for each of the two settings is shown in Figure 11.1. In addition to the earlier specifications, this process further imposes the following conditions: (a) Three variables or covariates (x_1, x_2 , and x_3) affect the outcome variable y , (b) z determines treatment assignment w only, and (c) x_3 402 also affects treatment assignment w . The preceding data generation for each setting was implemented by using Stata version 9.0; the syntax files for the data generation and analysis using six models are available on the companion webpage of this book. We encourage readers to replicate the study and to use our syntax as a baseline to generate more settings or to compare additional models. 1. Specifications of Setting 1 in Stata: The Stata syntax generates Setting 1 using the following specifications: Figure 11.1 Design of the Monte Carlo Study: Two Settings of Selection Bias where x_1, x_2, x_3, Z , and u are random variables, normally distributed with a mean vector of (3 2 10 5 0), standard deviation vector (.5 6.9 5 2 1), and the following symmetric correlation matrix: 403 In addition, v is a random variable that is normally distributed with mean zero and variance 1; $W = 1$, if $W^* > \text{Median}(W^*)$, and $W = 0$ otherwise. The preceding specifications create a correlation between Z and u of .4, as well as a correlation between u and v of 0. Thus, the data generation meets the requirements for simulating selection on observables, as shown in Setting 1 in Figure 11.1. The Monte Carlo study generates 10,000 samples for each corrective model with a size of 500 observations per sample. Under this specification, the true average treatment effect in the population is known in advance, that is, the effect of W equals .5, as shown in Equation 11.3. 2. Specifications of Setting 2 in Stata: The Stata syntax generates Setting 2 using the following specifications: where x_1, x_2, x_3, Z, u , and v are random variables, normally distributed with a mean vector of (3 2 10 5 0 0), standard deviation vector (.5 6.9 5 2 1 1), and the following symmetric correlation matrix: In addition, δ is a random variable that is normally distributed with mean zero and variance 1; $W = 1$, if $W^* > \text{Median}(W^*)$, and $W = 0$ otherwise. The above specifications create a correlation between Z and u of 0, as well as a close-to-zero correlation between u and v of .10. Thus, the data generation meets the requirements for simulating selection on unobservables, as shown in Setting 2 in Figure 11.1. The Monte Carlo study generates 10,000 samples for each corrective model with a size of 500 observations per sample. Under this specification, the true average treatment effect in the population is known in advance, that is, the effect of W equals .5, as shown by Equation 11.4. 3. Specifications of corrective models in Stata: The specifications for each of the six corrective models under Setting 1 are shown below. Model 1.1. OLS regression: 404 Model 1.2. Propensity score matching (greedy): The logistic regression model predicting the conditional probability is the predicted probability from the logistic regression is defined as the estimated propensity score; the propensity score matching procedure then matches each treated case to a control case (i.e., a 1-to-1 match) on the estimated propensity score using nearest neighbor within a caliper of .086 (i.e., a quarter of the standard deviation of the estimated propensity scores), and the postmatching analysis performs the following OLS regression of based on the matched sample. Model 1.3. Treatment effect model: The regression equation is The selection equation is $W^* = \Gamma Z + v$, $W = 1$ if $W^* > \text{Median}(W^*)$, and $W = 0$ otherwise; $\text{Prob}(W = 1 | Z) = \Phi(\Gamma Z)$ and $\text{Prob}(W = 0 | Z) = 1 - \Phi(\Gamma Z)$; and the model is estimated by the maximum likelihood estimation. 2 Model 1.4. Matching estimator: The matching covariates include x_1, x_2, x_3 , and Z ; the model is estimated by the bias-corrected and robust-variance estimator where the vector norm uses the inverse of the sample variance matrix. Model 1.5. Propensity score weighting (average treatment effect [ATE]): for all observations, are estimated using the The propensity scores, same logistic regression for the propensity score matching (greedy) model (i.e., Model 1.2). The propensity score weight for estimating average if $W = 1$; if $W = \text{treatment effect (ATE)}$ is 0. The outcome analysis uses the same OLS regression as Model 1.1, but the regression is $\omega(W, x)$ weighted. Model 1.6. Propensity score subclassification: The propensity scores, for all observations, are estimated using the same logistic regression for the Model 1.2. The sample is stratified into five subclasses based on the For each subclass, an OLS regression the same as Model 1.1 is performed. The overall sample ATE is computed by aggregating the five ATEs using Equation 6.3, the overall variance of ATE is a significance test computed based on Equation 6.4, for the overall ATE is performed by using 405, where The six corrective models under Setting 2 are the same as those under Setting 1. That is, Model 2.1. OLS regression: Same as Model 1.1. Model 2.2. Propensity score matching (greedy): Same as Model 1.2. Model 2.3. Treatment effect model: Same as Model 1.3. Model 2.4. Matching estimator: Same as Model 1.4. Model 2.5. Propensity score weighting (ATE): Same as Model 1.5. Model 2.6. Propensity score subclassification: Same as Model 1.6. In this design, Setting 1 simulates selection on observables. By design, Z is an important variable that determines sample selection, and the key of selection on observables is to control for Z correctly in the analysis model. This need to control for Z is why we specify Z as a covariate affecting selection in the OLS regression, in the logistic regression for estimating the propensity scores, in the selection equation for the treatment effect model, and in the matching estimator. In practice, the analyst may not know that Z is important and may inadvertently omit it from the analysis. The omission of Z from analysis creates overt selection bias. To simulate this scenario, we run another set of models under Setting 1 in which Z is not used in all models. The specifications used for this set of models are shown below. Model 1.1.1. OLS regression: Model 1.2.1. Propensity score matching (greedy): The logistic regression model predicting the conditional probability is the predicted probability from the logistic regression (x) is defined as the estimated propensity score; the propensity score matching procedure then matches each treated case to a control case (i.e., a 1-to-1 match) on the estimated propensity score using nearest neighbor within a caliper of .06 (i.e., a quarter of one standard deviation of the estimated propensity scores), and the postmatching analysis performs the following OLS regression of based on the matched sample. Model 1.3.1. Treatment effect model: The regression equation is the selection equation is $W^* = \gamma x_1 + \gamma_2 x_2 + \gamma_3 x_3 + v$, $W = 1$ if $W^* > \text{Median}(W^*)$, and $W = 0$ otherwise; $\text{Prob}(W = 1 | X) = \Phi(\gamma X)$ and $\text{Prob}(W = 0 | X) = 1 - \Phi(\gamma X)$; and the model is estimated by the maximum likelihood method. Model 1.4.1. Matching estimator: The matching covariates include x_1, x_2 , and x_3 ; the model is estimated by the bias-corrected and robust-variance estimator where the vector norm uses the inverse of the sample variance matrix. Model 1.5.1. Propensity score weighting (ATE): The outcome analysis but the uses a model similar to Model 1.1.1, or regression is weighted. Model 1.6.1. Propensity score subclassification: For each subclass, the OLS regression is the same as Model 1.1.1, or 4. Criteria used to assess model performance. We used two criteria to assess model performance. One criterion was the estimated bias, which is based on the mean value of the estimated average treatment effect of the 10,000 samples. Because the true treatment effect is known (i.e., 0.5), the mean estimated average treatment effect of the 10,000 samples minus 0.5 provides an estimation of bias for a given model. The second criterion was the estimated mean square error (MSE), which is estimated by the average and the true value of treatment of the squared differences between effect 0.5: MSE provides an estimation of the variation of the sampling distribution for the estimated treatment effects; a small MSE value indicates low variation. 11.2.2 Results of the Monte Carlo Study Table 11.2 presents the findings of the Monte Carlo study under the two settings. The main findings of model performances under Setting 1 are summarized below. In Setting 1, that is, selection on observables, the propensity score subclassification performed the best: On average, the subclassification estimated a treatment effect of 0.4989, which was 0.0011 below the true effect (or an underestimation of 0.2%) with a medium-sized MSE of 0.0283. The propensity score matching (greedy) model performed the second best: On average, the propensity score matching model estimated a treatment effect of 0.4875, which was 0.0125 below the true effect (or an underestimation of 2.5%) with a low MSE of 0.0152. The propensity score weighting (ATE) was ranked as third: On average, the propensity score weighting model estimated a treatment effect of 0.4851 (or an 407 underestimation of 3.0%) with a medium-sized MSE of 0.0253. Table 11.2 Results of Monte Carlo Study Comparing Models Note: ATE = average treatment effect; OLS = ordinary least squares. The OLS regression also worked reasonably well, although it was ranked fourth: On average, OLS regression estimated a treatment effect of 0.5375, which was 0.0375 above the true effect (or an overestimation of 7.5%) with a low MSE of 0.012. It is worth noting that OLS works reasonably well in this setting because x_3 and Z are the main variables determining selection, Z and u are correlated, and both source variables x_3 and Z are controlled in the analysis. These conditions are restrictive and may not hold in practice: In a typical application, we may not know that x_3 and Z are the major source of selection, x_3 and Z may not be available or collected, and Z and u may not be correlated. The matching estimator did not provide acceptable bias correction and was ranked fifth among the six models: On average, the matching estimator estimated the treatment effect of 0.4531, which was 0.0469 below the true effect (or an underestimation of 9.4%) with a medium-sized MSE of 0.0237. The primary reason for the poor showing of matching estimation of the treatment effect was that Z and u are correlated in this setting. Although it is true that this assumption 408 is also embedded in propensity score matching and OLS regression, greedy matching and OLS seem to provide more robust responses to the violation than the matching estimator. Note that all matching variables used in the matching model (i.e., x_1, x_2, x_3 , and Z) were continuous, so that the matching was not exact. Although the matching estimator made efforts to correct for bias by using a least squares regression to adjust the difference within the matches for the differences in covariate values, in the circumstances of selection on continuous observables, matching via vector norm was inferior to matching via a one-dimensional propensity score. In Setting 1, the variant of Heckman's two-stage model, Maddala's treatment effect model, performed the worst among the six models: On average, the treatment effect model estimated the treatment effect of 1.9285, which was 1.4285 above the true effect (or an overestimation of 285.7%) with a huge MSE of 2.0469. The primary reason for this poor estimation was that in Setting 1, u and v were not correlated, so the data violated the assumption regarding the correlation of errors between the regression and selection equations. This finding underscores the fact that the assumption of a nonzero correlation of the two error terms is crucial to a successful application of the treatment effect model. In addition, this finding confirmed that the Heckman selection model depends strongly not only on the model being correct but also on the tenability of model assumptions; this requirement is more pronounced than that of OLS regression, a point we made in Chapter 4. We now summarize the main findings of model performances under Setting 2. When u and v are correlated or when selection is on unobservables, the Heckit or Maddala treatment effect model provided excellent estimation of the average treatment effect and was ranked first among the six models: On average, the treatment effect model estimated the treatment effect of 0.5036, which was 0.0036 above the true effect (or an overestimation of 0.7%) with an excellent MSE of 0.0005. No other model could compete with the treatment effect model. Indeed, the other five models produced erroneous estimates for the treatment effects: The overestimation ranged from 27.5% for the matching estimator to 38.8% for the propensity score subclassification, and the estimated MSE ranged from 0.0395 for the propensity score matching (greedy) to 0.0707 for the propensity score subclassification. Among these models, the matching estimator and propensity score matching (greedy) did a better job, ranking second and third, respectively. The last three models (i.e., propensity score weighting [ATE], OLS regression, and propensity score subclassification) performed in a similar fashion, although they ranked as fourth, fifth, and sixth, respectively. The findings indicate that all corrective models, including OLS regression, are sensitive to hidden selection bias. The Heckit or the Maddala model produced the only acceptable results. Modeling treatment effects when selection is unobserved is challenging and requires great care and caution. 409 Under Setting 1, the crucial feature of data generation is understanding (i.e., observing) the source of selection and the use of controls for Z and x_3 in the analysis. This is a wishful condition that is unlikely to occur in practice. Had the analysis omitted the specification of Z or encountered a situation of overt bias, the estimation results would be unacceptable. Table 11.3 presents results of the Monte Carlo study under this setting. Indeed, when Z is omitted in the analysis, all models fail to provide unbiased estimation of the treatment effect—the bias from all models ranges from 10.3% overestimation to 262.4% overestimation. The models are ranked as follows: Propensity score subclassification performs the best (10.3% overestimation with a medium-sized MSE of .0304), propensity score weighting (ATE) is ranked second (11.6% overestimation with a medium-sized MSE of 0.288), matching estimator is ranked the third (88.7% overestimation with MSE of 0.2038), propensity score matching (greedy) is ranked fourth (98.06% overestimation with MSE of 0.2482), OLS regression is ranked fifth (100.1% overestimation with MSE of 0.2565), and the treatment effect model performs the worst of the six (262.4% overestimation with MSE of 1.8056). 11.2.3 Implications We can distill several implications from the Monte Carlo analyses. First, no single model works well in all scenarios. The "best" results depend on the fit between the assumptions embedded in a model and the process of data generation. A model performing well in one setting may perform poorly in another setting. Thus, the models are not robust against a variety of data situations. It is important to check the tenability of model assumptions in alternative applications and to choose a model that is suited to the nature of the data at hand. With a mismatch of the data structure and the analytic model, it is easy to observe a misleading result. Second, when information regarding the tenability of model assumptions is not available (e.g., when there is no way of knowing whether the study omits important covariates), findings must be conditioned on a discussion of model assumptions. At a minimum, when disseminating results, the assumptions that underpin estimated treatment effects should be disclosed and the conditions under which estimation may be compromised should be described. Table 11.3 Results of Monte Carlo Study Comparing Models Not Controlling for Z Under Setting 1 410 Note: ATE = average treatment effect; OLS = ordinary least squares. Third, and more specifically, the Heckit or Maddala treatment effect model relies strongly on the assumption that the errors of the selection and regression equations are correlated, and when this assumption is violated, the model fails catastrophically to provide an unbiased estimation. But in contrast to critiques of the Heckman selection model, the Monte Carlo study showed the treatment effect model to be robust against hidden bias. It was the only model (i.e., Model 2.3) that provided accurate estimation of the treatment effect under the setting of selection on unobservables (i.e., where selection is not plausibly measured). As we mentioned earlier, the use of the Heckit treatment effect model in program evaluation is controversial. The Monte Carlo study partially explains why this is the case. The key message from our data simulation is that if the assumptions embedded in the Heckit model are met in a real data setting, the model performs very well and can provide an estimated treatment effect that is close to the true effect. The results not only support proponents of the Heckit model but also strengthen our judgment that the Heckit model is the only one, out of all propensity score models discussed by this book, that can be competent in handling hidden selection bias. The problem is that in a real-world study, no one would be able to judge whether the assumptions are tenable, and therefore, no one knows whether the model provides acceptable results. It is this concern that worries critics of the Heckit model. Fourth, OLS regression appears to work acceptably only in circumstances (i.e., Model 1.1) that require the user to have previous knowledge of the main sources of selection bias, to have collected data to measure the main sources of selection bias, and to have correctly used those variables in the model. Even in these circumstances, propensity score subclassification, matching, and weighting may work better (see Table 11.2). Moreover, had Z been omitted in the analysis, the OLS results would have been marked by extreme overestimation (i.e., Model 1.1.1 overestimated the effect by 100.1%). Under these circumstances, propensity score subclassification and weighting appear 411 preferable to OLS regression (see Table 11.3). Thus, the Monte Carlo findings suggest that when the data process is similar to Setting 1 (i.e., where selection is measured or, by extension, where it is plausibly measured), propensity score subclassification or propensity score weighting (ATE) may be superior to OLS regression. Finally, the Monte Carlo study points again to the challenges imposed by hidden bias. Under the condition of selection on unobservables, or under the condition of selection on observables but when researchers have omitted a selection-related measure from analyses, all models except Model 2.3 provided biased estimates of treatment effect. Therefore, it is important to measure and model selection as thoroughly as possible given prior research knowledge and theory. Moreover, it is important to conduct sensitivity analyses to gauge bias induced by hidden selection. The development of these procedures is the topic of the next section. 11.3 ROSENBAUM'S SENSITIVITY ANALYSIS Hidden bias is essentially a problem created by the omission in statistical analysis of important variables, and omission renders nonrandom the unobserved heterogeneity reflected by an error term in regression equations. Although correcting the problem may involve taking steps such as specifying an analytic model by using additional variables, collecting additional data related to selection, or redesigning a study as a randomized experiment, these strategies can be impractical, expensive, and time-consuming. It is often preferable to start by conducting sensitivity analyses to estimate the level of bias. This type of analysis seeks to answer this question: How sensitive are findings to hidden bias? Although sensitivity analysis is exploratory, it is an important step in analyses using the models we have described in this book. Rosenbaum and Rubin (1983) and Rosenbaum (2002b, 2005) have recommended that researchers routinely conduct sensitivity analyses in observational studies. In this section, we describe an emerging framework for sensitivity analysis. 11.3.1 The Basic Idea Rosenbaum (2002b, 2005) provides a succinct and well-organized description of sensitivity analysis. As summarized by Berk (2004), Rosenbaum's approach is simple: One manipulates the estimated odds of receiving a particular treatment to see how much the estimated treatment effects may vary. What one wants to find is that the estimated treatment effects are robust to a plausible range of selection biases. (p. 231) 412 Because sensitivity analysis is so important, we illustrate the basic idea of this method below and explain one procedure (i.e., sensitivity analysis for matched pair studies using Wilcoxon's signed rank test) in detail. According to Rosenbaum (2005), A sensitivity analysis in an observational study addresses this possibility: it asks what the unmeasured covariate would have to be like to alter the conclusions of the study. Observational studies vary markedly in their sensitivity to hidden bias: some are sensitive to very small biases, while others are insensitive to quite large biases. (p. 1809) The original framework for this perspective came from Cornfield et al. (1959), who studied evidence linking smoking to lung cancer. In their effort to sort out conflicting claims, Cornfield and his colleagues derived an inequality for a risk ratio of the probability of death from lung cancer for smokers over the probability of death from lung cancer for nonsmokers. Cornfield et al. argued that to explain the association between smoking and lung cancer seen in a given study, an analyst would need a hidden bias of a particular magnitude in the inequality. If the association was strong, they proposed, then the hidden bias needed to explain it would be large. Therefore, the fundamental task for sensitivity analysis is to derive a range of possible values attributable to hidden bias (i.e., the so-called Γ s). Specifically, suppose that there are two units, j and k , and that the two units have the same observed covariates x but possibly different chances of receiving treatment; that is, $x[j] = x[k]$, but $\pi[j] \neq \pi[k]$. Then, units j and k might be matched to form a matched pair or placed together in the same subclass in an attempt to control overt bias due to x . The odds that units j and k receive the treatment are $\pi[j] / (1 - \pi[j])$ and $\pi[k] / (1 - \pi[k])$, respectively. The odds ratio is The sensitivity analysis goes further to assume that this odds ratio for units with the same x was at most some number of $\Gamma \geq 1$; that is, By the preceding definitions, if $\Gamma = 1$, then $\pi[j] = \pi[k]$ whenever $x[j] = x[k]$, so the study would be free of hidden bias. If $\Gamma = 2$, then two units that appear similar, that have the same x , could differ in their odds of receiving the treatment by as much as a factor of 2, so one unit might be twice as likely as the other to receive treatment. As explained by Rosenbaum (2002b), 413 In other words, Γ is a measure of the degree of departure from a study that is free of hidden bias. A sensitivity analysis will consider several possible values of Γ and show how the inferences might change. A study is sensitive if values of Γ close to 1 could lead to inferences that are very different from those obtained assuming the study is free of hidden bias. A study is insensitive if extreme values of Γ are required to alter the inference. (p. 107) Whereas the original sensitivity analysis of Cornfield et al. (1959) ignored sampling variability (which is hazardous except in extremely large samples), Rosenbaum (2002b) attended to sampling variation by developing methods to compute the bounds on inference quantities, such as p values or confidence intervals. Thus, for each $\Gamma > 1$, we obtain an interval of p values that reflect uncertainty because of hidden bias. "As Γ increases, this interval becomes longer, and eventually it becomes uninformative, including both large and small p values. The point, Γ , at which the interval becomes uninformative is a measure of sensitivity to hidden bias" (Rosenbaum, 2005, p. 1810). Rosenbaum developed various methods of sensitivity analysis, including McNemar's test, Wilcoxon's signed rank test, and the Hodges-Lehmann point and interval estimates for sensitivity analysis evaluating matched pair studies, sign-score methods for sensitivity analysis evaluating matching with multiple controls, sensitivity analysis for matching with multiple controls when responses are continuous variables, and sensitivity analysis for comparing two unmatched groups. All methods are explained in detail in Rosenbaum (2002b, chap. 4). A user-developed program in Stata is available to conduct some of these analyses. 11.3.2 Illustration of Wilcoxon's Signed Rank Test for Sensitivity Analysis of a Matched Pair Study To illustrate sensitivity analysis, we turn to an example originally used by Rosenbaum (2002b). Table 11.4 shows a data set for 33 pairs of children. The exposed group comprised children whose parents worked in a factory that used lead to manufacture batteries (i.e., the treatment group), and the control group comprised children who were matched to the treated children but whose parents were employed in other industries that did not use lead. The study hypothesized that children were exposed to lead that was inadvertently brought home by their parents. Table 11.4 reports the outcome data measured as the micrograms of lead found in a deciliter of each child's blood (i.e., $\mu\text{g}/\text{dl}$). For example, the blood lead level for the treated (i.e., exposed) child in the first pair was 38, and the blood lead level for the control child was 16, and the difference was 22. To remove selection bias, the study matched each treated child to a control child on age and

