Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives
Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives

An essential journey with Donald Rubin’s statistical family

Edited by

Andrew Gelman
Department of Statistics, Columbia University, USA

Xiao-Li Meng
Department of Statistics, Harvard University, USA
## Contents

**Preface**  xiii  

**I Casual inference and observational studies**  1  

1 An overview of methods for causal inference from observational studies, by Sander Greenland  3  
1.1 Introduction  3  
1.2 Approaches based on causal models  3  
1.3 Canonical inference  9  
1.4 Methodologic modeling  10  
1.5 Conclusion  13  

2 Matching in observational studies, by Paul R. Rosenbaum  15  
2.1 The role of matching in observational studies  15  
2.2 Why match?  16  
2.3 Two key issues: balance and structure  17  
2.4 Additional issues  21  

3 Estimating causal effects in nonexperimental studies, by Rajeev Dehejia  25  
3.1 Introduction  25  
3.2 Identifying and estimating the average treatment effect  27  
3.3 The NSW data  29  
3.4 Propensity score estimates  31  
3.5 Conclusions  35  

4 Medication cost sharing and drug spending in Medicare, by Alyce S. Adams  37  
4.1 Methods  38  
4.2 Results  40  
4.3 Study limitations  45  
4.4 Conclusions and policy implications  46  

5 A comparison of experimental and observational data analyses, by Jennifer L. Hill, Jerome P. Reiter, and Elaine L. Zanutto  49  
5.1 Experimental sample  50
5.2 Constructed observational study ...................................... 51
5.3 Concluding remarks ...................................................... 60

6 Fixing broken experiments using the propensity score,
by Bruce Sacerdote ......................................................... 61
   6.1 Introduction .......................................................... 61
   6.2 The lottery data ....................................................... 62
   6.3 Estimating the propensity scores ................................... 63
   6.4 Results .............................................................. 65
   6.5 Concluding remarks .................................................. 71

7 The propensity score with continuous treatments,
by Keisuke Hirano and Guido W. Imbens ............................. 73
   7.1 Introduction .......................................................... 73
   7.2 The basic framework ................................................ 74
   7.3 Bias removal using the GPS ......................................... 76
   7.4 Estimation and inference ............................................ 78
   7.5 Application: the Imbens–Rubin–Sacerdote lottery sample ....... 79
   7.6 Conclusion .......................................................... 83

8 Causal inference with instrumental variables, by Junni L. Zhang 85
   8.1 Introduction .......................................................... 85
   8.2 Key assumptions for the LATE interpretation of the IV estimand . 87
   8.3 Estimating causal effects with IV ................................... 90
   8.4 Some recent applications ............................................ 95
   8.5 Discussion .......................................................... 95

9 Principal stratification, by Constantine E. Frangakis ............. 97
   9.1 Introduction: partially controlled studies ......................... 97
   9.2 Examples of partially controlled studies .......................... 97
   9.3 Principal stratification .............................................. 101
   9.4 Estimands .......................................................... 102
   9.5 Assumptions ........................................................ 104
   9.6 Designs and polydesigns ............................................. 107

II Missing data modeling ....................................................... 109

10 Nonresponse adjustment in government statistical agencies:
   constraints, inferential goals, and robustness issues,
by John L. Eltinge .......................................................... 111
   10.1 Introduction: a wide spectrum of nonresponse adjustment efforts in
government statistical agencies ....................................... 111
   10.2 Constraints ........................................................ 112
   10.3 Complex estimand structures, inferential goals, and utility functions 112
10.4 Robustness ................................................................. 113
10.5 Closing remarks ......................................................... 113

11 Bridging across changes in classification systems, by Nathaniel Shenker 
11.1 Introduction ............................................................... 117
11.2 Multiple imputation to achieve comparability of industry and occupation codes ................................................................. 118
11.3 Bridging the transition from single-race reporting to multiple-race reporting ........................................................................ 123
11.4 Conclusion .................................................................. 128

12 Representing the Census undercount by multiple imputation of households, by Alan M. Zaslavsky 129
12.1 Introduction ............................................................... 129
12.2 Models .................................................................. 131
12.3 Inference ................................................................ 134
12.4 Simulation evaluations ................................................ 138
12.5 Conclusion ................................................................ 140

13 Statistical disclosure techniques based on multiple imputation, 
by Roderick J. A. Little, Fang Liu, and Trivellore E. Raghunathan 141
13.1 Introduction ............................................................... 141
13.2 Full synthesis ............................................................ 143
13.3 SMIKe and MIKe ...................................................... 144
13.4 Analysis of synthetic samples ........................................ 147
13.5 An application ........................................................... 149
13.6 Conclusions ............................................................... 152

14 Designs producing balanced missing data: examples from the National Assessment of Educational Progress, by Neal Thomas 153
14.1 Introduction ............................................................... 153
14.2 Statistical methods in NAEP ........................................ 155
14.3 Split and balanced designs for estimating population parameters ................................................................. 157
14.4 Maximum likelihood estimation ................................... 159
14.5 The role of secondary covariates ................................. 160
14.6 Conclusions ............................................................... 162