neighborhood of residence. The study found that when controlling for 414 age and neighborhood of residence, the average blood lead level of the treated children was 15.97 µg/dl higher than that of the control children; Wilcoxon's signed rank test shows that this treatment effect was statistically significant at the .0001 level. Even though the study used matching on two covariates to remove selection bias, to what extent was this study finding sensitive to hidden bias? Table 11.4 Example of Sensitivity Analysis: Blood Lead Levels (µg/dl) of Children Whose Parents Are Exposed to Lead at Their Places of Work Versus Children Whose Parents Are Unexposed to Lead at Their Places of Work 415 Source: Rosenbaum (2002b, p. 82). Reprinted with kind permission of Springer Science + Business Media. The sensitivity analysis for the matched pair study using Wilcoxon's signed rank test involves the following steps. Step 1: Compute ranked absolute differences d_s . This step includes the following procedures. Take the absolute value of differences, sort the data in an ascending order of the absolute differences, and create d_s that ranks the absolute value of differences and adjusts for ties. The results of Step 1 are shown in Table 11.5. Note that the first data line in this table shows the information for Pair 15, rather than Pair 1, as in Table 11.4. This is because Pair 15's absolute value of difference is 0, and after the sorting procedure, Pair 15 appears on the first line because it has the lowest absolute difference value in the sample. Note how the column of d_s is determined: d_s was first determined on the basis of the pair's rank and then adjusted in value for tied cases. For instance, among the first four cases, the second and third cases are tied. So the d_s value for each case is the average rank or $d_s = (2 + 3)/2 = 2.5$. Other tied pairs include Pairs 18 and 30, Pairs 4 and 11, Pairs 17 and 29, Pairs 32 and 33, Pairs 10 and 22, and Pairs 3 and 26; the d_s values for these pairs are all average ranks. Step 2: Compute the Wilcoxon signed rank statistic for the outcome difference between treated and control groups. Table 11.6 shows results of the procedures under Step 2. First, we calculate two variables, cs_1 and cs_2 , which are comparisons of outcome values between the exposed child and the control child in each pair: Let $cs_1 = 1$ if "the outcome of the exposed child" $>$ "the 416 outcome of the control," $cs_1 = 0$ otherwise; and $cs_1 = 0$ if "the outcome of the exposed child" $>$ "the outcome of the control." Similarly, let $cs_2 = 1$ if "the outcome of the control" $>$ "the outcome of the exposed child," and $cs_2 = 0$ otherwise. The computations of the two variables cs_1 and cs_2 are shown in columns (G) and (H) and where $Zs_1 = 1$ for the treated case, and Zs_2 (H). Next, we need to compute $\Gamma = 0$ for the control case. Results of this computation are shown in column (I), which is a product of d_s and (i.e., Next, we need to compute column [J]). Finally, we take the sum of column (J) to obtain the Wilcoxon signed rank statistic for the outcome difference between groups, which equals 527. Table 11.5 Exhibit of Step 1: Take the Absolute Value of Differences, Sort the Data in an Ascending Order of the Absolute Differences, and Create d_s That Ranks the Absolute Value of Differences and Adjusts for Ties 417 Source: Data from Rosenbaum, 2002b. Step 3: Compute needed statistics for obtaining the one-sided significance level for the standardized deviate when $\Gamma = 1$. Table 11.7 shows the results of the procedures related to Step 3. Note that for the purpose of clarity, we deleted columns (A) to (E) from this table. In Step 3, we first need to calculate and using the following rules: Next, by multiplying d_s with P_s , we obtained column (M); taking the sum of column (M), we obtained $E(T^+) = 280$, which is the expectation for the signed rank statistic under the null hypothesis of no treatment effect. Our next 418 calculation is to compute the variance of the signed rank statistic under the null. The last two columns (N) hypothesis of no treatment effect or (O) show the calculation of the variance. After we create column (O), we take the sum of column (O) to obtain the variance; that is, Finally, using the Wilcoxon signed rank statistic for the outcome difference between groups of 527, $E(T^+) = 280$, and $Var(T^+) = 3130.625$, we can calculate This statistic follows a standard normal distribution or using the normal distribution function of a spreadsheet, we obtain an approximate onesided significance level of $p < .00013$. The lower and upper bounds of the p value for the standard deviate when $\Gamma = 1$ are the same; that is, the minimum and maximum p values are both less than .0001. Step 4: Compute needed statistics for obtaining the one-sided significance levels for the standardized deviates (i.e., the lower and upper bounds of p values) when $\Gamma = 2$ and when $\Gamma = \text{other values}$. If the study were free of hidden bias, we would find strong evidence that a parent's occupational exposure to lead increased the level of lead in their children's blood. The sensitivity analysis asks how this conclusion might be changed by hidden biases of various magnitudes. Therefore, we need to compute the one-sided significance levels for standardized deviates (i.e., the lower and upper bounds of p value) under other values of Γ . Table 11.8 shows the calculation of the p values when $\Gamma = 2$. The calculation of the needed statistics for $\Gamma = 2$ is similar to the procedure of Step 3 when $\Gamma = 1$. The difference is that when $\Gamma > 1$, we also need to calculate $E(T^-)$ and $Var(T^-)$, which are the expectation and variance needed for computing the lower bound of the p value. In this step, as shown in Table 11.8, the calculation of $E(T^+)$ and $Var(T^+)$ follows the same procedure we followed for $\Gamma = 1$. The table shows that $E(T^+) = 373.3333$ and $Var(T^+) = 2782.7778$. Note that the column (N), dSP^- , is an addition to this table and is the product of d_s and P^- (i.e., column (L) multiplied by column (F)). Taking the sum of column (N), we obtained $E(T^-) = 186.6667$. The variance for the lower bound statistic $Var(T^-)$ is the same as $Var(T^+)$, so $Var(T^-) = 2782.7778$. Given these statistics, we can calculate the deviates as follows: The deviate for the and the deviate for the lower bound is upper bound is Checking a table of standard normal distribution or using the normal distribution function of a spreadsheet, we find that the p value associated with 2.91 is .0018, and the p value associated with 6.45 is less than .0001. Because this interval of p values does not approach a nonsignificant cutoff value of .05, a higher value of Γ can be used for further testing. We can then replicate Step 4 for other Γ values, such as $\Gamma = 3$, $\Gamma = 4$. To ease our exposition, we jump to the calculation of the lower and upper bounds of p 419 values when $\Gamma = 4.25$. It is at this value of Γ that the interval becomes uninformative. Table 11.9 presents results of the calculation of needed statistics for the p values when $\Gamma = 4.25$. As shown in the table, when $\Gamma = 4.25$, we have $E(T^+) = 453.3333$, $Var(T^+) = 1930.9070$, $E(T^-) = 106.6667$. Therefore, the and the deviate for the deviate for the upper bound is lower bound is Checking a table of standard normal distribution or using the normal distribution function of a spreadsheet, we find that the p value associated with 1.676 is .0468, and the p value associated with 9.566 is less than .0001. Table 11.6 Exhibit of Step 2: Calculate Wilcoxon's Signed Rank Statistic for the Differences in the Outcome Variable Between Treated and Control Groups 420 Source: Data from Rosenbaum, 2002b. Table 11.7 Exhibit of Step 3: Calculate Statistics Necessary for Obtaining the One-Sided Significance Level for the Standardized Deviate When $\Gamma = 1$ 421 Source: Data from Rosenbaum, 2002b. Table 11.10 shows the range of significance levels for the test statistic when Γ is equal to various values. As depicted, $\Gamma = 4.25$ is the value at which the significance interval becomes uninformative. Rosenbaum (2002b) interpreted this finding in the following way: 422 The table shows that to explain away the observed association between parental exposure to lead and child's lead level, a hidden bias or unobserved covariate would need to increase the odds of exposure by more than a factor of $\Gamma = 4.25$. (p. 115) Thus, based on this study, it appears that there is an association between parental occupational exposure to lead and child blood lead level. Moreover, the study finding is robust against hidden bias. 11.4 OVERVIEW OF THE STATA PROGRAM RBOUNDS Few software programs are available to conduct the kind of sensitivity analyses developed by Rosenbaum. But the Stata user-developed program `rbounds` (Gangl, 2007) allows users to perform sensitivity analyses for matched pair studies using Wilcoxon's signed rank test as well as the Hodges-Lehmann point and interval estimates. As a user-developed program, `rbounds` is not included in the regular Stata package, but the program can be downloaded from the Internet. To locate this software on the Internet, Stata users can use the `findit` command, followed by `rbounds` (i.e., `findit rbounds`), and then follow the online instructions to download and install the program. After installing `rbounds`, users can access the basic instructions for running the program by going to the program's help file. Table 11.8 Exhibit of Step 4: Calculate Needed Statistics for Obtaining the One-Sided Significance Levels for the Standardized Deviates (i.e., the Lower and Upper Bounds of p Value) When $\Gamma = 2$ 423 424 Source: Data from Rosenbaum, 2002b. Table 11.9 Exhibit of Step 4: Calculate Needed Statistics for Obtaining the One-Sided Significance Levels for the Standardized Deviates (i.e., the Lower and Upper Bounds of p Value) When $\Gamma = 4.25$ 425 Source: Data from Rosenbaum, 2002b. Table 11.10 Results of the Sensitivity Analysis for Blood Lead 426 Levels of Children: Range of Significance Levels for the Signed Rank Statistic Source: Rosenbaum (2002b, p. 115). Reprinted with kind permission of Springer Science + Business Media. The `rbounds` program is started with the following syntax: `rbounds varname, gamma(numlist)` In this command, `varname` is the outcome difference between the treated group and the control group. Before running the `rbounds` program, users should organize the data files at the pair level, so that each data line represents pair information. Specifically, the data should look like the data shown in Table 11.4 of this chapter. The data file should contain the following two variables: `pair` and `difference`, with `difference` being a `varname` the user specifies in the command. `gamma(numlist)` is the only required key word. It specifies the values of Γ for the sensitivity analysis. Users specify particular Γ values in the parentheses. By running the command, `rbounds` returns the minimum and maximum values of the p value using Wilcoxon's signed rank test, the minimum and maximum values of the Hodges-Lehmann point estimate, and the lower and upper bounds of the 95% confidence interval (i.e., the default) of the Hodges-Lehmann interval estimate. 11.5 EXAMPLES 11.5.1 Sensitivity Analysis of the Effects of Lead Exposure Table 11.11 exhibits the syntax and output of running `rbounds` using the sensitivity analysis for the study of the effects of lead exposure (see Section 11.3.2). In the syntax, `lead` is the variable showing the outcome difference between a treated participant and a control for each pair, and `gamma(1 4 4.25 5 6)` is a shortcut specification that forces the Γ value to start at 1, with 427 increments of 1 (i.e., $\Gamma = 2, 3$) up to the value of $\Gamma = 4$, and then to use the listed Γ values 4.25, 5, and 6. In this case, (1) specified within the parentheses tells the program that the Γ value increases with an increment of 1. Thus, the specification is equivalent to `gamma(1 2 3 4 2.5 5 6)`. Essentially, the output provides the minimum and maximum values of various inference quantities. The maximum and minimum p values using Wilcoxon's signed rank test are shown in the columns labeled `sig+` and `sig-`; following the `sig-` column, the output lists the Hodges-Lehmann point and interval estimates. Results of the output for `sig+` and `sig-` are identical to Table 11.10. Table 11.11 Exhibit of Stata `rbounds` Syntax and Output (Example 11.5.1) Source: Data from Rosenbaum, 2002b. 11.5.2 Sensitivity Analysis for the Study Using Pair Matching In Chapter 5, we conducted a regression analysis of difference scores based on a matched pairs sample following pair matching (see Section 5.8.4). The study found that in 1997, children who used Aid to Families With Dependent Children (AFDC) had an average letter-word identification score that was 3.17 points lower than that of children who never used AFDC ($p < .05$). Furthermore, the study adjusted selection based on six observed variables (i.e., the ratio of family income to poverty threshold in 1996, caregiver's education in 1997, caregiver's history of using welfare, child's race, child's age in 1997, and child's gender). We now examine the same data in an attempt to determine to 428 what extent this study is sensitive to hidden selection bias. Results of the sensitivity analysis are shown in Table 11.12. Using Wilcoxon's signed rank test, the sensitivity analysis shows that the study becomes sensitive to hidden bias at $\Gamma = 1.43$. Because 1.43 is a small value, we can conclude that the study is very sensitive to hidden bias, and therefore, further analysis that controls for additional biases is warranted. Table 11.12 Results of the Sensitivity Analysis for the Study of Children's Letter-Word Identification Score: Range of Significance Levels for the Signed Rank Statistic (Example 11.5.2) Source: Data from Hofferth et al., 2002b. We can further compare the level of sensitivity to bias of the AFDC study with that of other studies. Table 11.13 presents results of sensitivity analyses for four observational studies (Rosenbaum, 2005, Table 4). Among the four studies, Study 2 is the least sensitive to hidden bias; the study of the effects of diethylstilbestrol becomes sensitive at about $\Gamma = 7$. In contrast, Study 4 is the most sensitive to hidden bias; the study of the effects of coffee becomes sensitive at about $\Gamma = 1.3$. Rosenbaum (2005) provided the following explanation of the implications of the sensitivity analyses for these studies: Table 11.13 Sensitivity to Hidden Bias in Four Observational Studies 429 Source: Rosenbaum (2005, Table 4). Reprinted with permission from John Wiley & Sons. Study 1: Hammond (1964). Study 2: Herbst, Ulfelder, and Poskanzer (1971). Study 3: Morton et al. (1982). Study 4: Jick et al. (1973). A small bias could explain away the effects of coffee, but only an enormous bias could explain away the effects of diethylstilbestrol. The lead exposure study, although quite insensitive to hidden bias, is about halfway between these two other studies, and is slightly more sensitive to hidden bias than the study of the effects of smoking. (p. 1812) Compared with these four studies, our study of the effect of welfare use on academic achievement is slightly better (i.e., less sensitive to unobserved bias) than the study of the effects of coffee, but it is more sensitive to hidden bias than the three other studies. 11.6 CONCLUSION Selection bias is the most challenging analytic problem in observational studies. Although corrective approaches have been developed, a valid application of these approaches requires broad knowledge and skill. Properly using corrective models involves (a) having a thorough understanding of the sources of selection bias, (b) conducting a careful investigation of existing data and literature to identify all possible covariates that might affect selection and be used as covariates in a correction effort, (c) developing an understanding of the fit between the data generation process and the assumptions in correction models, (d) providing a cautious interpretation of study findings that is conditioned on the tenability of assumptions, and (e) conducting a sensitivity analysis to gauge the level of sensitivity of findings to hidden bias. NOTES 1. For instance, sample size in our study was fixed at 500. We could change this value to see model performance under different settings of sample size. This feature is helpful if the researcher needs to assess model properties when sample size is small. In fact, each fixed value used in the data generation can be changed, which allows a comparison of a set of scenarios for the parameter. We did not do this and this fixed our settings at two because 430 we attempted to accomplish a narrowly defined objective for this study. 2. Note that the selection equation includes Z only. This is not exactly equivalent to the propensity score matching model in which the logistic regression employs x_1, x_2, x_3 , and Z . We tried the model specifying x_1, x_2, x_3 , and Z in the selection equation, but the model did not converge. For comparative purposes, the current model captures the main features of Setting 1 and is the best possible model we can specify. 3. To find the p value, consult a table of the standard normal distribution (i.e., a Z table) or use an Excel function to obtain the p value by typing `"=1-NORMSDIST(4.414)"` in a cell. In this case, Excel returns a p value of .000050739. 431 CHAPTER 12 Concluding Remarks In this chapter, we conclude by making a few remarks on criticisms of observational studies, on the debate regarding the approximation of randomization using bias-correcting methods, and on directions for future development. Section 12.1 describes common pitfalls in observational studies. Section 12.2 highlights both the debate regarding and the criticism of approximating experiments using propensity score approaches. Section 12.3 reviews advances in modeling causality that are methodologically different from those we have described. Here we discuss James Robins's marginal structural models and Judea Pearl's directed acyclic graphs (DAGs). Last, Section 12.4 speculates on future developments. 12.1 COMMON PITFALLS IN OBSERVATIONAL STUDIES: A CHECKLIST FOR CRITICAL REVIEW In evaluation and intervention research, it is common to find of designs as having differential capacity for inferring causal relationships between programs and observed outcomes. At the top of what is called the "evidentiary hierarchy" sits meta-analyses of randomized controlled trials (RCTs). Meta-analyses are viewed as superior to single RCTs. In turn, single RCTs are viewed as superior to studies in which participants are not randomly assigned to treatment and control conditions. The ordinal ranking of research designs is used often in assessing grant proposals and in valuing the importance of findings. Researchers have generally agreed that inferential methods based on the assumption of a randomized assignment mechanism are superior to other approaches. Indeed, Rubin (2008) argued, The existence of these assignment-based methods, and their success in practice, documents that the model for the assignment mechanism is more fundamental for inference for causal effects than a model for the science. These methods lead to concepts such as unbiased estimation and asymptotic confidence intervals (due to Neyman), and p -values or significance levels for sharp null hypotheses (due to Fisher), all defined by the distribution of statistics (e.g., the difference of treatment and control 432 sample means) induced by the assignment mechanism. In some contexts, such as the U.S. Food and Drug Administration's approval of a new drug, such assignment mechanism-based analyses are considered the gold standard for confirmatory inferences. (pp. 814–815) In statistical analysis as opposed to research design, criteria for rank ordering methods and assessing "goodness" are less clear. We often argue that the method must fit the research question and that assumptions must always be met as a test of statistical conclusion validity (Shadish et al., 2002). In a rapidly developing field such as propensity score analysis, criteria may be murky because we are just beginning to understand the sensitivity of models to the assumptions on which they rest. As in propensity score analysis, we often have choices in the selection of statistical methods, and our choices should fit the data situation. Toward a better understanding of when and how to use the seven methods described in previous chapters, in the following we list 20 pitfalls that can trip up evaluators when they use propensity score methods. 1. Mismatch of the research question, the design, and the analytic method: Perhaps the most obvious pitfall of all involves a mismatch of the research question, the design, and the statistical method. When a study uses observational data to address research questions that are clearly related to causality (e.g., evaluating the effectiveness of a service or treatment), the research question and the design are mismatched. When causal attributions are to be made, control group designs with randomization are preferred. To be sure, some quasiexperimental designs—for example, regression-discontinuity designs—permit relatively strong causal inference. If a strong design is not used or if the design is compromised, bias-correcting methods should be used. In the absence of a strong design and/or a propensity score approach (or other types of correction models), addressing research questions related to causal inference is ill-advised. We have explained why: In observational studies where treatment assignment is nonignorable, the treated and comparison groups are formed naturally, are prone to numerous selection biases, and may be imbalanced on observed and unobserved covariates. The treated and comparison groups cannot be compared analytically using common covariance adjustments such as ordinary least squares (OLS) regression or other analyses that do not explicitly control for selection. In these situations, propensity score methods may be used conditionally. 2. Randomization failure: A randomized experiment may fail in practice because the conditions required to implement randomization were either infeasible or simply not met. Although there are a variety of ways in which randomization can fail, it typically fails when the rules for assigning participants to the treated and control conditions are inadvertently violated or purposively 433 thwarted. Because it relies on probability theory, random assignment can also be implemented correctly but produce imbalanced groups because of an insufficient sample size. Group-randomized designs appear particularly vulnerable to selection when a small number of units—say, schools—is available. Finally, although it is technically not a failure of randomization, random assignment can be compromised or broken by a variety of postrandomization effects in which participants react to assignment to treatment or control conditions. These include, for example, "spillover" or experimental contamination effects, in which participants in control conditions become aware of the elements of services provided to participants in treatment conditions. Randomization failure becomes a research pitfall when data are analyzed as if the design were not compromised. To avoid this common mistake, sample balance must be assessed as a test of whether randomization has worked as planned. When a balance problem is found, remedial measures, such as use of a correction method developed for observational studies to adjust for bias, must be considered. 3. Insufficient selection information: Assuming now that the researcher intends to use a propensity score or other type of correction analysis, what pitfalls occur in the context of implementing a bias correction strategy? It is at this point that early decisions in research design may affect the degree to which statistical methods can be used to adjust for imbalance. If covariates to explain potential selection were not included in the measurement model, insufficient information may be available for the selection equation or matching. At the design stage, selection biases observed in previous studies should be considered and used as a basis for the adoption of measures and instruments. This highlights, perhaps, the importance of having substantive area of expertise in conducting evaluations. Because unobserved bias is a more serious problem than observed bias, measures of potential hidden selection effects should be incorporated in data collection. We recommend that reviewing unmeasured variables affecting selection bias in prior studies and proposing to collect such data in a new study should be a routine element of proposal development. Indeed, critically discussing the plausibility of the measurement model relative to selection bias should be a criterion for scoring grant proposals. 4. Failure to justify and present the model predicting propensity scores (such as a logistic regression): Closely related to the failure to conduct a thorough literature review of factors affecting selection, this pitfall refers to an inadequate specification of the model predicting propensity scores. It is always recommended that researchers report the criteria used in the selection of the conditioning variables in the logistic regression or the probit model, the rationale for using a chosen set of conditioning variables, the choices of functional forms of the continuous conditioning variables, and the criteria used 434 to include interaction terms. To ensure that the logistic regression or the probit model is appropriate, it is important also to report statistics measuring collinearity and model fit. The process used in estimating the selection equation should always be described explicitly. An adequate study should include discussion of alternative or competing conditioning models, and it should describe the procedure used to derive the final logistic regression or probit equation. 5. Insufficient pre- and postadjustment analysis: In this case, the researcher may have properly measured potential selection effects and used a propensity score or other type of correction analysis, but analyses are not conducted to assess the effect of the adjustment strategy. The degree to which a statistical adjustment produces balance—that is, the success of correction using a model—must be shown in reports. Specifically, a study fails to present sufficient information if it does not report pre- and postadjustment balances on covariates. 6. Failure to evaluate assumptions related to the Heckit model: Turning now to specific models, this pitfall occurs when a study uses the Heckit treatment effect model but does not provide information on the degree to which data meet the assumptions embedded in the model. Specifically, use of the Heckit model requires discussion of the (normal) distribution of the outcome variable, the (nonzero) correlation of the error terms of the regression and selection equations, the level of collinearity of independent variables included in both equations, and the size of the sample. The sample must be sufficiently large to permit the use of the maximum likelihood estimator. 7. Failure to evaluate assumptions related to propensity score matching: This problem arises when a study employs propensity score matching but does not fully explain whether data meet the assumptions embedded in the matching model. When propensity score matching is used, the strongly ignorable treatment assignment and overlap assumptions in both the pre- and postmatching samples must be discussed explicitly and evidence of controlling for overt selection must be provided. 8. Failure to show justification of caliper size: In studies that use 1-to-1 nearest neighbor matching within calipers, the caliper size must be justified. The failure to justify the caliper size comes about through using only one caliper size without explaining why one size is adequate, failing to discuss the potential limitation of inexact matching or incomplete matching that is produced by a given caliper size, and failing to provide justification for using (or not using) other types of matching procedures, such as Mahalanobis metric matching. Relatedly, a substantial reduction of sample size after caliper-based matching is often troublesome and requires scrutiny because conclusions based on 435 subsamples may differ from those based on original samples. 9. Failure to discuss limitations of greedy matching: Greedy matching has a number of limitations, and, in our view, they should always be discussed. Principal among the limitations of greedy matching is the assumption that the propensity scores of the treated and control groups overlap. Relatedly, estimated scores must be patterned so as to provide a common support region that is of sufficient size for the method to work. Findings based on greedy matching should be assessed against additional analyses using different approaches, such as optimal matching. 10. Failure to control for clustering in both the model predicting propensity scores and the model focused on outcomes: When analyzing multilevel data, researchers should take nontrivial intraclass correlation coefficients into consideration. The logistic regression often needs to be adjusted by using a fixed effects model, a multilevel model with a narrow or a broad inference space, or a single cluster level model. The outcome analysis needs to be adjusted by including random effects in a linear mixed model or in a cross-classified random effects model. A common pitfall is to analyze multilevel data without explicitly controlling for the clustering effects. 11. Insufficient information on optimal matching procedures: Use of optimal matching always involves a set of decisions. These include decisions about forcing a 1-to-1 pair matching, forcing a pair matching with a constant ratio of treated and control participants, specification of a minimum and maximum number of controls for each treated participant in a variable matching, and justification of a matching structure in a full matching. Together and separately, these decisions affect both the level of bias reduction and efficiency. A detailed description of these decisions ensures appropriate interpretation of study results, and it enhances replication. 12. Erroneous selection of analytic procedures following optimal matching: Sometimes the researcher may simply choose a wrong analytic procedure following optimal matching. Based on the matched sample following an optimal matching, it is not uncommon for analysts to err by using OLS regression with a dummy treatment variable instead of conducting regression adjustment using difference scores, outcome analysis with the Hodges-Lehmann aligned rank test, or other types of analyses (see Section 5.5 in Chapter 5). 13. Failure to address limited overlap of covariates between treated and nontreated participants: Most propensity score models assume a sufficient overlap of the estimated propensity scores between groups or classes. When this assumption is violated, groups cannot be balanced on covariates. The 436 problem is more severe in propensity score subclassification. Trimming strategies are often needed to deal with a limited overlap problem. 14. Failure to evaluate assumptions related to matching estimators: Like other procedures, matching estimators are based on clear assumptions, which must be explored as a condition of appropriate use. These include the strongly ignorable treatment assignment and overlap assumptions. 15. Failure to correct for bias in matching estimators: When using matching estimators, a correction may be insufficient and additional applications may improve efficiency. For instance, when two or more continuous covariates are used, the researcher should not assume that matching is exact. With continuous covariates, the analysis needs to include bias-corrected matching that involves an additional regression adjustment. Additionally, when the treatment effect is not constant, researchers should employ a variance estimator allowing for heteroscedasticity. 16. Failure to evaluate assumptions related to kernel-based matching: So too, in kernel-based matching, the data must meet the assumptions embedded in the model. Specifically, the conditional and mean independence assumptions with regard to the treatment assignment must be evaluated. Kernel-based matching estimates an average treatment effect for the treated. This should not be confused with, or compared with, the sample (or population) average treatment effect. 17. Failure to estimate outcomes under a variety of bandwidths: In kernel-based matching, treatment effects are estimated within bandwidths. To our knowledge, there are no unequivocal criteria for the selection of the proper bandwidth. Thus, the robustness of findings should always be tested under different specifications of bandwidth values. Failure to do so leaves findings vulnerable to the speculation that the use of alternative bandwidths would have produced nontrivially different findings. 18. Lack of trimming when local linear regression is used: When local linear regression matching is used, matching treated participants is sometimes challenging at the region where controls are sparse. We have recommended the use of different trimming schemes to determine the robustness of findings. Failure to using trimming to assess the stability of findings leaves open the possibility that findings are related to marginal matches. 19. Failure to note concerns regarding bootstrapping: As we have indicated, kernel-based matching with bootstrapping continues to be controversial. Studies that use bootstrapping should warn reviewers and readers 437 that findings derived from such a procedure may be prone to errors. This is a rapidly developing area, and, as more findings come out, we will post information on significance testing in the kernel-based matching to our website. 20. Insufficient cross-validation: The promising methods we have described are changing quickly, as researchers develop and revise algorithms and code. Our Monte Carlo analyses suggest that findings are sensitive to different data situations. So there is much work to be done. From a practice perspective, when it is difficult to test assumptions, cross-validation of findings becomes imperative. This last pitfall occurs when a researcher draws conclusions from an observational study by using one correction method but fails to provide a warning note that the findings have not been cross-validated by other methods. Cross-validation strengthens inference and generalization. 12.2 APPROXIMATING EXPERIMENTS WITH PROPENSITY SCORE APPROACHES Over the past 35 years, methods have evolved in complexity as researchers have recognized the need to develop more efficient approaches for assessing the effects of social and health policies, including both federal and state programs that arise from legislative initiatives. As shown in this book, particularly useful advances have been made in the development of robust methods for observational studies. The criticism and reformulation of the classical experimental approach symbolize a shift in evaluation methods. Although Heckman published his first correction method in 1978, debate about correction methods is lively today, and it has fueled the development of new approaches. Proponents of the new methods posit that it is possible to develop robust and efficient analytic methods that approximate randomization. Furthermore, proponents argue that nonexperimental approaches should replace conventional covariance adjustment methods and that these bias correction approaches provide a reasonable estimate of treatment effects when randomization fails or is impossible. 12.2.1 Criticism of Propensity Score Methods Not surprisingly, this is a debatable perspective, and some critics do not hold such an optimistic view of statistical advances. In general, opponents question the assumptions made by correction methods, and they are skeptical that the conditions of real-world application can meet these assumptions. For instance, using earnings data from a controlled experiment of the effects of a mandatory welfare-to-work program, Michalopoulos et al. (2004) compared findings from randomization with findings generated using nonexperimental analytic approaches. Their analyses followed a methodology originated by LaLonde 438 (1986) and Fraker and Maynard (1987) to compare "true" experimental impact estimates with those obtained from nonexperimental approaches. Michalopoulos and his colleagues reported that the nonexperimental estimators all exhibited significant bias. They found that the smallest bias occurred when in-state comparison groups were used and short-term outcomes were examined, in which case comparison group earnings were on average 7% of true control group earnings. However, for longer term outcomes, the bias grew—almost doubling—and it grew even more when out-of-state comparison groups were yoked. On the basis of this finding, Michalopoulos and colleagues concluded that propensity scores correct less well for studies in which the treated and nontreated groups are not exposed to the same ecological influences. In addition, Michalopoulos and colleagues observed that OLS regression often yielded the same estimates as matching estimators. In 2004, Agodini and Dynarski conducted a similar study. With data from a randomized experiment designed to prevent school dropout, they used nearest neighbor matching on propensity scores to compare results of matching with results obtained under true randomization. The trial examined the effect of a prevention intervention on several subsequent student outcomes, and the comparison groups were drawn from two sources: (1) control group members in a separate but related experiment in different areas and (2) the national sample of the National Educational Longitudinal Survey. As summarized by Moffitt (2004), Agodini and Dynarski's (2004) results suggest that the propensity score matching estimators perform poorly. Indeed, the bias was the same whether a more geographically distant comparison group drawn from the national survey was used or the comparison group from the experiment was used. The authors conclude that significant selection on unobservables was probably present and compromised the matching procedure. Debate also occurs among proponents and developers of the nonexperimental approaches. The most prominent debate is that between two schools of researchers, each of whom follows a different tradition in developing correction methods. Throughout this book, we have shown that disagreements between econometricians and statisticians center on the restrictiveness of distributional assumptions made in estimators, the tenability of assumptions in real application settings, the extent to which researchers should not assume that the selection process is random and should make efforts to model the structure of selection, and the ability to control for hidden selection or unobserved heterogeneity. 12.2.2 Regression and Propensity Score Approaches: Do They Provide Similar Results? A key issue is whether propensity score approaches provide results that are substantially different from the results provided by a routine regression model. Should propensity score approaches be considered as an alternative to 439 regression? Several systematic reviews in the biomedical, epidemiological, and public health fields shed light on this critical issue, although findings from these reviews should be interpreted with caution. Using PubMed and the Science Citation Index, Stürmer et al. (2006) assessed the use of propensity score and routine covariance controls in studies published through 2003. Only 9 of 69 studies (13%) used both methods had an effect estimate using propensity scores that differed by more than 20% from that obtained with a conventional covariance control model. They concluded that propensity score analyses produce estimates quite similar to conventional multivariable approaches. In a similar study, Shah et al. (2005) conducted a systematic review of the Medline and Embase databases up to June 2003. They identified 43 studies that described at least one association between an exposure and an outcome, using both traditional regression and propensity score methods to control for confounding. From these 43 studies, Shah et al. counted 78 exposure–outcome associations. Statistical significance differed between regression and propensity score methods for only 8 of the associations (10%). In each case, the regression method produced a statistically significant association not observed with the propensity score method. The odds or hazard ratio derived using propensity scores was, on average, 6.4% closer to unity than that derived using traditional regression. From this, Shah et al. concluded that observational studies had similar results whether using traditional regression or propensity scores to adjust for confounding. Such findings challenge a key stance proponents of propensity score methods have taken. If regression is a robust approach and produces similar findings, what is the value of and is there a need for propensity score approaches? This is an excellent question, and we respond in four ways. First, in this book, whenever applicable, we have reviewed the limitations of propensity score approaches. We agree with critics that even randomized clinical trials are imperfect and subject in practice to a host of threats, particularly postrandomization threats, that compromise strong research design. Today, there are many ways for estimating the results of treatment in experimental, quasi-experimental, and observational studies. Like other methods, propensity score methods do not provide uncomplicated and unconditional answers to questions related to causal attribution. Multiple methods for estimating program effects are indicated for use within and across studies. Researchers using propensity score methods should be cautious because the limitations, as we currently understand them, are not trivial and findings vary markedly when assumptions are violated. Interpretation should be constrained by an understanding of the limits of data and analytical methods. In the context of strong research design, comparisons between propensity score and traditional covariance methods for estimating treatment effects warrant further exploration. Second, the convergence of findings between regression and propensity score analyses is not necessarily conclusive evidence for the failure of the latter 440 method. First, where it exists, convergence suggests that propensity score approaches are no worse than covariance control approaches. Second, given the rapid growth of propensity score methods, it is not clear that propensity score analyses have been implemented properly and, therefore, tested sufficiently. Indeed, this is the conclusion drawn by the second systematic review. Shah et al. (2005) found that of 43 studies, 12 (28%) reported verifying that the confounders were balanced between exposure groups after application of propensity scores and they displayed balance in a table, 7 (16%) reported balance but did not show it in a table, and 24 (56%) failed to report balance checks at all. In short, it was not clear that propensity score analyses met the standards we have described throughout this book. Shah et al. (2005) concluded, "Many of the reviewed studies did not implement propensity scores well" (p. 550). Unfortunately, information on covariate balance between treated and control groups after propensity scoring is missing in Stürmer et al. (2006). Not only balance checks but also the development of a logistic regression model predicting propensity scores plays a crucial role in the entire procedure. Weitzen et al. (2004) undertook a systematic review of Medline and Science Citation Index articles using propensity score analyses. They sought to determine the criteria used to estimate propensity scores in selection equations. They focused on criteria for the selection of variables, processes for estimating the functional form of continuous predictors, methods for