15 Propensity score estimation with missing data, 
by Ralph B. D’Agostino Jr. ................................................... 163
15.1 Introduction ............................................................... 163
15.2 Notation ................................................................. 165
15.3 Applied example: March of Dimes data ....................... 168
15.4 Conclusion and future directions .................................. 174
16 Sensitivity to nonignorability in frequentist inference,  
by Guoguang Ma and Daniel F. Heitjan 175
16.1 Missing data in clinical trials .................................. 175
16.2 Ignorability and bias ............................................. 175
16.3 A nonignorable selection model ................................. 176
16.4 Sensitivity of the mean and variance ......................... 177
16.5 Sensitivity of the power ........................................ 178
16.6 Sensitivity of the coverage probability ...................... 180
16.7 An example ...................................................... 184
16.8 Discussion ...................................................... 185

III Statistical modeling and computation 187

17 Statistical modeling and computation, by D. Michael Titterington 189
17.1 Regression models .............................................. 190
17.2 Latent-variable problems ...................................... 191
17.3 Computation: non-Bayesian .................................. 191
17.4 Computation: Bayesian ....................................... 192
17.5 Prospects for the future ...................................... 193

18 Treatment effects in before-after data, by Andrew Gelman 195
18.1 Default statistical models of treatment effects .............. 195
18.2 Before-after correlation is typically larger for controls than for treated units ........................................ 196
18.3 A class of models for varying treatment effects ........... 200
18.4 Discussion ...................................................... 201

19 Multimodality in mixture models and factor models, by Eric Loken 203
19.1 Multimodality in mixture models .............................. 204
19.2 Multimodal posterior distributions in continuous latent variable models .............................................. 209
19.3 Summary ........................................................ 212

20 Modeling the covariance and correlation matrix of repeated measures, 
by W. John Boscardin and Xiao Zhang 215
20.1 Introduction ....................................................... 215
20.2 Modeling the covariance matrix ............................... 216
20.3 Modeling the correlation matrix ............................... 218
20.4 Modeling a mixed covariance-correlation matrix ........... 220
20.5 Nonzero means and unbalanced data ......................... 220
20.6 Multivariate probit model ..................................... 221
20.7 Example: covariance modeling ................................ 222
20.8 Example: mixed data ........................................... 225
21 Robit regression: a simple robust alternative to logistic and probit regression, by Chuanhai Liu  227  
21.1 Introduction .................................................. 227  
21.2 The robit model ................................................. 228  
21.3 Robustness of likelihood-based inference using logistic, probit, and robit regression models ................................................. 230  
21.4 Complete data for simple maximum likelihood estimation ........ 231  
21.5 Maximum likelihood estimation using EM-type algorithms ........ 233  
21.6 A numerical example ............................................. 235  
21.7 Conclusion .................................................. 238  

22 Using EM and data augmentation for the competing risks model, by Radu V. Craiu and Thierry Duchesne  239  
22.1 Introduction .................................................. 239  
22.2 The model .................................................. 240  
22.3 EM-based analysis ............................................. 243  
22.4 Bayesian analysis ............................................. 244  
22.5 Example .................................................. 248  
22.6 Discussion and further work .................................. 250  

23 Mixed effects models and the EM algorithm, by Florin Vaida, Xiao-Li Meng, and Ronghui Xu  253  
23.1 Introduction .................................................. 253  
23.2 Binary regression with random effects .......................... 254  
23.3 Proportional hazards mixed-effects models ...................... 259  

24 The sampling/importance resampling algorithm, by Kim-Hung Li  265  
24.1 Introduction .................................................. 265  
24.2 SIR algorithm .................................................. 266  
24.3 Selection of the pool size ..................................... 267  
24.4 Selection criterion of the importance sampling distribution .... 271  
24.5 The resampling algorithms .................................... 272  
24.6 Discussion .................................................. 276  

IV Applied Bayesian inference  277  

25 Whither applied Bayesian inference?, by Bradley P. Carlin  279  
25.1 Where we’ve been ............................................. 279  
25.2 Where we are .................................................. 281  
25.3 Where we’re going ............................................. 282  

26 Efficient EM-type algorithms for fitting spectral lines in high-energy astrophysics, by David A. van Dyk and Taeyoung Park  285  
26.1 Application-specific statistical methods .......................... 285
26.2 The Chandra X-ray observatory ........................................ 287
26.3 Fitting narrow emission lines ........................................... 289
26.4 Model checking and model selection ................................. 294

27 Improved predictions of lynx trappings using a biological model, 
by Cavan Reilly and Angelique Zeringue .................................. 297
27.1 Introduction ............................................................... 297
27.2 The current best model .................................................. 298
27.3 Biological models for predator prey systems ....................... 299
27.4 Some statistical models based on the Lotka-Volterra system ..... 300
27.5 Computational aspects of posterior inference ..................... 302
27.6 Posterior predictive checks and model expansion ............... 304
27.7 Prediction with the posterior mode .................................... 307
27.8 Discussion ...................................................................... 308

28 Record linkage using finite mixture models, by Michael D. Larsen 309
28.1 Introduction to record linkage ........................................... 309
28.2 Record linkage ............................................................... 310
28.3 Mixture models ............................................................. 311
28.4 Application ...................................................................... 314
28.5 Analysis of linked files ..................................................... 316
28.6 Bayesian hierarchical record linkage ................................. 317
28.7 Summary ........................................................................ 318

29 Identifying likely duplicates by record linkage in a survey 
of prostitutes, by Thomas R. Belin, Hemant Ishwaran, Naihua Duan, 
Sandra H. Berry, and David E. Kanouse .................................. 319
29.1 Concern about duplicates in an anonymous survey ............. 319
29.2 General frameworks for record linkage ............................. 321
29.3 Estimating probabilities of duplication in the Los Angeles Women’s 
Health Risk Study .............................................................. 322
29.4 Discussion ...................................................................... 328

30 Applying structural equation models with incomplete data, 
by Hal S. Stern and Yoonsook Jeon ........................................ 331
30.1 Structural equation models ................................................ 332
30.2 Bayesian inference for structural equation models ............. 334
30.3 Iowa Youth and Families Project example .......................... 339
30.4 Summary and discussion .................................................. 342

31 Perceptual scaling, by Ying Nian Wu, Cheng-En Guo, 
and Song Chun Zhu ............................................................... 343
31.1 Introduction .................................................................... 343
31.2 Sparsity and minimax entropy ......................................... 347
31.3 Complexity scaling law .................................................... 353
CONTENTS

31.4 Perceptibility scaling law ........................................ 356
31.5 Texture = imperceptible structures .......................... 358
31.6 Perceptibility and sparsity .................................... 359

References 361

Index 401
Preface

This volume came into existence because of our long-held desire to produce a “showcase” book on the ways in which complex statistical theories and methods are actually applied in the real world. By “showcase,” we do not imply in any way that this volume presents the best possible analyses or applications—any such claim would only demonstrate grotesque lack of understanding of the complexity and artistic nature of statistical analysis. The world’s top five statisticians, however selected, could never produce identical “solutions” to any real-life statistical problem. Putting it differently, if they were all to arrive at the same answer, in the usual mathematical sense, then the problem must be of a toy nature.

Just as objects displayed in a museum showcase are often collectibles from various sources attracting different degrees of appreciation by different viewers, readers of this volume may walk away with different degrees of intellectual stimulation and satisfaction. Nevertheless, we have tried to provide something for almost everyone. To put it another way, it would be difficult to find an individual, statistician or otherwise, who could successfully deal with a real-life statistical problem without having the frustration of dealing with missing data, or the need for some sophistication in modeling and computation, or the urge, possibly subconscious, to learn about underlying causal questions. The substantive areas touched upon by the chapters in this volume are also wide-ranging, including astrophysics, biology, economics, education, medicine, neuroscience, political science, psychology, public policy, sociology, visual learning, and so forth. The Summary of Contents below provides a more detailed account.

Like any showcase display, there is a general theme underlying the chapters in this volume. Almost all the methods discussed in this volume benefited from the incomplete-data perspective. This is certainly true for the counterfactual model for causal inference, for multiple imputation, for the EM algorithm and more generally for data augmentation methods, for mixture modeling, for latent variables, for Bayes hierarchical models, and so forth. Most of the chapters also share a common feature in that out of the total of 31 chapters, 24 are authored or coauthored by Donald Rubin’s students and grandstudents. Their names are indicated in the “family tree” on page xix. Three of the remaining seven chapters are coauthored by Don’s long-time collaborators: Guido Imbens, Rod Little, and Hal Stern. The remaining four chapters are written by specially invited distinguished experts who are not part of the “Rubin statistical family”: Sander Greenland, John Eltinge, Mike Titterington, and Brad Carlin. Each of these “outsiders” provides an overview article to lead
the four parts of the volume. No matter how large any statistical family is, it is obvious that readers will benefit from both within and between-family variability, sometimes dominantly so by the latter.

The immediate motivation for this volume is to celebrate Don Rubin’s 60th birthday, and it is scheduled to appear just in time for the 2004 Joint Statistical Meetings in Toronto, during which Don will deliver the Fisher lecture. As his students, we obviously wish to dedicate this volume to Don, whose enormous contribution to statistics and impact on general quantitative scientific studies are more than evident from the chapters presented in this volume. We checked the Science Citation Index and found that his papers have been cited over 8,000 times—but Don claims that what he really likes is that his ideas such as ignorability and multiple imputation are so accepted that people use them without even citing him.¹ (A quick look through Parts 3 and 4 of this volume, along with the reference list, reveals that Bayes, and Metropolis are similarly honored but not cited by our contributors.)

Indeed, Don’s work is so wide-ranging that it was not an easy task to come up with an accurate but attractive title for this volume. Titles we considered include “Blue-label Statistics (60 years): Sipping with Donald Rubin,” “Defenders of Tobacco Companies: From R. A. Fisher to D. B. Rubin,” and so forth. We finally settled on the current title, not as amusing as some of us would have liked, but conveying the serious objective of this volume: to showcase a range of applications and topics in applied statistics related to inference and missing data and to take the reader to the frontiers of research in these areas.

Summary of Contents

Part 1: Causal inference and observational studies

Part 1 contains nine chapters, leading with Sander Greenland’s overview of three common approaches to causal inference from observational studies. Greenland’s chapter is followed with a chapter on the role of matching in observational studies, and a chapter reviewing the basics of the most popular method of performing matching, based on propensity scores, with illustrations using data from the National Supported Work Demonstration and the Current Population Survey. Propensity score matching is in some ways as fundamental to observational studies as randomization is to experimental studies, for it provides “the next to the best thing”—a remarkably simple and effective method for reducing or even eliminating confounding factors when randomization is not possible or not used in the design stage.