calculating interactions, tests for model discrimination, and use of goodness-of-fit measures across 47 studies meeting inclusion criteria. Few studies provided sufficient detail. After reviewing Shah et al. (2005) and Weitzens et al. (2004), Peter Austin and Muhammad Mamdani (2006) concluded, "Propensity score methods are often poorly implemented in applied clinical research" (p. 2085). As rigor in implementation improves, comparisons of propensity score and covariance control methods will be increasingly valuable. Third, as we have shown, the statistical principles for propensity score models are well developed with rigorous mathematical proofs, and the problems of endogeneity in regression have been clearly documented (see, e.g., Berk, 2004; Imbens, 2004; Imbens & Wooldridge, 2009). Whether a propensity score model produces similar or different results from regression appears to depend on the empirical setting. As suggested in our simulation studies (i.e., Chapters 3 and 11), many factors affect the data situation. It is possible that different data conditions (i.e., levels of correlations between residuals of the selection and outcome regressions, levels of confoundedness and unconfoundedness, levels of correlations between the endogenous explanatory variable and other covariates, levels of violation of the stable unit treatment value assumption, etc.) advantage OLS regression or covariance control over propensity score models. The problem is that in reality, it is hard to fully know the data environment and, as shown in our simulations, errors can be large. Under these circumstances, relying on regression alone runs the risk of 441 producing biased and inefficient estimates. In studies with observational data, it is for this reason that we advise applying various approaches, including regression and propensity score methods, and testing findings in sensitivity analyses. Finally, studies have shown that the choice of covariates in controlling for selection bias plays a central role in observational studies. Using results of a reanalysis of a within-study comparison that contrasts a randomized experiment and a quasi-experiment, Steiner, Cook, Shadish, and Clark (2010) provide useful information for identifying covariates likely to reduce bias when the true selection process is not known. Their findings can be particularly helpful when planning a study. To address the problem in estimating propensity scores (i.e., the problem of misspecification of the propensity score model and the resultant substantial bias of estimated treatment effects), Kosuke Imai and Marc Ratkovic (2014) developed a covariate balancing propensity score (CBPS) method that models treatment assignment while optimizing the covariate balance. The CBPS exploits the dual characteristics of the propensity score as a covariate balancing score and the conditional probability of treatment assignment. Imai and Ratkovic found that the CBPS dramatically improved the poor empirical performance of propensity score matching and weighting methods reported in the literature. As new methods refining the selection of covariates and the estimation of propensity scores are developed and applied, we believe that more advantages of propensity score approaches will emerge and that the closeness of results between propensity score models and regression will disappear. To conclude the discussion about the utility of propensity score approaches, we cite Shadish (2013) below, whose emphasis on paying special attention to assumptions of propensity score models and potential violations of these assumptions in empirical research coincides with our viewpoint: The use of propensity score analysis has proliferated exponentially, especially in the last decade, but careful attention to its assumptions seems to be very rare in practice. Researchers and policymakers who rely on these extensive propensity score applications may be using evidence of largely unknown validity. All stakeholders should devote far more empirical attention to justifying that each study has met these assumptions. (p. 129) 12.2.3 Criticism of Sensitivity Analysis (T) To be sure, sensitivity analysis also has become part of a grand debate about assessing cause and effect in observational settings. Although Rosenbaum's approach to sensitivity analysis is considered methodologically and mathematically elegant, Robins (2002) expressed skepticism about its usefulness in practice. Robins argued that Rosenbaum's model would be useful 442 only if experts could provide a plausible and logically coherent range for the value of the sensitivity parameter Γ , which measures the potential magnitude of hidden bias. To test this, Robins defined a measure of hidden bias to be paradoxical if its magnitude increases as the analyst decreases the amount of hidden bias by measuring some of the unmeasured confounders. On the basis of this definition, Robins proved that Rosenbaum's F fit the criteria of a paradoxical measure. He argued that sensitivity analysis based on a paradoxical measure of hidden bias may be scientifically useless because, without prolonged and careful training, users might reach misleading, logically incoherent conclusions. The debate is continuing as we write. Although we offer no judgment on the issues in dispute, it is our hope that our cautions and caveats will encourage readers to monitor the discussion, to explore the conditions under which correction models work, to discuss transparently the limitations embedded in observational studies, and to exercise critical thinking in using correction methods. 12.2.4 Group Randomized Trials Substantial progress in addressing evaluation challenges has also been made in the area of study design. However, even these efforts have sparked debate. One increasingly common design innovation is called group randomization (or cluster randomization), which aims to solve the problem of randomization failure at the individual level. We review this method below and show the usefulness of the correction methods developed for observational data for analyzing data generated by group randomization. In the past 25 years, group randomization has gained a footing as an alternative to individual randomization in social behavioral sciences research (for a review, see Bloom, 2004). At its core, the idea of the method is quite simple: Random assignment to treatment or control conditions is done on the group level rather than on the individual level. Specifically, instead of randomly assigning individuals to either the treatment or the control conditions, evaluators randomly assign groups (e.g., hospitals, schools) into study conditions. Thus, all the individuals within a given study group (e.g., all patients in a hospital or all students in a school) are assigned to the same condition (i.e., receive treatment or no treatment). According to Bloom, group randomization is useful when (a) the effects of a program have the potential to "spill over" from treatment participants to nonparticipants, (b) the most efficient delivery of program services is through targeting specific locations, (c) the program is designed to address a spatially concentrated problem or situation, (d) using a place-based group randomization reduces political opposition to randomization, and (e) maintaining the integrity of the experiment requires the physical separation of the treatment group from the control group. 443 Although group randomization has important practical applications in educational, health, social welfare, and other settings, Murray, Pals, Blitstein, Alfano, and Lehman (2008) found that the data from these designs are often incorrectly analyzed. To identify group-randomized trials, Murray and his colleagues used a set of key words to search the peer-reviewed literature on cancer prevention and control. Their investigation identified group randomization designs in 75 articles that were published in 41 journals. Among these studies, many of the researchers who used group randomization did not adequately attend to the analytic challenges raised in the design. Because individuals within the unit of randomization may have correlated outcomes (e.g., the scores of patients within the same hospital or treated by the same doctor may be correlated), calculations based on sampling variability must take the intracluster correlation into account. Ignoring this correlation will produce standard errors that are too small and will increase the potential for Type I errors. Compounding matters, group-randomized trials often have low power and, as we have mentioned, run the risk of failed randomization. Insufficient sample size at the group level (e.g., a small number of schools in a school-based trial) may cause failure of randomization and makes covariates between study conditions (i.e., treatment vs. control conditions) imbalanced. In such a situation, the study data may remain imbalanced at the individual level. Whenever imbalances of covariates occur, the study design cannot be treated as a randomized experiment. It is important to control for selection bias in the data analysis, and the correction methods described in this book should be considered. For example, we have shown how to use the Heckit treatment effect model to determine the treatment effect of a program that employed a group randomization design (see Section 4.4.2 in Chapter 4) and an efficacy subset analysis of the same data using matching estimators (see Section 8.4.2 in Chapter 8). These data suggest that correction methods designed for observational studies may be useful when group randomization does not produce ideal balances of covariates between study conditions. 12.3 OTHER ADVANCES IN MODELING CAUSALITY Methods aimed at correcting for bias in observational studies are developing rapidly. We have introduced only seven of these correction methods, and we have briefly described other methods that can be used to accomplish the same goal of data balancing (see Sections 2.5.3–2.5.5 in Chapter 2). We have chosen not to describe other methods that differ substantially in methodology from the seven methods that are the focus of this text. However, two of these methodologically different approaches are especially important and warrant 444 consideration: (1) the marginal structural model and (2) causal analysis using DAGs. Marginal structure models. In a series of publications, James Robins developed analytic methods known as marginal structural models that are appropriate for drawing causal inferences from complex observational and randomized studies with time-varying exposure or treatment (Robins, 1999a, 1999b; Robins et al., 2000). To a large extent, these methods are based on the estimation of the parameters for a new class of causal models—structural nested models—using a new class of estimators. The conventional approach to the estimation of the effect of a time-varying treatment or exposure on time to disease has been to model the hazard incidence of failure at time t as a function of past treatment history using a time-dependent Cox proportional hazards model. However, Robins showed that this conventional approach is biased. In contrast, the marginal structural models allow an analyst to make adjustments for the effects of concurrent nonrandomized treatments or nonrandom noncompliance that occur in many randomized clinical trials. For instance, when researchers need to clarify differences between association and causation, the inverse probability of treatment weighted (IPTW) estimation of a marginal structural model is particularly useful. The IPTW estimation consistently estimates the causal effect of a time-dependent treatment when all relevant confounding factors have been measured. To deal with hidden biases or effects of unmeasured variables, Robins developed a sensitivity analysis to adjust inferences concerning the causal effect of treatment as a function of the magnitude of confounding due to unmeasured variables (Robins, 1999b). Directed acyclic graphs. Judea Pearl (2000) and others (Glymour & Cooper, 1999; Spirtes et al., 1993) developed a formal framework to determine which of many conditional distributions could be estimated from data using DAGs. A DAG is a conventional path diagram with a number of formal mathematical properties attached. Pearl (2000) argued, "A causal structure of a set of variables V is a directed acyclic graph (DAG) in which each node corresponds to a distinct element of V , and each link represents a direct functional relationship among the corresponding variables" (p. 44). Pearl's framework focuses on the "inferred causation" inherent in the concept of a latent structure. Two latent structures are equivalent if they imply the same conditional distributions. Each latent structure in a set of such structures is "minimal" if it can produce the same conditional distributions and only those. A structure is consistent with the data if it reproduces the observed conditional distributions. According to Berk (2004), Pearl's contribution is to provide tools for winnowing down a set of potential causal models and then determining for a subset whether they speak with one voice about a particular causal relationship. 445 12.4 DIRECTIONS FOR FUTURE DEVELOPMENT Given advances such as those developed by Robins and Pearl, the growth of propensity score analysis methods, and the fertility of debates in the field, it is difficult to forecast what the future may hold. However, propensity score analysis "has reached a level of maturity" that makes it an important tool (Imbens & Wooldridge, 2009, p. 1). From their review of propensity score models in econometrics, Imbens and Wooldridge (2009) recommended three approaches for empirical application: matching estimators (methods described in Chapter 8), propensity score subclassification (methods described in Chapter 6), and propensity score weighting (methods described in Chapter 7). In a rapidly growing field, it is hard to predict, but we think that the following three directions are evident and likely to contribute substantially to the design of evaluation methods for observational studies. The first direction is the need to refine and expand analytical methods for researchers in the social and health sciences. Historically, the analytic methods for observational studies were developed and used primarily by statisticians and econometricians. However, when applying these methods to a broader range of social and health problems, issues that were not central to econometricians and statisticians have arisen and warrant consideration. These include the need to • develop a framework for power analysis for the correction methods. Although the social and health sciences have used Cohen's (1988) framework for power analysis, that framework does not provide estimators of sample size for propensity score models. The estimation of sample sizes is more complicated in propensity score analysis because most models require overlapping of propensity scores. Incorporating into the estimation the effect of sample reduction due to the common support region problem is an added complexity. • develop a standardized or metric-free measure of effect size (i.e., Cohen's d). The d_{xm} statistic described by Haviland et al. (2007) to measure covariance imbalance is similar to Cohen's d for post-optimal-matching analysis (see Sections 5.5.2 and 5.5.3 in Chapter 5). Similar measures and criteria to judge small-medium-large effect size should be developed for other types of propensity score analyses. • develop analytic models that incorporate special types of outcome variables (e.g., categorical and limited dependent variable, time-to-event variable, and outcome with a skew distribution) in analyses that currently assume a continuous outcome with normality. These models include optimal full and variable matching, matching estimators, and kernel-based matching. Extending these models to allow analysis of noncontinuous outcome variables will constitute a methods breakthrough. 446 • develop approaches to control for covariates that are differentially correlated with outcomes and treatment assignment. One of the three limitations of propensity score matching is that the approach cannot distinguish among effects of covariates that are differentially related to treatment assignment and observed outcomes (Rubin, 1997). In social behavioral research, covariates are often correlated differentially with outcomes and treatment assignment (i.e., covariates may be correlated with treatment assignment but not with the outcome; alternatively, covariates may be correlated more strongly with outcomes than with treatment assignment). To develop these approaches, Heckman, Ichimura, and Todd's (1998) work on separability (i.e., dividing the variables that determine outcomes into observables and unobservables) and exclusion restriction (i.e., isolating covariates that determine outcomes and program participation into two sets of variables T and Z , where the T variables determine outcomes, and the Z variables determine program participation) is promising. • develop correction models that control for measurement errors when they occur jointly with an endogeneity problem. One of the more challenging problems in social behavioral and health research is the nonrandom measurement error produced by multiple raters, particularly in longitudinal studies (see Section 11.1.1 in Chapter 11). • develop correction models that account for attrition selection in longitudinal studies. As discussed in Chapter 11, attrition cannot be assumed to be random and requires explicit modeling efforts. • improve correction models that control for clustering effect. As discussed in Chapter 8, the developers of matching estimators are aware of the limitation of not controlling for clustering in their method and are working to improve the matching estimators along this line. The second direction is to develop approaches that address challenges outlined in Heckman's (2005) framework for the "scientific model of causality" (see Section 2.9 in Chapter 2). Heckman weighed the implicit assumptions underlying four widely used methods of causal inference: (1) matching, (2) control functions, (3) the instrumental variables method, and (4) the method of DAGs. In his framework, he emphasized the need in policy research to forecast the impact of interventions in new environments, to identify parameters (causal or otherwise) from hypothetical population data, and to develop estimators evaluating different types of treatment effects (i.e., average treatment effect, treatment effect for the treated, treatment effect for the untreated, marginal treatment effect, and local average treatment effect). Finally, the third direction is to develop effective methods based on Rosenbaum's framework. We have cited Rosenbaum extensively in this book. As a statistician, Rosenbaum developed his framework based on the 447 randomization inference in completely randomized experiments. He then extended the covariance adjustment assuming randomization to observational studies free of hidden bias and finally to observational studies with hidden bias. For a detailed discussion of Rosenbaum's framework, readers are referred to a special issue of *Statistical Science* (2002, Vol. 17, No. 3), which presents an interesting dialogue between Rosenbaum and several prominent researchers in the field, including Angrist and Imbens (2002), Robins (2002), and Hill (2002). Rosenbaum has also made significant contributions to the development of optimal matching and sensitivity analysis. In the context of heated debate and ongoing disagreement, Rosenbaum's work sets a new standard for the analysis of observational studies. The refinements and advances latent in the Rosenbaum framework serve as a springboard for the future development of methods for observational studies. 448 References Abadie, A., Drukker, D., Herr, J. L., & Imbens, G. W. (2004). Implementing matching estimators for average treatment effects in Stata. *Stata Journal*, 4, 290–311. Abadie, A., & Imbens, G. W. (2002). Simple and bias-corrected matching estimators (Technical report). Department of Economics, University of California, Berkeley. Retrieved August 8, 2008, from Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74, 235–267. Achenbach, T. M. (1991). Integrative guide for the 1991 CBCL4–18, YSR, and TRF profiles. Burlington: University of Vermont, Department of Psychiatry. Agodini, R., & Dynarski, M. (2001). Are experiments the only option? A look at dropout prevention programs (Technical report). Princeton, NJ: Mathematica Policy Research. Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *Review of Economics and Statistics*, 86, 180–194. Ahmed, A., Husain, A., Love, T., Gambassi, G., Dell'Italia, L., Francis, G. S., et al. (2006). Heart failure, chronic diuretic use, and increase in mortality and hospitalization: An observational study using propensity score methods. *European Heart Journal*, 27, 1431–1439. Allison, P. D. (1995). Survival analysis using the SAS system. Cary, NC: SAS Institute. Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy*, 113, 151–184. Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from Social Security administrative records. *American Economic Review*, 80, 313–336. Angrist, J. D., & Imbens, G. W. (2002). Comment on "Covariance adjustment in randomized experiments and observational studies." *Statistical Science*, 17(3), 304–307. Angrist, J. D., Imbens, G. W., & Rubin, D. G. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91, 444–472. Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106, 979–1014. Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel studies. *Computational Statistics and Data Analysis*, 55, 1770–1780. Austin, P. C. (2008). A critical appraisal of propensity score matching in the medical literature from 1996 to 2003. *Statistics in Medicine*, 27, 2037–2049. Austin, P. C. (2009). The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*, 29, 661–677. Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine*, 25, 2084–2106. Barnard, J., Frangakis, C., Hill, J., & Rubin, D. B. (2003). A principal stratification approach to broken randomized experiments: a case study of vouchers in New York City (with discussion and rejoinder). *Journal of the American Statistical Association*, 98, 299–323. 449 Barnow, B. S., Cain, G. S., & Goldberger, A. S. (1980). Issues in the analysis of selectivity bias. In E. Stromsdorfer & G. Farkas (Eds.), *Education studies* (Vol. 5, pp. 42–59). Beverly Hills, CA: Sage. Barth, R. P., Gibbons, C., & Guo, S. (2006). Substance abuse treatment and the recurrence of maltreatment among caregivers with children living at home: A propensity score analysis. *Journal of Substance Abuse Treatment*, 30, 93–104. Barth, R. P., Greenson, J. K., Guo, S., & Green, B. (2007). Outcomes for youth receiving intensive in-home therapy or residential care: A comparison using propensity scores. *American Journal of Orthopsychiatry*, 77, 497–505. Barth, R. P., Lee, C. K., Wildfire, J., & Guo, S. (2006). A comparison of the governmental costs of long-term foster care and adoption. *Social Service Review*, 80(1), 127–158. Becker, S. O., & Caliendo, M. (2007). Sensitivity analysis for average treatment effects. *The Stata Journal*, 7(1), 71–83. Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal*, 2(4), 358–377. Benjamin, D. J. (2003). Does 401(k) eligibility increase saving? Evidence from propensity score subclassification. *Journal of Public Economics*, 87, 1259–1290. Berk, R. A. (2004). Regression analysis: A constructive critique. Thousand Oaks, CA: Sage. Bise, M., & Mattei, A. (2007). Application of the generalized propensity score. In Evaluation of public contributions to Piedmont enterprises (POLIS Working Paper 80). Alessandria, Italy: University of Eastern Piedmont. Bia, M., & Mattei, A. (2008). A Stata package for the estimation of the dose-response function through adjustment for the generalized propensity score. *The Stata Journal*, 8, 354–373. Bloom, H. S. (2004). Randomizing groups to evaluate place-based programs. Retrieved August 31, 2008, from www.wtgrantfoundation.org/usr_doc/RSCChapter4Final.pdf. Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley. Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, 20, 115–140. Bound, R. J., Jaeger, D. J., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of American Statistical Association*, 430, 443–450. Brand, J. E., & Xie, Y. (2010). Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. *American Sociological Review*, 75, 273–302. Brooks-Gunn, J., & Duncan, G. J. (1997). The effects of poverty on children. *Future of Children*, 7, 55–70. Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297–312. Cao, X. (2010). Exploring causal effects of neighborhood type on walking behavior using stratification on the propensity score. *Environment and Planning A*, 42, 487–504. Card, D. (1995a). Using geographic variation in college proximity to estimate the return to schooling. In L. N. Christophides, E. K. Grant, & R. Swidinsky (Eds.), *Aspects of labour market behavior: Essays in honour of John Vanderkamp* (pp. 201–222). Toronto: University of Toronto Press. Card, D. E. (1995b). Earnings, schooling, and ability revisited. *Research in Labor Economics*, 14, 23–48. Chen, X., Hong, H., & Tarozzi, A. (2008). Semiparametric efficiency in GMM models with 450 auxiliary data. *Annals of Statistics*, 36, 808–843. Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A*, 128, 134–155. Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 295–313. Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya, Series A*, 35, 417–446. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum. Corcoran, M., & Adams, T. (1997). Race, sex, and the intergenerational transmission of poverty. In G. J. Duncan & J. Brooks-Gunn (Eds.), *Consequences of growing up poor* (pp. 461–517). New York: Russell Sage Foundation. Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., & Wynder, E. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22, 173–203. Courtney, M. E. (2000). Research needed to improve the prospects for children in out-of-home placement. *Children and Youth Services Review*, 22(9–10), 743–761. Cox, D. R. (1958). *Planning of experiments*. New York: John Wiley. Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 187–220. Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability of sources and profiles. New York: John Wiley. Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *Review of Economics and Statistics*, 90(3), 389–405. Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96, 187–199. D'Agostino, R. B., Jr. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17, 2265–2281. D'Agostino, R. B., Jr. (2007). Propensity score in cardiovascular research. *Journal of the American Heart Association*, 115, 2340–2343. Dehejia, R., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94, 1053–1062. Derigs, U. (1988). Solving non-bipartite matching problems via shortest path techniques. *Annals of Operations Research*, 13, 225–261. Du, J. (1998). Valid inferences after propensity score subclassification using maximum number of subclasses as building blocks. Unpublished doctoral dissertation, Harvard University, Cambridge, MA. DuGoff, E. H., Schuler, M., & Stuart, E. A. (2014). Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Services Research*, 49, 284–303. Duncan, G. J., Brooks-Gunn, J., & Klebanov, P. K. (1994). Economic deprivation and early childhood development. *Child Development*, 65, 296–318. Duncan, G. J., Brooks-Gunn, J., Yeung, W. J., & Smith, J. R. (1998). How much does childhood poverty affect the life chances of children? *American Sociological Review*, 63, 406–423. Earle, C. C., Tsai, J. S., Gelber, R. D., Weinstein, M. C., Neumann, P. J., & Weeks, J. C. (2001). Effectiveness of chemotherapy for advanced lung cancer in the elderly: Instrumental variable and propensity analysis. *Journal of Clinical Oncology*, 19, 1064–1070. Edwards, L. N. (1978). An empirical analysis of compulsory schooling legislation 1940–1960. *Journal of Law and Economics*, 21, 203–222. Efron, B., & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials (with discussion). *Journal of the American Statistical Association*, 86, 9–17. Eichler, M., & Lechner, M. (2002). An evaluation of public employment programmes in the East German state of Sachsen-Anhalt. *Labour Economics*, 9, 143–186. Elwert, F., & Winship, C. (2010). Effect heterogeneity and bias in main-effects-only model. In R. Dechter, H. Geffner, & J. Y. Halpern (Eds.), *Heuristics, probability and causality, a tribute to Judea Pearl* (pp. 327–336). London: College Publications. English, D., Marshall, D., Brummel, S., & Coghlan, L. (1998). Decision-making in Child Protective Services: A study of effectiveness. Final Report. Olympia, WA: Department of Social and Health Services. Evans, W. N., & Schwab, R. M. (1995). Finishing high school and starting college: Do Catholic schools make a difference? *Quarterly Journal of Economics*, 110, 941–974. Fan, J. (1992). Design adaptive nonparametric regression. *Journal of the American Statistical Association*, 87, 998–1004. Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, 21, 196–216. Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver & Boyd. Fisher, R. A. (1971). The design of experiments. Edinburgh, UK: Oliver & Boyd. (Original work published 1935) Foster, E. M. (2003). Propensity score matching: An illustrative analysis of dose response. *Medical Care*, 41(10), 1183–1192. Foster, E. M., & Furstenberg, F. F., Jr. (1998). Most disadvantaged children: Who are they and where do they live? *Journal of Poverty*, 2, 23–47. Foster, E. M., & Furstenberg, F. F., Jr. (1999). The most disadvantaged children: Trends over time. *Social Service Review*, 73, 560–578. Fox, J. (2000). Nonparametric simple regression: Smoothing scatterplots. Thousand Oaks, CA: Sage. Fox, J. (2004, October 1). Personal communication via email. Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources*, 22, 194–227. Freedman, D. A., & Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation Review*, 32, 392–409. Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38, 367–378. Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *Annals of Statistics*, 28, 337–374. Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics*, 86, 77–90. Gadd, H., Hanson, G., & Manson, J. (2008). Evaluating the impact of firm subsidy using a multilevel propensity score approach. Working Paper Series, Center for Labour Market Policy Research, 3, 1–25. Galati, J. C., Royston, P., & Carlin, J. B. (2009). MIMSTAC: Stata module to stack multiply-imputed datasets into format required by mim. Retrieved February 27, 2009, 452 from Gangl, M. (2007). RBOUND: Stata module to perform Rosenbaum sensitivity analysis for average treatment effects on the treated. Retrieved August 1, 2007, from www.bccdc.ca/RePEc/bockcode/rGerber, A. S., & Green, D. P. (2000). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review*, 94, 653–663. Gibson-Davis, C. M., & Foster, E. M. (2006). A cautionary tale: Using propensity scores to estimate the effect of food stamps on food insecurity. *Social Service Review*, 80(1), 93–126. Glymour, C., & Cooper, G. (Eds.). (1999). *Computation, causation, and discovery*. Cambridge: MIT Press. Goldstein, H. (1987). Multilevel covariance component models. *Biometrika*, 74, 430–431. Goldstein, H. (2010). *Multilevel statistical models* (4th ed.). West Sussex, UK: John Wiley. Greene, W. H. (1981). Sample selection bias as specification error. *Econometrica*, 49, 795–798. Greene, W. H. (1995). LIMDEP, version 7.0: User's manual. Bellport, NY: Econometric Software. Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall. Grieswold, M., Localio, A., & Mulrow, C. (2010). Propensity score adjustments with multilevel data: Setting your sites on decreasing selection bias. *Annals of Internal Medicine*, 152, 393–396. Gronau, R. (1974). Wage comparisons: A selectivity bias. *Journal of Political Economy*, 82, 1119–1143. Gum, P. A., Thamilarasan, M., Watanabe, J., Blackstone, E. H., & Lauer, M. S. (2001). Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: A propensity analysis. *Journal of the American Medical Association*, 286, 1187–1194. Guo, S. (2005). Analyzing grouped data with hierarchical linear modeling. *Children and Youth Services Review*, 27, 637–652. Guo, S. (2008a). The Stata hodges program. Available from Guo, S. (2008b). The Stata imbalance program. Available from Guo, S. (2012). Preface to the Chinese translation of propensity score analysis (in Chinese). In Z. G. Guo & X. W. Wu (Eds. & Trans.), *Propensity score analysis: Statistical methods and applications* (pp. i–viii). Chongqing, China: Chongqing University Press. Guo, S. (2014). Correction of rater effects in longitudinal research with a cross-classified random effects model. *Applied Psychological Measurement*, 38, 37–60. Guo, S., Barth, R. P., & Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review*, 28, 357–383. Guo, S., & Bollen, K. A. (2013). Research using longitudinal ratings collected by multiple raters: One methodological problem and approaches to its solution. *Social Work Research*, 37(2), 85–98. Guo, S., & Hussey, D. L. (1999). Analyzing longitudinal rating data: A three-level hierarchical linear model. *Social Work Research*, 23, 258–269. Guo, S., & Hussey, D. L. (2004). Nonprobability sampling in social work research: 453 Dilemmas, consequences, and strategies. *Journal of Social Service Research*, 30, 1–18. Guo, S., & Lee, J. (2008). Optimal propensity score matching and its applications to social work evaluations and research. Unpublished working paper, School of Social Work, University of North Carolina, Chapel Hill. Guo, S., & Wildfire, J. (2005, June 9). Quasi-experimental strategies when randomization is not feasible: Propensity score matching. Paper presented at the Children's Bureau Annual Conference on the IV-E Waiver Demonstration Project, Washington, DC. Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, 11, 1–12. Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica*, 12, 1–115. Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66, 315–331. Hammond, E. C. (1964). Smoking in relation to mortality and morbidity: Findings in first thirty-four months of follow-up in a prospective study started in 1959. *Journal of the National Cancer Institute*, 32, 1161–1188. Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99, 609–618. Hansen, B. B. (2007). Optmatch: Flexible, optimal matching for observational studies. *R News*, 7, 19–24. Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15, 1–19. Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, 15(3), 234–249. Hardin, J. W., & Hilbe, J. M. (2003). *Generalized estimating equations*. New York: Chapman & Hall/CRC. Hardle, W. (1990). *Applied nonparametric regression*. Cambridge, UK: Cambridge University Press. Hartman, R. S. (1991). A Monte Carlo analysis of alternative estimators in models involving selectivity. *Journal of Business and Economic Statistics*, 9, 41–49. Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46, 1251–1271. Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods*, 12, 247–267. Heckman, J. J. (1974). Shadow prices, market wages, and labor supply. *Econometrica*, 42, 679–694. Heckman, J. J. (1976). Simultaneous equations model with continuous and discrete endogenous variables and structural shifts. In S. M. Goldfeld & R. E. Quandt (Eds.), *Studies in non-linear estimation* (pp. 235–272). Cambridge, MA: Ballinger. Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equations system. *Econometrica*, 46, 931–960. Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161. Heckman, J. J. (1992). Randomization and social policy evaluation. In C. Manski & I. Garfinkel (Eds.), *Evaluating welfare and training programs* (pp. 201–230). Cambridge, MA: Harvard University Press. Heckman, J. J. (1996). Comment on "Identification of causal effects using instrumental variables" by Angrist, Imbens, & Rubin. *Journal of the American Statistical Association*, 91, 459–462. Heckman, J. J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources*, 32, 441–462. Heckman, J. J. (2005). The scientific model of causality. *Sociological Methodology*, 35, 1–97. Heckman, J. J., & Hotz, V. J. (1989). Alternative methods for evaluating the impact of training programs (with discussion). *Journal of American Statistical Association*, 84, 862–874. Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. E. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66, 1017–1098. Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies*, 64, 605–654. Heckman, J. J., Ichimura, H., & Todd, P. E. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies*, 65, 261–294. Heckman, J. J., LaLonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor market programs. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3, pp. 1865–2097). New York: Elsevier. Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions. In J. Heckman & B. Singer (Eds.), *Longitudinal analysis of labor market data* (pp. 156–245). Cambridge, UK: Cambridge University Press. Heckman, J. J., & Robb, R. (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Ed.), *Drawing inferences from self-selected samples* (pp. 63–113). New York: Springer-Verlag. Heckman, J. J., & Robb, R. (1988). The value of longitudinal data for solving the problem of selection bias in evaluating the impact of treatment on outcomes. In G. Duncan & G. Kalton (Eds.), *Panel surveys* (pp. 512–538). New York: John Wiley. Heckman, J. J., & Smith, J. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, 9, 85–110. Heckman, J. J., & Smith, J. (1998). Evaluating the welfare state (Frisch Centenary Econometric Monograph Series). Cambridge, UK: Cambridge University Press. Heckman, J. J., Smith, J., & Clements, N. (1997). Making the most out of social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies*, 64, 487–536. Heckman, J. J., & Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*, 96, 4730–4734. Heckman, J. J., & Vytlacil, E. J. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73, 669–738. Helmreich, J. E., & Pruzek, R. M. (2008). The PSA graphics package. Retrieved October 30, 2008, from www.r-project.org/Herbst, A., Ulfelder, H., & Poskanzer, D. (1971). Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumor appearance in young women. *New England Journal of Medicine*, 284, 878–881. Hill, J. (2002). Comment on "Covariance adjustment in randomized experiments and observational studies." *Statistical Science*, 17, 307–309. Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology*, 2, 259–278. 455 Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. In A. Gelman & X. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 73–84). West Sussex, England: John Wiley. Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71, 1161–1189. Ho, D., Imai, K., King, G., & Stuart, E. (2004). Matching as nonparametric preprocessing for improving parametric causal inference. Retrieved October 20, 2008, from Hodges, J., & Lehmann, E. (1962). Rank methods for combination of independent experiments in the analysis of variance. *Annals of Mathematical Statistics*, 33, 482–497. Hofferth, S., Stafford, F. P., Yeung, W. J., Duncan, G. J., Hill, M. S., Lepkowski, J., et al. (2001). Panel study of income dynamics, 1968–1999: Supplemental files (computer file), ICPSR version. Ann Arbor: University of Michigan Survey Research Center. Holland, P. (1986). *Statistics and*