The next three chapters apply the propensity score method to three studies in public health and economics: a Medicare cost-sharing and drug-spending study, an infant health development study, and a Massachusetts lottery study. Along the way,

¹For the record, it’s Rubin (1976) and Rubin (1978).
these chapters also demonstrate how to use propensity score matching to construct observational studies and to fix “broken experiments.” The seventh chapter of Part 1 shows how propensity scores can be extended to continuous treatments, and the methods are applied to the aforementioned lottery study.

The eighth chapter provides an introduction to another popular method in causal inference, the method of instrumental variables. The last chapter of Part 1 investigates the use of instrumental variables for dealing with “partially controlled” studies, a rather difficult class of problems where extra caution is needed in order to arrive at meaningful estimates for treatment effects. The fundamental concept of “principal stratification” is introduced and illustrated with a study on the effectiveness of a needle exchange program in reducing HIV transmission.

Part 2: Missing data modeling

The second part of the book begins with a review by John Eltinge of methods used to adjust for nonresponse in government surveys. The next three chapters provide three accounts of applications of multiple imputation, one of the most popular methods for dealing with missing data, especially in the context of producing public-use data files. The first of the three applications concerns the use of multiple imputation for the purposes of bridging across changes in classification systems, from the earliest application of multiple imputation for achieving comparability between 1970 and 1980 industry and occupation codes, to one of the latest applications involving bridging the transition from single-race reporting to multiple-race reporting in the 2000 Census.

The second of the three applications concerns the use of multiple imputation for representing Census undercount, an extremely contentious issue due to the use of census data in allocation of congressional seats and federal funding. The third application touches on data confidentiality—another controversial issue that has received much attention among the public and in government. Multiple imputation provides a flexible framework for dealing with the conflict between confidentiality and the informativeness of released data by replacing parts or all of the data with synthetic imputations.

The remaining three chapters of Part 2 address design and estimation issues in the presence of missing data. The first of the three investigates the “missing by design” issue with the National Assessment of Educational Progress, an ongoing collection of surveys of students (and teachers) in the U.S. that uses a matrix sampling design to reduce the burden on each student—namely, different students are administered different small subsets of a large collection of test items. The next chapter presents models and computation methods for dealing with the problem of missing data in estimating propensity scores, with application to the March of Dimes observational study examining various effects of post-term birth versus term birth on preteen development. The last chapter of Part 2 describes a convenient method for diagnosing the sensitivity of inferences to nonignorability in missing data modeling, a thorny but essential issue in almost any real-life missing-data problem.
Part 3: Statistical modeling and computation

The third part of the book begins with an overview by Mike Titterington of modeling and computation, which between them cover much of applied statistics nowadays. As Titterington notes, although the more cerebral activity of modeling is rather different from the nuts-and-bolts issues of computation, the two lines of research are in practice closely interwoven. General ideas of modeling are immediately worthwhile only if they are computationally feasible. On the other hand, the need for fitting realistic models in complex settings (which essentially is the case for most applied activities when we take them seriously) has been the strongest stimulus for more advanced and statistical computational methods, the availability of which in turn promote investigators to consider models beyond what are traditionally available (to a point that there is a growing concern that we fit more complex models simply because we can).

The remaining chapters in Part 3 clearly demonstrate this interweaving, both in methodological research and in practice. The second chapter proposes a class of variance-component models for dealing with interactions between treatment and pretreatment covariates, a problem motivated by several examples including observational studies of the effects of redistricting and incumbency in electoral systems in the United States. The next chapter investigates a novel “preclassifying” method for dealing with the tricky computational issue of label-switching with mixtures and other models with latent categories. The investigation involves both the EM algorithm and Markov chain simulation, and the method is illustrated with a well-known factor analysis in educational testing. The following chapter deals with the complicated problem of modeling covariance and correlation matrices, giving specific steps for a Markov chain simulation to fit the proposed models in a variety of settings, and providing a detailed application to a repeated measurement problem arising from a study of long-term neuropsychological impacts after head trauma, using data from UCLA Brain Injury Research Center.

Part 3 continues with a new class of regression models for analyzing binary outcomes, the “robit regression” model, which replaces the normal model underlying the probit regressions by the more robust (hence the term “robit”) t models. The models are then fitted by the EM algorithm and several of its recent extensions. The next two chapters detail how to use both the EM algorithm and Markov chain simulation methods for fitting competing risk models and mixed-effect models, including generalized mixed-effect models and the so-called frailty models for estimating hazard rates in survival analysis. Again, all methods are illustrated in detail using simulated and real data sets. The concluding chapter of Part 3 provides a comprehensive overview and investigation of the sampling/importance resampling algorithm for Bayesian and general computation.

Part 4: Applied Bayesian inference

The final part of the book begins with an entertaining survey by Brad Carlin on the past, present, and future of applied Bayesian inference, followed by six
chapters on how Bayesian methods are applied to address substantive questions in natural and social sciences. The first study is an inference on emission lines in high-energy astrophysics, based on photon counts collected by the Chandra X-ray observatory. Carefully constructed and problem-specific hierarchical models are developed to handle the complex nature of the sources and the collection process of the data. Complications include, but are not limited to, the mixing of continuum and line emission sources, background contamination, restrictions due to “effective area” and instruments, and absorption by interstellar or intergalactic media. The next chapter demonstrates how and why statistical modeling should be integrated with scientific modeling in addressing substantive questions.

In studying a famous example from time series analysis—the dynamic of the Canadian lynx population—a simple biological “prey-predator” type of model combined with empirical time-series models provides a more realistic depiction of the lynx population (with forecasts substantially outperforming previously proposed models), even without the availability of actual data on its prey, the snowshoe hare population!

The next two chapters apply Bayesian methods for record linkage—the problem of matching subjects from different data files or even within the same data file. The methods were originally developed for the purposes of linking various governmental files, such as for estimating undercount by identifying individuals who were counted in both the decennial Census and a Post-Enumeration Survey and those who were only counted in one of the canvasses. However, as Chapter 29 demonstrates, the methods are also useful in identifying duplicates in anonymous surveys. A case in point is the Los Angeles Women’s Health Risk Study, where it was found that about 10% of surveyed prostitutes appeared more than once in the data source because of the monetary incentive given for drawing blood, which was necessary in order to estimate the prevalence of HIV and other diseases in this population.

The next chapter discusses Bayesian inference for structural equation models with incomplete data, as applied to a longitudinal study of rural families using data from the Iowa Youth and Families Project. The final chapter of the book provides a fascinating framework, inspired by the Bayesian philosophy, for studying a profound problem in visual learning, namely, how to model humans’ perceptual transition over scale. For an image of, say, trees, at near distance, we perceive individual leaves, including their edges and shapes. For the same image at far distance, however, we only perceive a collective foliage impression, even though the natural scene itself is the same. The proposed entropy-based framework provides an elegant theoretical explanation of this common-sense perception change. More importantly, it leads to statistical methods for creating synthesized images by effectively separating sparse structures from collective textures in natural images. The pictures on the cover of this volume, supplied by Zijian Xu and Yingnian Wu, show an illustrative comparison of sketch images at two different levels of resolution.
Acknowledgments

Our foremost thanks, of course, go to all the contributors. We are particularly grateful to the four “outsiders” for their willingness to write capsule overviews of huge chunks of statistics, especially given the stringent time constraints. Dale Rinkel’s editorial assistance cannot be over-thanked, for without it, we simply could not have been able to make the deadline for publication. We also thank our editors, Elizabeth Johnston, Rob Calver, and Kathryn Sharples, and the others at Wiley who have helped make this book happen, along with the National Science Foundation for partial support.

And of course we thank Don for turning both of us from students of textbooks into editors of this volume.

— Andrew Gelman and Xiao-Li Meng, April, 2004
THE RUBIN STATISTICAL FAMILY TREE

William G. Cochran, Advisor†

Donald B. Rubin

* Paul Rosenbaum ('80)       * Michael Larsen ('96)
  David Lax ('81)              Heng Li ('96)
* Nathaniel Schenker ('85)    * Ying Nian Wu ('97)
* Kim-Hung Li ('85)           * Rajeev Dehejia ('97)
* Daniel Heitjan ('85)        * Guoguang Ma
  ** Guoguang Ma               * Bruce Sacerdote ('97)
* T.E. Raghunathan ('87)      * Jiangtao Du ('98)
  Leisa Weld ('87)             Leonard Roseman ('98)
* Neal Thomas ('88)           * Keisuke Hirano ('98)
* Alan Zaslavsky ('89)        * Elaine Zanutto ('98)
* Andrew Gelman ('90)         * Constantine Frangakis ('99)
  ** John Boscardin            * Jerome Reiter ('99)
  ** Cavan Reilly              * Jaap P.L. Brand ('99)
* Xiao-Li Meng ('90)          * Alyce Adams ('99)
  ** Radu Craiu                * Jennifer Hill ('00)
  ** Florin Vaida              * C.J. Zijin Shen ('00)
  ** David van Dyk             * Eric Loken ('01)
* Thomas Belin ('91)          * Junni Zhang ('02)
* Joseph Schafer ('92)        * Samantha Cook ('04)
* Ralph D’Agostino, Jr. ('94) * Elizabeth Stuart ('04)
* Chuanhai Liu ('94)

* Rubin’s advisees (with year of Ph.D. graduation) contributing chapters to this book.
** Rubin Statistical Family “grandchildren” contributing chapters. (Other “grandchildren” are not shown.)
† William G. Cochran’s statistical mentors were Ronald A. Fisher, John Wishart, and Frank Yates.
Part I

Casual inference and observational studies
An overview of methods for causal inference from observational studies

Sander Greenland¹

1.1 Introduction

This chapter provides a brief overview of causal-inference methods found in the health sciences. It is convenient to divide these methods into a few broad classes: Those based on formal models of causation, especially potential outcomes; those based on canonical considerations, in which causality is a property of an association to be diagnosed by symptoms and signs; and those based on methodologic modeling. These are by no means mutually exclusive approaches; for example, one may (though need not) base a methodologic model on potential outcomes, and a canonical approach may use modeling methods to address specific considerations. Rather, the categories reflect historical traditions that until recently had only limited intersection.