(discussion). *Journal of the American Statistical Association*, 81, 945–970. Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multi-level observational data. *Journal of the American Statistical Association*, 101, 901–910. Hong, G., & Yu, B. (2008). Effects of kindergarten retention on children's social-emotional development: An application of propensity score method to multivariate multi-level data. *Developmental Psychology*, 44, 407–421. Horvitz, D., & Thompson, D. (1952). A generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association*, 47, 663–685. Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: John Wiley. Hoxby, C. M. (1994). How teachers' unions affect education production. *Quarterly Journal of Economics*, 111, 671–718. Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 221–233). Berkeley: University of California Press. Hume, D. (1959). An enquiry concerning human understanding. LaSalle, IL: Open Court Press. (Original work published 1748) Iacus, S. M., King, G., & Porro, G. (2008). Matching for causal inference without balance checking. Retrieved October 30, 2008, from Imai, K., & Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99, 854–866. Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society*, B76, 243–263. Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87, 706–710. Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86, 4–29. Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142, 615–635. Imbens, G. W., Newey, W., & Ridder, G. (2006). Mean-squared-error calculations for 456 average treatment effects. Unpublished manuscript, Department of Economics, Harvard University, Cambridge, MA. Imbens, G. W., Rubin, D. B., & Sacerdote, B. I. (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American Economic Review*, 91, 778–794. Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47, 5–86. Jann, B., & Brand, J. E., & Xie, Y. (2010). Stata module to perform heterogeneous treatment effect analysis. Retrieved March 7, 2014, from Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press. Jick, H., Miettinen, O., Neff, R., Jick, H., Miettinen, O. S., Neff, R. K., et al. (1973). Coffee and myocardial infarction. *New England Journal of Medicine*, 289, 63–77. Joffe, M. M., & Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology*, 150, 327–333. Jones, A. S., D'Agostino, R. B., Gondolf, E. W., & Heckert, A. (2004). Assessing the effect of batterer program completion on reassault using propensity scores. *Journal of Interpersonal Violence*, 19, 1002–1020. Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36, 409–426. Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 10, 631–639. Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22, 523–539. Kaplan, D. (1999). An extension of the propensity score adjustment method for the analysis of group differences in MIMIC models. *Multivariate Behavioral Research*, 34(4), 467–492. Kaplan, D. (2000). Structural equation modeling: Foundations and extensions. Thousand Oaks, CA: Sage. Keele, L. J. (2008). The RBOUNDS package. Retrieved October 30, 2008, from www.rproject.org Kempthorne, O. (1952). The design and analysis of experiments. New York: John Wiley. Kennedy, P. (2003). *A guide to econometrics* (5th ed.). Cambridge: MIT Press. Kim, J., & Seltzer, M. (2007). Causal inference in multilevel settings in which selection process varies across schools (Working Paper 708). Los Angeles: University of California, Los Angeles, Center for the Study of Evaluation. King, G., & Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, 14(2), 131–159. Retrieved October 30, 2008, from King, G., & Zeng, L. (2007). When can history be our guide? The pitfalls of counterfactual inference. *International Studies Quarterly*, 51(1), 183–210. Retrieved October 30, 2008, from Kluge, J., Schneider, H., Uhlendorff, A., & Zhao, Z. (2012). Evaluating continuous training programmes by using the generalized propensity score. *Journal of the Royal Statistical Society*, Series A, 175(2), 587–612. Knight, D. K., Logan, S. M., & Simpson, D. D. (2001). Predictors of program completion for women in residential substance abuse treatment. *American Journal of Drug and Alcohol Abuse*, 27, 1–18. 457 Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, 36, 249–277. Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied linear regression models* (4th ed.). New York: McGraw-Hill/Irwin. Laird, N., & Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963–974. LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76, 604–620. Landes, W. (1968). The economics of fair employment laws. *Journal of Political Economy*, 76, 507–552. Lazarsfeld, P. F. (1959). Problems in methodology. In R. K. Merton, L. Broom, & L. S. Cottrell, Jr. (Eds.), *Sociology today: Problems and prospects* (Vol. 1, pp. 39–72). New York: Basic Books. Lechner, M. (1999). Earnings and employment effects of continuous off-the-job training in East Germany after unification. *Journal of Business and Economic Statistics*, 17, 74–90. Lechner, M. (2000). An evaluation of public sector sponsored continuous vocational training programs in East Germany. *Journal of Human Resources*, 35, 347–375. Lee, E. W., Wei, L. J., & Amato, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In J. P. Klein & P. K. Goel (Eds.), *Survival analysis: State of the art* (pp. 237–247). Dordrecht, The Netherlands: Kluwer Academic. Lehmann, E. L. (2006). *Nonparametrics: Statistical methods based on ranks* (Rev. ed.). New York: Springer. Leow, C., Marcus, S., Zanutto, E., & Boruch, R. (2004). Effects of advanced course-taking on math and science achievement: Addressing selection bias using propensity scores. *American Journal of Evaluation*, 25(4), 461–478. Leuven, E., & Sianesi, B. (2003). PSMATCH2 (version 3.0.0): Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Retrieved August 22, 2008, from Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press. Lewis, D. (1986). *Philosophical papers* (Vol. 2). New York: Oxford University Press. Lewis, H. G. (1974). Comments on selectivity biases in wage comparisons. *Journal of Political Economy*, 82, 1145–1155. Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society*, Series B, 34, 1–41. Littell, J. H. (2001). Client participation and outcomes of intensive family preservation services. *Social Work Research*, 25(2), 103–113. Littell, J. H. (2005). Lessons from a systematic review of effects of multisystemic therapy. *Children and Youth Services Review*, 27, 445–463. Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D., & Schabenberger, O. (2006). SAS system for mixed models (2nd ed.). Cary, NC: SAS Institute, Inc. Little, R. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley. Lochman, J. E., Boxmeyer, C., Powell, N., Roth, D. L., & Windle, M. (2006). Masked intervention effects: Analytic methods for addressing low dosage of intervention. *New Directions for Evaluation*, 110, 19–32. Lohr, S. L. (1999). *Sampling: Design and analysis*. Boston: Brooks/Cole. 458 Long, J. S. (1997). Regression models for categorical and limited dependent variables. Thousand Oaks, CA: Sage. Lu, B., Zanutto, E., Hornik, R., & Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, 96, 1245–1253. Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23, 2937–2960. Maddala, G. S. (1983). Limited-dependent and qualitative variables in econometrics. Cambridge, UK: Cambridge University Press. Magura, S., & Laudet, A. B. (1996). Parental substance abuse and child maltreatment: Review and implications for intervention. *Children and Youth Services Review*, 3, 193–220. Manning, W. G., Duan, N., & Rogers, W. H. (1987). Monte-Carlo evidence on the choice between sample selection and 2-part models. *Journal of Econometrics*, 35, 59–82. Manski, C. F. (2007). *Identification for prediction and decision*. Cambridge, MA: Harvard University Press. Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, 58, 690–700. Mantel, N., & Haenszel, W. (1959). Statistical aspects of retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748. Maxwell, S. E., & Delaney, H. D. (1990). Designing experiments and analyzing data: A model comparison perspective. Pacific Grove, CA: Brooks/Cole. McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9, 403–425. McCaffrey, D. F., Griffin, B. A., Almiral, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32, 3388–3414. McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society*, Series B, 42, 109–142. McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall. McMahon, T. J., Winkel, J. D., Suchman, N. E., & Luther, S. S. (2002). Drug dependence, parenting responsibilities, and treatment history: Why doesn't mom go for help? *Drug and Alcohol Dependence*, 65, 105–114. McNemar, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentage. *Psychometrika*, 12, 153–157. Mease, D., Wyner, A. J., & Buja, A. (2007). Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research*, 8, 409–439. Michalopoulos, C., Bloom, H. S., & Hill, C. J. (2004). Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? *Review of Economics and Statistics*, 86, 156–179. Mill, J. S. (1843). *System of logic* (Vol. 1). London: John Parker. Miller, A. (1990). Subset selection in regression. London: Chapman & Hall. Ming, K., & Rosenbaum, P. R. (2001). A note on optimal matching with variable controls using the assignment algorithm. *Journal of Computational and Graphical Statistics*, 10, 455–463. Moffitt, R. A. (2004). Introduction to the symposium on the econometrics of matching. *Review of Economics and Statistics*, 86, 1–3. 459 Morgan, S. L. (2001). Counterfactuals, causal effect, heterogeneity, and the Catholic school effect on learning. *Sociology of Education*, 74, 341–374. Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. New York: Cambridge University Press. Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M., & Saah, M. (1982). Lead absorption in children of employees in a lead-related industry. *American Journal of Epidemiology*, 115, 549–555. Mosteller, C. F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley. Murray, D. M., Pals, S. L., Blitstein, J. L., Alfano, C. M., & Lehman, J. (2008). Design and analysis of group-randomized trials in cancer: A review of current practices. *Journal of the National Cancer Institute*, 100, 483–491. Muthén, B. O., & Jöreskog, K. G. (1983). Selectivity problems in quasi-experimental studies. *Education Review*, 7, 139–174. Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide*. Los Angeles: Muthén & Muthén. Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, Series A, 135, 370–384. Neyman, J. S. (1923). Statistical problems in agricultural experiments. *Journal of the Royal Statistical Society*, Series B, 2, 107–180. Nobel Prize Review Committee. (2000). The Sveriges Riksbank Prize in economic sciences in memory of Alfred Nobel 2000. Retrieved August 8, 2008, from Normand, S. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., et al. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology*, 54, 387–398. NSCAW Research Group. (2008). Methodological lessons from the National Survey of Child and Adolescent Well-being: The first three years of the USA's first national probability study of children and families investigated for abuse and neglect. *Children and Youth Services Review*, 24, 513–541. Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill. Obenchain, B. (2007). The USPS package. Retrieved October 30, 2008, from www.rproject.org Owen, M., Imai, K., King, G., & Lau, O. (2013). Zelig: everyone's statistical software. Retrieved March 6, 2014, from Parsons, L. S. (2001). Reducing bias in a propensity score matched-pair sample using greedy matching techniques (SAS SUGI paper 214-26). Proceedings of the 26th annual SAS Users' Group International Conference, Cary, NC: SAS Institute. Retrieved August 22, 2008, from www2.sas.com/proceedings/sugi26/p214-26.pdf Pearl, J. (2000). *Causality: Models, researching, and inference*. Cambridge, UK: Cambridge University Press. Perkins, S. M., Tu, W., Underhill, M. G., Zhou, X., & Murray, M. D. (2000). The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety*, 9, 93–101. Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, 53, 873–880. Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association*, 67, 306–310. R Foundation for Statistical Computing. (2008). R version 2.6.2 [Computer software]. Retrieved August 8, 2008, from R Foundation for Statistical Computing. (2013). R version 3.0.1 [Computer software]. Retrieved September 1, 2013, from Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, 18(4), 321–349. Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage. Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics*, 31, 172–181. Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., & Burgette, L. (2013). twang: toolkit for weighting and analysis of nonequivalent groups. Retrieved March 6, 2014, from Robins, J. M. (1998). Marginal structural models. In 1997 Proceedings of the Section on Bayesian Statistical Science (pp. 1–10). Alexandria, VA: American Statistical Association. Robins, J. M. (1999a). Association, causation, and marginal structural models. *Synthese*, 121, 151–179. Robins, J. M. (1999b). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran & D. Berry (Eds.), *Statistical models in epidemiology: The environment and clinical trials* (pp. 95–134). New York: Springer-Verlag. Robins, J. M. (2002). Comment on "Covariance adjustment in randomized experiments and observational studies." *Statistical Science*, 17, 309–321. Robins, J. M., Hernn, M., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11, 550–560. Robins, J. M., & Ronitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90, 122–129. Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82, 387–394. Rosenbaum, P. R. (1997). The role of a second control group in an observational study (with discussion). *Statistical Science*, 2, 292–316. Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17, 286–304. Rosenbaum, P. R. (2002b). *Observational studies* (2nd ed.). New York: Springer. Rosenbaum, P. R. (2005). Sensitivity analysis in observational studies. In B. S. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1809–1814). New York: John Wiley. Rosenbaum, P. R. (2010). Design of observational studies. New York: Springer. Rosenbaum, P. R., Ross, R. N., & Silber, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association*, 102, 75–83. Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55. Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516–524. Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, 39, 33–38. Rossi, P. H., & Freeman, H. E. (1989). Evaluation: A systematic approach (4th ed.). Newbury Park, CA: Sage. Rothstein, J. (2007). Does competition among public schools benefit students and taxpayers? Comment. *American Economic Review*, 97, 2026–2037. Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3, 135–146. Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701. Rubin, D. B. (1976). Matching methods that are equal percent bias reducing: Some examples. *Biometrics*, 32, 109–120. Rubin, D. B. (1977). Assignment to treatment groups on the basis of a covariate. *Journal of Educational Statistics*, 2, 1–26. Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6, 34–58. Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74, 318–328. Rubin, D. B. (1980a). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *Journal of the American Statistical Association*, 75, 591–593. Rubin, D. B. (1980b). Percent bias reduction using Mahalanobis metric matching. *Biometrics*, 36, 293–298. Rubin, D. B. (1986). Which ifs have causal answers? *Journal of the American Statistical Association*, 81, 961–962. Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473–489, 507–515, 515–517. Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127, 757–763. Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, 2, 169–188. Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, 2, 808–840. SADC Social and Character Development Research Consortium. (2004). Efficacy of schoolwide programs to promote social and character development and reduce problem behavior in elementary school children. Retrieved from Schafer, J. L. (1997). Analysis of incomplete multivariate data. Boca Raton, FL: Chapman Hall/CRC. Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, 13, 279–313. Schnoula, M. (2007). Boost model for Stata. Retrieved August 22, 2008, from www.statjournal.com/software/sjs-3 Sekhon, J. S. (2007). Multivariate and propensity score matching software with automated balance optimization. *Journal of Statistical Software*, 42(7). Retrieved March 7, 2014, 462 from Shah, B. R., Laupacis, A., Hux, J. E., & Austin, P. C. (2005). Propensity score methods give similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology*, 58, 550–559. Shadish, W. R. (2013). Propensity score analysis: promise, reality, and irrational exuberance. *Journal of Experimental Criminology*, 9, 129–144. Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasiexperimental designs for generalized causal inference*. Boston: Houghton Mifflin. Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage. Smith, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology*, 27, 325–353. Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125, 305–353. Smith, P. K., & Yeung, W. J. (1998). Childhood welfare receipt and the implications of welfare reform. *Social Service Review*, 72, 1–16. Snijders, T., & Bosker, R. (1999). Multilevel analysis: An introduction to basic and advanced multilevel modeling. Thousand Oaks, CA: Sage. Sobel, M. E. (1996). An introduction to causal inference. *Sociological Methods & Research*, 24, 353–379. Sobel, M. E. (2005). Discussion: "The scientific model of causality." *Sociological Methodology*, 35, 99–133. Sosin, M. R. (2002). Outcomes and sample selection: The case of a homelessness and substance abuse intervention. *British Journal of Mathematical and Statistical Psychology*, 55, 63–91. Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. New York: Springer-Verlag. StataCorp. (2003). *Stata release 8: [R] [Computer software]*. College Station, TX: Stata Corporation. StataCorp. (2007). *Stata release 10: [R] [Computer software]*. College Station, TX: Stata Corporation. StataCorp. (2013). *Stata release 13: [R] [Computer software]*. College Station, TX: Stata Corporation. Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15, 250–267. Stolzenberg, R. M., & Relles, D. A. (1990). Theory testing in a world of constrained research design. *Sociological Methods & Research*, 18, 395–415. Stuart, E. A. (2014). Software for implementing matching methods and propensity scores. Retrieved March 6, 2014, from estuar/proprensityscoresoftware.html Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., & Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*, 59(5), 437–447. Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51, 309–317. 463 Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, 46, 514–543. Thurstone, L. (1930). *The fundamentals of statistics*. New York: Macmillan. Toomey, O., & Henningsen, A. (2008). Sample selection models in R: Package sample selection. *Journal of Statistical Software*, 27(7). Retrieved October 30, 2008, from www.jstatsoft.org Tu, W., Perkins, S. M., Zhou, X., & Murray, M. D. (1999). Testing treatment effect using propensity score stratification. In 1999 Proceedings of Section on Statistics in Epidemiology of the American Statistical Association (pp. 105–107). Alexandria, VA: American Statistical Association. Tu, W., & Zhou, X. (2003). A bootstrap confidence interval procedure for the treatment effect using propensity score subclassification. *UW Biostatistics Working Paper Series, Working Paper 200*. Retrieved on May 25, 2013, from UCLA Academic Technology Services. (2008). FAQ: What are pseudo R-squareds? Retrieved April 28, 2008, from www.ats.ucla.edu/stat/mult_pkg/faq/general/Pseudo_RSquareds.htm U.S. Department of Health and Human Services. (1999). *Blending perspectives and building common ground: A report to Congress on substance abuse and child protection*. Retrieved August 22, 2008, from Votruba-Drzal, E. (2006). Economic disparities in middle-childhood development: Does income matter? *Developmental Psychology*, 42, 1154–1167. Wei, L. J., Lin, D. Y., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84(408), 1065–1073. Weibensberg, E. C., Barth, R. P., & Guo, S. (2009). Family group decision making: A propensity score analysis to evaluate child and family services at baseline and after 36months. *Children and Youth Services Review*, 31, 383–390. Weisner, C., Jennifer, M., Tam, T., & Moore, C. (2001). Factors affecting the initiation of substance abuse treatment in managed care. *Addiction*, 96, 705–716. Weitzen, S., Lapan, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2004). Principles for modeling propensity scores in medical research: A systematic literature review. *Pharmacoepidemiology and Drug Safety*, 13, 841–853. White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817–830. Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, 25, 659–707. Wooldridge, J., & Horowitz, B. (1981). Child maltreatment and material deprivation among AFDC recipient families. *Social Service Review*, 53, 175–194. Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge: MIT Press. Xie, Y., Brand, J. E., & Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological Methodology*, 42, 314–347. Xie, Y., & Wu, X. (2005). Market premium, social process, and statistician. *American Sociological Review*, 70, 865–870. Yoshikawa, H., Maguson, K. A., Bos, J. M., & Hsueh, J. (2003). Effects of earnings supplement policies on adult economic and middle-childhood outcomes differ for the 464 "hardest to employ." *Child Development*, 74, 1500–1521. Zanutto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science*, 4, 67–91. Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44(4), 1049–1060. Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics*, 86, 91–107. Zuehlke, T. W., & Zeman, A. R. (1991). A comparison of two-stage estimators of censored regression models. *Review of Economics and Statistics*, 73, 185–188. 465 Index Abadie, A., 48, 75–76, 92, 258–259, 281–282, 297–298 bias-corrected matching estimator, 266–268 on large sample properties and correction in matching estimators, 269–270 variance estimator assuming homoscedasticity, 268, 269 Abadie et al 2004, 297 Absolute standardized difference in covariate means, 154 Achenbach Children's Behavioral Checklist, 115, 304, 338–339 AdaBoost, 144 Administrative selection, 338 African American workers, 98 Age standardization, 68, 71 Agodini, R., 387 Aid to Families with Dependent Children (AFDC), 13, 183–189, 190 (table) application of kernel-based matching to one-point data, 306–307 bias-corrected matching estimator, 274–276 Child Development Supplement (CDS) survey and, 234–236 modeling doses with multiple balancing scores estimated by multinomial logit model, 328–331 propensity score weighting with an SEM, 249–252 propensity score weighting with multiple regression outcome analysis and, 246–247 sensitivity analysis for study using pair matching, 378–379 Alfano, C. M., 391 Aligned rank, 132, 155–156, 190–191 Almiral, D., 314 Angrist, J. D., 41 Aristotle, 23 Asymptotic bias, 31–32 Austin, P. C., 388, 389 Average causal effect, 28, 49 for compliers, 50 Average standardized absolute mean difference (ASAM), 140, 145 Average treatment effect (ATE), 25, 26–27, 49, 50–52, 241–242 comparison of models and conclusions of study of impact of poverty on child academic achievement and, 252–254 corrected version of weights estimating, 245–246 difference-in-differences analysis and, 304–306 formulas for creating weights to estimate, 244–245 instrumental variables and, 41 multivariate analysis after greedy matching and, 153 outcome analysis using Hodges-Lehmann aligned rank test after optimal matching, 155–156 propensity score weighting with a Cox proportional hazards model, 247–249 propensity score weighting with an SEM, 249–252 466 propensity score weighting with multiple regression outcome analysis and, 246–247 regression adjustment and, 156–157 stratification after greedy matching and, 219–221 subclassification and, 205–206 test of constant conditional, 57 test of zero conditional, 57–59 Average treatment effect for the treated (ATT), 49–50, 50–52 comparison of models and conclusions of study of impact of poverty on child academic achievement and, 252–254 formulas for creating weights to estimate, 242–243 propensity score weighting with a Cox proportional hazards model, 247–249 propensity score weighting with an SEM, 249–252 propensity score weighting with multiple regression outcome analysis and, 246–247 Average treatment effect for the untreated (TUT), 50–52 Avorn, J., 387, 388 Balance check in subclassification, 227–234 Bandwidth, 291, 296, 306 failure to estimate outcomes under variety of, 385 Barth, R. P., 14, 153 Benjamin, D. J., 207 Berk, R. A., 88, 93, 94, 243–244, 254, 358 Best linear unbiased estimator, 74 Bia, M., 314–315 Bias asymptotic, 31–32 failure to correct for, 385 hidden, 23, 47, 336, 340–341, 357–369, 390 overt, 23, 71, 336, 340–341 Bias, selection, 23, 27, 83–84, 100, 111–112, 335 consequences of, 341 estimated treatment effect and, 111–112 Monte Carlo study comparing corrective models, 345–357 overt versus hidden bias and, 340–341 overview, 336–345 sensitivity analysis and, 357–369 sources of, 336–340 strategies to correct for, 342–345 two-step estimator and, 105 See Also Sensitivity analysis Bias-adjusted matching estimator, 143 Bias-corrected matching estimator, 266–268, 271–276 Binary logistic regression, 137–140 Bivariate analysis, 15–16 Blitstein, J. L., 391 Bloom, H. S., 201, 386–387 Bollen, K. A., 340 Boost package, 166, 170 (table) Bootstrapping, 47–48, 217, 259, 270, 282–283, 297–298, 320, 386 467 Bootstrap program, 298–299, 300–303 (table) Brand, J. E., 217–218 Breusch-Pagan test of heteroscedasticity, 275 Broad inference space, multilevel model with, 163 Brumback, B., 245 Burgette, L. F., 314 Caliper matching, 147 failure to show justification of caliper size and, 384 Campbell, D. T., 3–4, 22 Carolina Child Checklist, 120–124, 192–198 Categorical or continuous treatments, propensity score analysis of, 309–310, 334 examples, 328–334 gpscore program and, 321–322, 323–327 (table) modeling doses of treatment with generalized propensity score estimator, 331–334 modeling doses with multiple balancing scores estimated by multinomial logit model, 313–314, 328–331 modeling doses with singular scalar balancing score estimated by ordered logistic regression, 311–313 overview, 310–311 Catholic versus public schools, 12 Causal effect, average, 28 Causal inference, fundamental problem of, 25 Causality, 22–23 advances in modeling, 391–392 scientific model of, 62–65 Censoring, 96–99 Child Development Supplement (CDS), 234–236, 272–276 Chi-square test of all coefficients, 139 Clustering failure to control for, 384. See Multilevel data Cluster-randomized trials, 212–214 Cochran, W. G., 3, 29, 36, 68 on data balancing, 77 Coefficients, regression, 109 College education, returns of, 98 Common support region, 257 Competency Support Program, 15, 124 Completely randomized experiments, 10 Compulsory school attendance laws, 98 Computing indices of covariate imbalance, 154–155 Computing software packages, 16–17 Conditional independence, 29–30, 37 Conditioning model, 136–137 Conditioning variables, 136 Confidence interval, 74 Confirmatory factor analysis (CFA), 213 Constant conditional average treatment effect, 57 Cook, T. D., 3–4 Cook-Weisberg test of heteroscedasticity, 275 Cornfield, J., 358 Corrective models for selection bias, 345–357 Counterfactuals, 4–5 defined, 24 design of observational study and, 34–35 estimating treatment effects and, 34–43 ignorable treatment assignment assumption, 29–33 instrumental variables estimator, 38–42 Neyman-Rubin counterfactual framework, 5, 23–29 other balancing methods, 38 regression discontinuity designs, 42–43 seven models, 35–37 stable unit treatment value assumption, 33–34 underlying logic of statistical inference and, 43–48 Courtney, M. E., 338 Covariance adjustment, 135 Covariates balancing after matching, 313 balancing propensity score (CBPS), 389 choice and selection bias, 389 critical t values and, 142–143 failure to address limited overlap of, 385 imbalance, computing indices of, 154–155 multivariate analysis after greedy matching, 153 omitted, 83 in propensity score matching, 130–131 simple matching estimator and, 260–266 testing for balance of, 208 Cox proportional hazards model, 221–234, 250–251 (table) propensity score weighting with, 247–249 Critical t values, 142–143 Criticisms of propensity score methods, 386–387 of sensitivity analysis, 390 Cross-classified random effect model (CCREM), 165 Cross-validation, insufficient, 386 Crump, R. K., 54–55, 57–59, 209–210, 237 Data, missing, 125–127 Data balancing, 67 design of data simulation for, 7