1.2 Approaches based on causal models

Background: potential outcomes

Most statistical methods, from orthodox Neyman–Pearsonian testing to radical subjective Bayesianism, have been labeled by their proponents as solutions to problems

¹Departments of Epidemiology and Statistics, University of California, Los Angeles. The author is grateful to Katherine Hoggatt, Andrew Gelman, James Robins, Marshall Joffe and Donald Rubin for helpful comments.
of inductive inference (Greenland, 1998), and causal inference may be classified as a prominent (if not the major) problem of induction. It would then seem that causal-inference methods ought to figure prominently in statistical theory and training. That this has not been so has been remarked on by other reviewers (Pearl, 2000). In fact, despite the long history of statistics up to that point, it was not until the 1920s that a formal statistical model for causal inference was proposed (Neyman, 1923), the first example of a potential-outcome model.

Skeptical that induction in general and causal inference in particular could be given a sound logical basis, David Hume nonetheless captured the foundation of potential-outcome models when he wrote:

“We may define a cause to be an object, followed by another, . . . where, if the first object had not been, the second had never existed.” (Hume, 1748, p. 115)

A key aspect of this view of causation is its counterfactual element: It refers to how a certain outcome event (the “second object,” or effect) would not have occurred if, contrary to fact, an earlier event (the “first object,” or cause) had not occurred. In this regard, it is no different from standard frequentist statistics (which refer to sample realizations that might have occurred, but did not) and some forms of competing-risk models (those involving a latent outcome that would have occurred, but for the competing risk). This counterfactual view of causation was adopted by numerous philosophers and scientists after Hume (e.g., Mill, 1843; Fisher, 1918; Cox, 1958; Simon and Rescher, 1966; MacMahon and Pugh, 1967; Lewis, 1973).

The development of this view into a statistical theory for causal inference is recounted by Rubin (1990), Greenland, Robin, and Pearl (1999), Greenland (2000), and Pearl (2000). To describe that theory, suppose we wish to study the effect of an intervention variable \( X \) with potential values (range) \( x_1, \ldots, x_J \) on a subsequent outcome variable \( Y \) defined on an observational unit or a population. The theory then supposes that there is a vector of potential outcomes \( y = (y(x_1), \ldots, y(x_J)) \), such that if \( X = x_j \) then \( Y = y(x_j) \); this vector is simply a mapping from the \( X \) range to the \( Y \) range for the unit. To say that intervention \( x_i \) causally affects \( Y \) relative to intervention \( x_j \) then means that \( y(x_i) \neq y(x_j) \); and the effect of intervention \( x_i \) relative to \( x_j \) on \( Y \) is measured by \( y(x_i) - y(x_j) \) or (if \( Y \) is strictly positive) by \( y(x_i)/y(x_j) \). Under this theory, assignment of a unit to a treatment level \( x_i \) is simply a choice of which coordinate of \( y \) to attempt to observe; regardless of assignment, the remaining coordinates are treated as existing pretreatment covariates on which data are missing (Rubin, 1978a). Formally, if we define the vector of potential treatments \( x = (x_1, \ldots, x_J) \), with treatment indicators \( r_i = 1 \) if the unit is given treatment \( x_i \), 0 otherwise, and \( r = (r_1, \ldots, r_J) \), then the actual treatment given is \( x_a = r'x \) and the actual outcome is \( y_a = y(x_a) = r'y \). Viewing \( r \) as the item-response vector for the items in \( y \), causal inference under potential outcomes can be seen as a special case of inference under item nonresponse in which \( \Sigma_i r_i = 0 \) or 1, that is, at most one item in \( y \) is observed per unit (Rubin, 1991).
The theory extends to stochastic outcomes by replacing the \( y(x_i) \) by probability mass functions \( p_i(y) \) (Greenland, 1987; Robins, 1988; Greenland, Robin, and Pearl, 1999), so the mapping is from \( X \) to the space of probability measures on \( Y \). This extension is embodied in the “set” or “do” calculus for causal actions (Pearl, 1995, 2000) described briefly below. The theory also extends to continuous \( X \) by allowing the potential-outcome vector to be infinite-dimensional with coordinates indexed by \( X \), and components \( y(x) \) or \( p_x(y) \). Finally, the theory extends to complex longitudinal data structures by allowing the treatments to be different event histories or processes (Robins, 1987, 1997).

**Limitations of potential-outcome models**

The power and controversy of this formalization derives in part from defining cause and effect in simple terms of interventions and potential outcomes, rather than leaving them informal or obscure. Judged on the basis of the number and breadth of applications, the potential-outcome approach is an unqualified success, as contributions to the present volume attest. Nonetheless, because only one of the treatments \( x_i \) can be administered to a unit, for each unit at most one potential outcome \( y(x_i) \) will become an observable quantity; the rest will remain counterfactual, and hence in some views less than scientific (Dawid, 2000). More specifically, the approach has been criticized for including structural elements that are in principle unidentifiable by randomized experiments alone. An example is the correlation among potential outcomes: Because no two potential outcomes \( y(x_i) \) and \( y(x_j) \) from distinct interventions \( x_i \neq x_j \) can be observed on one unit, nothing about the correlation of \( y(x_i) \) and \( y(x_j) \) across units can be inferred from observing interventions and outcomes alone; the correlation becomes unobservable and hence by some usage “metaphysical.”

This sort of problem has been presented as if it is a fatal flaw of potential outcomes models (Dawid, 2000). Most commentators, however, regard such problems as indicating inherent limits of inference on the basis of unrepeatable “black-box” observation: For some questions, one must go beyond observations of unit responses, to unit-specific investigation of the mechanisms of action (e.g., dissection and physiology). This need is familiar in industrial statistics in the context of destructive testing, although controversy does not arise there because the mechanisms of action are usually well understood. The potential-outcomes approach simply highlights the limits of what statistical analyses can show without background theory about causal mechanisms, even if treatment is randomized: standard statistical analyses address only the magnitude of associations and the average causal effects they represent, not the mechanisms underlying those effects.

**Translating potential outcomes into statistical methodology**

Among the earliest applications of potential outcomes were the randomization tests for causal effects. These applications illustrate the transparency potential outcomes
can bring to standard methods, and show their utility in revealing the assumptions
needed to give causal interpretations to standard statistical procedures.

Suppose we have \( N \) units indexed by \( n \) and we wish to test the strong (sharp)
null hypothesis that treatment \( X \) has no effect on \( Y \) for any unit, that is, for all
\( i, j, n, y_n(x_i) = y_n(x_j) \). Under this null, the observed distribution of \( Y \) among the
\( N \) units would not differ from its observed value, regardless of how treatment is
allocated among the units. Consequently, given the treatment-allocation probabil-
ities (propensity scores), we may compute the exact null distribution of any measure
of differences among treatment groups. In doing so, we can and should keep the
marginal distribution of \( Y \) at its observed value, for with no treatment effect on \( Y \),
changes in treatment allocation cannot alter the marginal distribution of \( Y \).

The classic examples of this reasoning are permutation tests based on uniform
allocation probabilities across units (simple randomization), such as Fisher’s exact
test (Cox and Hinkley, 1974, sec. 6.4). For these tests, the fixed \( Y \)-margin is often
viewed as a mysterious assumption by students, but can be easily deduced from the
potential-outcome formulation, with no need to appeal to obscure and controversi-
all conditionality principles (Greenland, 1991). Potential-outcome models can also be
used to derive classical confidence intervals (which involve nonnull hypotheses and
varying margins), superpopulation inferences (in which the \( N \) units are viewed as a
random sample from the actual population of interest), and posterior distributions
for causal effects of a randomized treatment (Robins, 1988; Rubin, 1978). The
models further reveal hidden assumptions and limitations of common procedures
for instrumental-variable estimation (Angrist, Imbens, and Rubin, 1996), for intent-
to-treat analyses (Goetghebeur and van Houwelingen, 1998), for separating direct
and indirect effects (Robins and Greenland, 1992, 1994; Frangakis and Rubin,
2002), for confounding identification (Greenland, Robins, and Pearl, 1999), for
estimating causation probabilities (Greenland and Robins, 2000), for handling time-
varying covariates (Robins, 1987, 1998; Robins et al., 1992), and for handling
time-varying outcomes (Robins, Greenland, and Hu, 1999a).

A case study: g-estimation

Potential-outcome models have contributed much more than conceptual clarifica-
tion. As documented elsewhere in this volume, they have been used extensively
by Rubin, his students, and his collaborators to develop novel statistical proce-
dures for estimating causal effects. Indeed, one defense of the approach is that it
stimulates insights which lead not only to the recognition of shortcomings of pre-
vious methods but also to development of new and more generally valid methods
(Wasserman, 2000).

Methods for modeling effects of time-varying treatment regimes (generalized
treatments, or “g-treatments”) provide a case study in which the potential-outcome
approach led to a very novel way of attacking an exceptionally difficult problem.
The difficulty arises because a time-varying regime may not only be influenced
by antecedent causes of the outcome (which leads to familiar issues of confounding) but may also influence later causes, which in turn may influence the regime. Robins (1987) identified a recursive “g-computation” formula as central to modeling treatment effects under these feedback conditions and derived nonparametric tests on the basis of this formula (a special case of which was first described by Morrison, 1985). These tests proved impractical beyond simple null-testing contexts, which led to the development of semiparametric modeling procedures for inferences about time-varying treatment effects (Robins, 1998).

The earliest of these procedures were based on the structural-nested failure-time model (SNFTM) for survival time $Y$ (Robins, Blevins et al., 1992; Robins and Greenland, 1994; Robins, 1998), a generalization of the strong accelerated-life model (Cox and Oakes, 1984). Suppressing the unit subscript $n$, suppose a unit is actually given fixed treatment $X = x_a$ and fails at time $Y_a = y(x_a)$, the potential outcome of the unit under $X = x_a$. The basic accelerated-life model assumes the survival time of the unit when given $X = 0$ instead would have been $Y_0 = e^{x_a \beta} Y_a$, where $Y_0$ is the potential outcome of the unit under $X = 0$, and the factor $e^{x_a \beta}$ is the amount by which setting $X = 0$ would have expanded (if $x_a \beta > 0$) or contracted (if $x_a \beta < 0$) survival time relative to setting $X = x_a$.