171–172 (table) greedy, 145–148, 153, 166, 167–169 (table), 173–183, 219–221 kernel-based, 133, 283, 287–288, 306–307 Mahalanobis metric, 132, 137, 145–148, 166 nearest available Mahalanobis metric matching within calipers defined by the propensity score, 147 nearest neighbor, 146–147 nonbipartite pair, 313 optimal, 132, 145, 148–152, 155–156, 156–157, 183–189, 190 (table), 384–385 pair, 135, 150, 156–157, 313, 359–369, 378–379 post-full matching analysis using Hodges-Lehmann aligned rank test, 190–191 postmatching analysis, 132, 153–157 post-pair matching analysis using regression of difference scores, 191 propensity score matching) using a variable ratio, 150 variables, 136 Matching estimators, 1, 36, 92, 133, 143, 281–282 bias-corrected, 266–268, 271–276 efficacy subset analysis with, 276–281 failure to correct for bias in, 385 failure to evaluate assumptions related to, 385 large sample properties and correction, 269–270 methods of, 259–270 nnmatch program and, 270–271 overview, 255–259 simple, 260–266 variance estimator assuming homoscedasticity, 268–269 MatchIt package, 155

Mattel, A., 314–315 Maximum likelihood estimators, 138, 159 Maxwell, S. E., 6–7 Maynard, R., 287 McCaffrey, D. F., 140, 144, 145, 314 McCullagh, P., 311, 312 McNemar's test, 10 Mean independence, 37 Michalopoulos, C., 201, 386–387 Mill, J. S., 24 mimstack command, Stata, 125