Suppose now $X$ could vary and the actual survival interval $S = (0, Y_a)$ is partitioned into $K$ successive intervals of length $\Delta t_1, \ldots, \Delta t_K$, such that $X = x_k$ in interval $k$, with a vector of covariates $Z = z_k$ in the interval. A basic SNFTM for the survival time of the unit had $X$ been held at zero over time is then $Y_0 = \Sigma_k \exp(x_k \beta) \Delta t_k$; the extension to a continuous treatment history $x(t)$ is $Y_0 = \int_S e^{x(t) \beta} dt$. The model is semiparametric insofar as the distribution of $Y_0$ across units is unspecified or incompletely specified, although this distribution may be modeled as a function of covariates, for example, by a proportional-hazards model for $Y_0$.

Likelihood-based inference on $\beta$ is unwieldy, but testing and estimation can be easily done with a clever two-step procedure called g-estimation (Robins et al., 1992; Robins and Greenland, 1994; Robins, 1998). To illustrate the basic idea, assume no censoring of $Y$, no measurement error, and let $X_k$ and $Z_k$ be the treatment and covariate random variables for interval $k$. Then, under the model, a hypothesized value $\beta_h$ for $\beta$ produces for each unit a computable value $Y_0(\beta_h) = \Sigma_k \exp(x_k \beta_h) \Delta t_k$ for $Y_0$. Next, suppose that for all $k$, $Y_0$ and $X_k$ are independent given past treatment history $X_1, \ldots, X_{k-1}$ and covariate history $Z_1, \ldots, Z_k$ (as would obtain if treatment were sequentially randomized given these histories). If $\beta = \beta_h$, then $Y_0(\beta_h) = Y_0$ and so must be independent of $X_k$ given the histories. One may test this conditional independence of $Y_0(\beta_h)$ and the $X_k$ with any standard method. For example, one could use a permutation test or some approximation to one (such as the usual logrank test) stratified on histories; subject to further modeling assumptions, one could instead use a test that the coefficient of $Y_0(\beta_h)$ is zero in a model for the regression of $X_k$ on $Y_0(\beta_h)$ and the histories. In either case, $\alpha$-level rejection of conditional independence of $X_k$ and $Y_0(\beta_h)$ implies $\alpha$-level rejection of $\beta = \beta_h$, and the set of all $\beta_h$ not so rejected form a $1 - \alpha$
confidence set for $\beta$. Furthermore, the random variable corresponding to the value $b$ for $\beta$ that makes $Y_0(b)$ and the $X_k$ conditionally independent is a consistent, asymptotically normal estimator of $\beta$ (Robins, 1998).

Of course, in observational studies, g-estimation shares all the usual limitations of standard methods. The assignment mechanism is not known, so inferences are only conditional on an uncertain assumption of “no sequential confounding”; more precisely, that $Y_0$ and the $X_k$ are independent given the treatment and covariate histories used for stratification or modeling of $Y_0$ and the $X_k$. If this independence is not assumed, then rejection of $\beta_h$ only entails that either $\beta \neq \beta_h$ or that $Y_0$ and the $X_k$ are dependent given the histories (i.e., there is residual confounding). Also, inferences are conditional on the form of the model being correct, which is not likely to be exactly true, even if fit appears good. Nonetheless, as in many standard testing contexts (such as the classical t-test), under broad conditions the asymptotic size of the stratified test of the no-effect hypothesis $\beta = 0$ will not exceed $\alpha$ if $Y_0$ and the $X_k$ are indeed independent given the histories (i.e., absent residual confounding), even if the chosen SNFTM for $Y_0$ is incorrect, although the power of the test may be severely impaired by the model misspecification (Robins, 1998). In light of this “null-robustness” property, g-null testing can be viewed as a natural extension of classical null testing to time-varying treatment comparisons.

If (as usual) censoring is present, g-estimation becomes more complex (Robins, 1998). As a simpler though more restrictive approach to censored longitudinal data with time-varying treatments, one may fit a marginal structural model (MSM) for the potential outcomes using a generalization of Horvitz–Thompson inverse-probability-of-selection weighting (Robins, 1999; Hernan, Brumback, and Robins, 2001). Unlike standard time-dependent Cox models, both SNFTM and MSM fitting require special attention to the censoring process, but make weaker assumptions about that process. Thus their greater complexity is the price one must pay for the generality of the procedures, for both can yield unconfounded effect estimates in situations in which standard models appear to fit well but yield very biased results (Robins et al., 1992; Robins and Greenland, 1994; Robins, Greenland, and Hu, 1999a; Hernan, Brumback, and Robins, 2001).

Other formal models of causation

Most statistical approaches to causal modeling incorporate elements formally equivalent to potential outcomes (Pearl, 2000). For example, the sufficient-component cause model found in epidemiology (Rothman and Greenland, 1998, Chapter 2) is a potential-outcome model. In structural-equation models (SEMs), the component equations can be interpreted as models for potential outcomes (Pearl, 1995, 2000), as in the SNFTM example. The identification calculus based on graphical models of causation (causal diagrams) has a direct mapping into the potential-outcomes framework, and yields the g-computation algorithm as a by-product (Pearl, 1995).