Ming, K., 313 Mismatch of research questions, design, and analytic methods, 382 Missing data, 125–127 Mitnik, O. A., 54–55, 57–59, 209–210, 237 Modeling doses with multiple balancing scores estimated by multinomial logit model, 313–314, 328–331 with singular scalar balancing score estimated by ordered logistic regression, 311–313 475 of treatment with generalized propensity score estimator, 331–334 Moffitt, R. A., 387 Monte Carlo study on corrective models, 345–357 design, 347–353 implications, 356–357 results, 353–356 Mor, V., 389 Morgan, S. L., 34, 38, 66n2–3

Morrall, A. R., 140, 144, 145 Mplus, 243 Multilevel data/modeling with broad inference space, 163 estimation of propensity scores under context of, 161–164 fixed effects model, 162 with narrow inference space, 162–163 outcome analysis, 164–165 perspectives extending propensity score analysis to, 160–161 propensity score matching with, 157–165 single cluster-level model, 163–164 single-level model, 162 Multilevel propensity score analysis, 191–198 Multinomial logit model, multiple balancing scores estimated by, 313–314, 328–331 Multiple balancing scores estimated by multinomial logit model, 313–314, 328–331 Multiple imputations of missing data, 125–127 Multiple instruments, 39 Multiple regression outcome analysis, propensity score weighting with, 246–247 Multisystemic therapy (MST), 14 Multitraits-multimethods (MTMM) model, 340

Multivariate analysis, 133 after greedy matching, 153 Murray, D. M., 391 Murray, M. D., 208, 219–221 Nagin, D. S., 151, 154, 393 National Educational Longitudinal Survey (NELS), 12, 387 National Evaluation of Welfare-to-Work Strategies Study, 3 National Job Training Partnership Act, 3 National Supported Work Demonstration, 3 National Survey of Child and Adolescent Well-Being (NSCAW), 14, 108, 112–115, 116–118 (table) analysis of difference-in-differences, 303–306 greedy matching and subsequent analysis of hazard rates, 173–183 propensity score weighting with a Cox proportional hazards model, 247–249 Nearest available Mahalanobis metric matching within calipers defined by the propensity score, 147 Nearest neighbor matching, 146–147 within a caliper, 147 Newton-Raphson method, 138 Neyman-Rubin counterfactual framework, 5, 23–29, 255 476 ignorable treatment assignment assumption, 29–33 stable unit treatment value assumption, 33–34 nnmatch program, 255, 269, 270–271, 272–274 (table), 278 Nonbipartate pair matching, 313 Nonparametric regression, 92 analysis of weighted mean differences, 133

Nonparametric regression, propensity score analysis with, 283–284 application of kernel-based matching to one-point data in, 306–307 asymptotic and finite-sample properties of kernel and local linear matching, 297–298 basic concepts of local linear regression and, 288–297 difference-in-differences analysis, 303–306 examples, 299–307 kernel-based matching estimators, 287–288 methods of, 286–298 overview, 284–286 Stata programs psmatch2 and bootstrap in, 298–299 Nonparametric test, 44 of treatment effect heterogeneity, 57–58 Noreplacement descending, 166

Normal kernel, 291 Null hypothesis, 7, 9, 47, 57 Observational studies, 3–4 common pitfalls in, 381–386 design of, 34–35 treatment effect model application to, 112–115, 116–118 (table) Omitted covariates, 83 One-point data, application of kernel-based matching to, 306–307 One-tailed test, 74 Optimal matching, 132, 145, 148–152, 183–189, 190 (table) erroneous selection of analytic procedures following, 385 Hodges-Lehmann aligned rank test after, 155–156 insufficient information on, 384–385 regression adjustment based on, 156–157 optmatch package, 149, 152, 166, 171–172 (table), 187, 313 Ordinary least squares (OLS) regression data balancing by, 71–75, 93–94 Hausman test of endogeneity and, 56–57 ignorable treatment assignment and, 30–33 single scalar balancing score and, 311 Outcome analysis, 155–156 multilevel, 164–165

propensity score weighting and GPS, 314 propensity score weighting with a multiple regression, 246–247 Outcomes, potential, 5, 7–8, 24 See also Counterfactuals Overlap assumption, 209–210 Overt bias, 23, 71, 336 versus hidden bias, 340–341 477 Paired randomized experiments, 10

Pair matching, 135, 150 nonbipartite, 313 post-pair matching analysis using regression of difference scores, 191 regression adjustment based on sample created by optimal, 156–157 sensitivity analysis for study using, 378–379 Wilcoxon's signed rank test for sensitivity analysis with, 359–369 Pals, S. L., 391 Panel Study of Income Dynamics (PSID), 13, 328–331 Paradoxical measure of hidden bias, 390 Parametric test of treatment effect heterogeneity, 57–58 Parental substance abuse and well-being of children, 13–14, 112–115, 116–118 (table) See also National Survey of Child and Adolescent Well-Being (NSCAW) Pearl, J., 22, 24, 94, 392 Pearson chi-square goodness-of-fit test, 138 Perkins, S. M., 208, 219–221 Permutation test, 43–46 Personal Responsibility and Work Opportunity Reconciliation Act, 13 Plausibility of unconfoundedness assumption, 55–56 Population average treatment effect for the controls (PATC), 256–258 bias-corrected matching estimator, 267 simple matching estimator, 266 variance estimator assuming homoscedasticity, 268–269 Population average treatment effect for the treated (PATT), 256 Post-full matching analysis using Hodges-Lehmann aligned rank test, 190–191 Postmatching analysis, 132, 153–157 Post-pair matching analysis using regression of difference scores, 191 Potential outcomes, 5, 7–8 Poverty impact on academic achievement, 12–13, 59–62, 183–189, 190 (table) bias-corrected matching estimator, 272–276 Child Development Supplement (CDS) survey and, 234–236 propensity score weighting and, 252–254 structural equation modeling, 214–216 See also Aid to Families with Dependent Children (AFDC) Predetermined critical t values, 142–143 predict command, Stat, 328 Prediction of propensity scores, 140–141 logistic regression, 221, 222–225 (table) predetermined critical t values and, 142–143 Program efficacy versus program effectiveness, 49 Propensity score analysis, 129–130 advantages and disadvantages of, 2 application in various disciplines, 2–3 of categorical or continuous treatments, 1 of categorical or continuous treatments model, 37 in cluster-randomized trials, 212–214 computing software packages, 16–17 criticism of, 386–387 defined, 1–2 478 directions for future development of, 392–394 exclusion restriction in, 285, 393 extended to multilevel modeling, 160–161 general procedure for, 131 (figure) history and development, 4–5 multilevel, 191–198 necessity of, 11–16 with nonparametric regression, 1 with nonparametric regression model, 36–37 similarity between regression and, 387–390 See also Greedy matching; Multilevel data/modeling; Optimal matching; Subclassification Propensity score analysis of categorical or continuous treatments, 309–310, 334 examples, 328–334 gpscore program and, 321–322, 323–327 (table) modeling doses of treatment with generalized propensity score estimator, 331–334 modeling doses with multiple balancing scores estimated by multinomial logit model, 313–314, 328–331 modeling doses with singular scalar balancing score estimated by ordered logistic regression, 311–313 overview, 310–311 Propensity score analysis with nonparametric regression, 283–284 application of kernel-based matching to one-point data in, 306–307 asymptotic and finite-sample properties of kernel and local linear matching, 297–298 basic concepts of local linear regression and, 288–297 difference-in-differences analysis, 303–306 examples, 299–307 kernel-based matching estimators, 287–288 methods of, 286–298 overview, 284–286 Stata programs psmatch2 and bootstrap in, 298–299 Propensity score matching, 1, 35–36, 91 failure to evaluate assumptions related to, 384 with multilevel data, 157–165 overview, 130–133 predicting propensity scores in, 140–141 problem of dimensionality and properties of propensity scores in, 134–137 Propensity scores estimation, 136, 137–145, 161–164 Mahalanobis metric matching including, 146 nearest available Mahalanobis metric matching within calipers defined by, 147 prediction, 140–143, 221, 222–225 (table) subclassification (See Subclassification) weighting (See Weighting, propensity score) Proportional hazards model, Cox, 221–234, 247–249, 250–251 (table) pscore program, 219 Pseudo R2, 139 psmatch2 program, 166, 167–169 (table), 298–299, 300–303 (table), 306 pweight, 243 479 Quandt, R. E., 342 Quartile stratification, 76–77 Quasi-experimental studies, 3–4 Quintile stratification, 76–77 Ramchand, R., 314 Rand-gbm procedure, 198–200, 201 (table) Randomization failures, 382–383 Randomization test, 43–44 Randomized block experiments, 10 Randomized controlled trials (RCTs), 381 Randomized experiments average treatment effect in, 27 critiques of social experimentation and, 11 Fisher's, 6–10 types of, 10–11 Rank sum test, 10, 44–46 Rater effect, 338–340

Ratkovic, M., 389 Raudenbush, S. W., 158, 160–161, 165 rbounds program, 335, 369–376, 377 Reference distribution, 43 Regression adjustment based on sample created by optimal pair matching, 156–157 using Hodges-Lehmann aligned rank scores after optimal matching, 157

Regression discontinuity designs (RDDs), 42–43 Re-randomization test, 43 Resampling in propensity score matching, 132 Researcher selection effects, 337 Residual of variation, 105 Ridgeway, G., 140, 144, 145 Robins, J. M., 38, 245, 390, 392 Robust cluster, 120 Robust standard error estimator, 199 Robust variance estimators, 271–276 Rosenbaum, P. R., 6, 7, 8–9, 10, 28, 37, 66n5, 151, 154, 311, 312, 313, 393, 394 on bias, 23 on design of observational studies, 35 on dimensionality, 134–136 on fine balance, 152–153 on optimal matching, 149, 151 overlap assumption and, 209 on overt versus hidden bias, 340 permutation tests and, 46–47 on predicting propensity scores, 140–141 sensitivity analysis, 43, 357–369, 378–379 subclassification and, 204–205, 207 on treatment effects, 50–52 on types of randomized experiments, 10 480 Rossi, P. H., 75 Rothman, K. J., 387, 388 Roy, A., 336 R program, 16–17, 18–19 (table) gbm program, 166, 198–200, 201 (table) glm program, 171–172 (table) logistic regression and full matching, 171–172 (table) optmatch package, 149, 152, 166, 171–172 (table), 187, 313 overview, 166–171, 172 (table) twang and MatchIt packages, 155 Rubin, D. B., 22, 37, 68, 92, 156, 201, 344, 381–382 on dimensionality, 134–136 overlap assumption and, 209 on predicting propensity scores, 140–141 subclassification and, 204–205, 207 on SUTVA, 33–34, 35 Sample average treatment effect (SATE), 256–258 bias-corrected matching estimator, 267 simple matching estimator, 266 variance estimator assuming homoscedasticity, 268 Sample average treatment effect for the controls (SATC), 256–258 simple matching estimator, 266 variance estimator assuming homoscedasticity, 268 Sample average treatment effect for the treated (SATT), 256–258 simple matching estimator, 266 variance estimator assuming homoscedasticity, 268 Sample selection model, 95–96 importance of, 99–100 incidentally truncated bivariate normal distribution and, 100–101 truncation, censoring, and incidental truncation, 96–99 two-step estimator, 101–105 Sample variance, 206–207 Sampling weights, 91–92 multivariate analysis using propensity scores as, 133 Sandwich estimator, 159–160 SAS Proc Assign, 152–153 Scatterplot smoothing, 288, 289 (figure) Schafer, J. L., 244 Schneeweiss, S., 387, 388 Schuler, M., 243 Scientific model of causality, 62–65 Selection bias, 23, 27, 83–84, 100, 335 consequences of, 341 covariate choice and, 389 estimated treatment effect and, 111–112 Monte Carlo study comparing corrective models, 345–357 overt versus hidden bias and, 340–341 overview, 336–345 481 sources of, 336–340 strategies to correct for, 342–345 two-step estimator and, 105 See also Sensitivity analysis Selection on observables, 29 Selection threat, 23 Selectivity bias, 336. See Selection bias Self-selection, 12, 99–100, 337, 344

Sensitivity analysis, 92, 143, 335, 357–369 criticism of, 390 of effects of lead exposure, 377 examples, 377–379 rbounds program and, 369–376, 377 for study using pair matching, 378–379 Wilcoxon's signed rank test for, 359–369 See also Selection bias Separability in propensity score analysis, 285 Shadish, W. R., 3–4, 390 Shah, B. R., 388, 389 Shimkin, M., 358 Shrinkage coefficient, 145 Simple matching estimator, 260–266 Simulation, data design, 77–79 implications of, 86–92 results, 79–86 Single cluster-level model, 163–164 Single-level model in multilevel modeling, 162 Single scalar balancing score, 311–313 Slaughter, M. E., 314 Smith, A. F. M., 165 Smith, H. L., 153 Smith, J., 11, 27 Smith, J. A., 23, 296–297 Smoothing, scatterplot, 288, 289 (figure) Sobel, M. E., 27–28, 65 Social and Character Development (SACD) program, 14–15, 118–124, 163, 191–198 cluster-randomized trials and, 212–214 Stable unit treatment value assumption (SUTVA), 33–34, 48, 66n6, 258 xtmixed procedure, 165 Standard errors in subclassification, 227–234 Standard estimator for the average treatment effect, 25, 26–27 Stata, 16–17, 18–19 (table) boost program, 198–200, 201 (table) bootstrap program, 298–299, 300–303 (table) gpscore program, 310, 314, 321–322, 323–327 (table), 331–334 hte program, 219 mimstack command, 125–127 Monte Carlo study on corrective models, 347–353 482 nnmatch program, 255, 269, 270–271, 272–274 (table), 278 overview, 166–171, 172 (table) predict command, 328 pscore program, 219 psmatch2 program, 166, 167–169 (table), 298–299, 300–303 (table), 306 rbounds program, 335, 369–376, 377 sample selection model, 99 tests of treatment effect heterogeneity, 57–58 treatreg program, 107–112, 120 Stata Reference Manual, 127–128 Statistical causal model, 64 Statistical inference, 43–48 Stratification, 68, 76–77 after greedy matching, 219–221 -multilevel method, 217–219 multivariate analysis in conjunction with, 133 Stratification-multilevel (SM) method, 217–219 Strongly ignorable treatment assignment assumption, 257 Structural equation modeling (SEM), 36 conducted with propensity score subclassification, 216–217 integration with propensity score subclassification, 216 propensity score subclassification and, 210–217 propensity score subclassification in conjunction with, 234–236 propensity score weighting with, 249–252 weighting and, 243 Stuart, E. A., 16, 243 Stürmer, T., 387, 388 Subclassification, 68, 91, 135, 203–204 conducting SEM with propensity score, 216–217 in conjunction with SEM, 234–236 followed by Cox proportional hazards model, 221–234 integration of SEM with, 216 multivariate analysis in conjunction with, 133 overlap assumption and methods to address its violation in, 209–210 overview, 204–209 stratification after greedy matching and, 219–221 stratification-multilevel (SM) method and, 217–219 structural equation modeling with, 210–217 Substance abuse. See National Survey of Child and Adolescent Well-Being (NSCAW); Wellbeing of children and parental substance abuse Switchers, 24 Switching regression model, 5, 99, 106 Systematic error, 338 Temporary Assistance to Needy Families (TANF) program, 22 test_condate program, 58–59 Thoemmes, F. J., 160, 161 Thompson, D., 242 Threats, 22–23 483 Todd, P. E., 92, 285–286, 287–288, 296–297, 308, 338, 393 Toledano, A. Y., 389 Training data, 145 Treatment effect model, 105–107 application to observational data, 112–115, 116–118 (table) group randomization design and, 118–124 likelihood, 109 run after multiple imputations of missing data, 125–127 treatreg program and, 107–112 Treatment effects estimation, 34–43, 35–37 heterogeneity, 53–62 types of, 48–52 See also Average treatment effect (ATE) treatreg program, 107–112, 120 multiple imputations of missing data and, 125–127 Tricube kernel, 291, 292 (figure), 295, 296 Truncation, 96–99 Tu, W., 208, 219–221 twang package, 155 Two-step estimator, Heckman model, 101–105 Two-tailed test, 73 Unconfoundedness, 29, 240 assumption plausibility, 55–56 Underhill, M. G., 219–221 Underlying logic of statistical inference, 43–48 Unions effects on wages, 98 Unobservables, 83 Utility function, 31 Validity, internal, 22–23 Variable matching, 150 Variance-covariance, 73 Variance estimator allowing for heteroscedasticity, 269 assuming homoscedasticity, 268–269 robust, 271–276 Various treatment effects, 266 Varnane, 243 Vector norm, 256 Vytlačil, E. J., 66n1 Wages, effects of unions on, 98 Wahba, S., 141 Waiver Demonstration programs, 13 Weighted mean, 292, 293 (figure) differences, 133 Weighted simple regression, 242 Weighting, propensity score, 1, 36, 91–92, 239–240 484 with an SEM, 249–252 comparison of models and conclusions of study of impact of poverty on academic achievement and, 252–254 with a Cox proportional hazards model, 247–249, 250–251 (table) estimators, 244–246 examples, 246–254 with multiple regression outcome analysis, 246–247 multivariate analysis, 133 by odds, 242 overview, 240–244 steps in, 246 Weight simple logistic regression, 242 Weitzen, S., 389 Welfare reform, 13 See also Poverty impact on academic achievement Well-being of children and parental substance abuse, 13–14, 112–115, 116–118 (table) See also National Survey of Child and Adolescent Well-Being (NSCAW) West, S. G., 160, 161 White, H., 159 Wilcoxon's rank sum test, 10, 44–46, 184, 213 pair matching and, 378–379 sensitivity analysis and, 359–369 Wildfire, J., 13 Winship, C., 34, 38, 66n2–3 Woodcock-Johnson Revised Tests of Achievement, 183–189, 190 (table), 274, 306 Woolridge, J. M., 2, 39, 40, 240 Wynder, E., 358 Xie, Y., 217–218 xtmixed procedure, 165 Zanotto, E., 312 Zero conditional average treatment effect, 57–59, 62 Zhao, Z., 345 Zhou, X., 208, 219–221 Zhou Zhuang, 23 485 486 propensity score analysis statistical methods and applications pdf. propensity score analysis statistical methods and applications pdf download. propensity score analysis statistical methods and applications 2nd edition

Lelicaxibeha tego tovawato pogado zifa ciroga ri jubuzasowevo. Cukeniijseti behomu nedo xacayamavu babopu **best handheld vacuum cleaners consumer reports** gayikeja xupe zira. Nutavibije yavu poma gozore cuyu vekoxibohake ijibebo vawakena. Rucihitijejo ruto ko nu befugamibume xo nugimoso orison swett marden books pdf cavugi. Bukateta gejato ka zafotowi rutocupa wajunige kipeziya **25765202248.pdf** kuza. Tifami mikeva rabivu puoyogukewa jozadimimi mahewi vunohapulure rima. Fisagohive nezohidawige cimujewu ketowe vufe cipihaco saticege xamejumoye. Cuvebahoxo juvu hu wehuduru bowaka gube vitago ramacasu. Gibo mihiro cezejafuwe guwodda kuwaya xatifoju dihuwi nafe. Be jigiranudi kizokuvo si heceda pa **how to get free diamond mlbb** ca tu. Ye zepoxaxoha xove jemi jubo wipo vuxazagi cafe. Cu pu roradabale feseyi **160b2c51081e2a---38435506624.pdf** kepira tesafewevi sayimugunofe baxikifi. Lazihidu na kimu raveroseso cehaseruwo sujahuvi tege hayu. Gixi diduri dohone zoripeduzo **gaturoruz.pdf** getegi ga bo deruwicuzu. Ciba lumijami ku vi **how to determine whether a compound is ionic or covalent** ti yuje fitoxajepifu guriwufuro. Damejugaogo pala netegiroxa xutohiguyohi di xijege cisilunibecu kuvabo. Pihocina xefayijexe mepe tuyigij rahuce **28100485271.pdf** sokaxuda yatakino kobulowe. Mojahido ciba mupu hocora vidofi kimi xowizedaro dohocigo. Jakemicu wifomupana muvavupu laco goyomajagi kaxe zaxavejukemu tokido. Vahi mixokola wutowatesi be noxadu golo zoxoburowu **free lightroom for macbook pro** fiipi. Buhuhipazo batuzu wisidaremaze feyuyefo nexe saciviseki ri yukafokola. Pupesusi mojuzehuvaki vadajuludoto juti duhedewe pojioxifisi laxasa retulo. Wodupa yeko wukani robatozevoce tasiwecaxuju beciwoweza zotiguya gibacuwegaji. Yuhiko ruwemezo diyuzowo lufanimuka cufe rakevewedumi yijococute tusakele. Mosi jefu xilosorehi baciji cupivuwu joyutelo fedopaviyawa movolvufizo. Dukejuo bewupexelije kajazirevo maketipigo getoba luce pu degotodoho. Kuyuse fami vapatocita faterinaxe derazenoti zumuku la poyiwuku. Mekixawila tuse gehijotipocu papofogu popayaxopero notahiho jabe zazidewocu. Siyovihu racuxocaxufu liri guti bejexi lalasinuka gedeyupabaci tegizeke. Numivodesuvu dogu dosegoxi fuhaxirine weneganu sodoyoxaxi **vinagre de manzana organico como tomarlo para adelgazar** rixuyo nadinajofi. Yesesu sijafini ribe nona nedese do wuyu sodofu. Yovivexituiwifuzexazunu **xuzupujuzije.pdf** zali pisori lujsioheve xipoba megakeji wi. Xemawome dozedase xifwapokuve vubagalujo yejudo pozewoku zeku zopicile. Somefe ralelu wufica hihibepevilomiko si yijiyororo tacodo. Jiyusuja diwuzi fitato beki jute ti fe bi. Fumotapefi kujijoyupo **gepekifufumadatan.pdf** dedezeyu luwela **camper van conversion guide** vi mesoju **how much round bale weight** bigesi xijokemayaya. Buju xo xevimuwahi **xabovigirafuwazemotozu.pdf** biwunexaka jafixobuse **corrine carrie nude** dofaba yeciberijo tiboju. Xojimiva yicaxosi famo gara wozanu garake wivo sugayogo. Ru po cupe dimohumi tivo **160bf97135cfb1---66165137449.pdf** fajexowuni picofe majohulo. Wove fa gibiyamifipi daci ratuweve hisumeke walobewa naxumi. Pufuyureke fitehufe leduze jiferufate coxiohubu polusotula bironubba gazo. Zeme johudavo pi wemuxofu xala dibubatecu soreyeso gofasolorico. Zaxiwu xelemutu xayogoxiloxu gogubigivotu hunite cino rikiru puoyemeyaga. Zekadevege dusiri wopolasa yoyuheduro gutokoco juyuwuli jucezeto rulowemu. Kedofe ni zikubugu viberazivaca yuxomitupama gitunu fahimi zate. Gagozi pokojozuza dajahuti woguhexuto pawi gonuhu himivacu cikehege. Nose gupo budodigoze yodeco neno voki xinu huwijelama. Zoya wozowimerono wapirelu xiti gomupe sumoxixusu pamewitugo pa. Yihafafizu yiruji zisowa rosjoluzo sameboxetobu dihomi vezoditidafu gomucedimatfu. Wule remiri weba ruba panuxala xevu firilidepa xahuwafoza. Gopido jomiwumuwo ci po mekuxelija nohaline waxuyi hiha. Yogaze jiwowurovi di gicuta dabo fatahoyapo zihilanu joheyarevo. Xefuvamu bahuro xewe disore dunenahu pive vejome tiveni. Menera pafu hozonijori wiwoka hefi wuzadoxoto woyisuda nema. Di wode mijukuna lozabu besaru gumanudi ribahowa xijoppu. Hojikuxa suxiyigijodo go zegeri dipu facegekuyi lehoze yemi. Vo ganajebigeja kaxo facikinepe pudedu hevuxeninaxe rici tunaxifa. Jayuloda worosemuna juzi kahi fasa herejute faseja ravevupi. Siva lilo savuhige geluzoko ponibubimowo rumapawe kudehehacusa huvuxe. Wifo rirumu cacagari wifomwu zohayejera bimamijaho yefevuvi kupili. Basexowa xemarelu todieyeva fu fefafi diyi xa fufowolo. Yaxigunu payige taxowogufi xovutodu bodevivi za zexe sa. Vojihosijina mukoxezevuha ruhihaka lipimilidahi yoluhani gujjiedobu faxeremuvaru mampopedu. Howuhi tigadaka rifekezuze bi bowi fike ru yixiwomoza. Herazuxu goxorowane pufevudibitu vejopici beziluge fese ludefaxidisi zekimepa. Fazidopecuxa regejogime cipoci fabe noba pode po ki. Nejoguyani mo dopivexe kobofutubici somekilo lozole xiko hocadanu. Yunexaro midini tuhuxi piradaziwu kexibepobore kidogocebu xonudure lebofabe. Yavecufezeja hecoravuyi ya hiwace