Title
Statistical Models and Causal Inference: A Dialogue with the Social Sciences

Author
Freedman, David A

Publication Date
2009-11-01

Peer reviewed

This file includes the Title Page, an Overview Page, Table of Contents, Preface, and Editors' Introduction.

Statistical Models and Causal Inference

A Dialogue with the Social Sciences

David A. Freedman

Edited by

David Collier
University of California, Berkeley

Jasjeet S. Sekhon
University of California, Berkeley

Philip B. Stark
University of California, Berkeley

CAMBRIDGE
UNIVERSITY PRESS

2009

STATISTICAL MODELS AND CAUSAL INFERENCE

A Dialogue with the Social Sciences

This volume collects twenty of David A. Freedman's most accessible and influential papers on the use and limits of statistical modeling in social science, policy, law, and epidemiology. Through this collection, Freedman offers an integrated synthesis of his views on causal inference. He explores the foundations and limitations of statistical modeling and evaluates research in political science, public policy, law, and epidemiology.

Freedman argues that many new technical approaches to statistical modeling constitute not progress, but regress, and he shows why these methods are not reliable. Instead, Freedman advocates "shoe leather" methodology, which exploits natural variation to mitigate confounding and relies on intimate knowledge of the subject matter to develop meticulous research designs and eliminate rival explanations.

When Freedman first enunciated this position, he was met with skepticism, in part because it was hard to believe that a mathematical statistician of his stature would favor "low-tech" approaches. But the tide is turning. Many social scientists now agree that statistical technique cannot substitute for good research design and subject matter knowledge. Freedman offers here a definitive synthesis of his approach.

David A. Freedman (1938–2008) was Professor of Statistics at the University of California, Berkeley. He was a distinguished mathematical statistician whose theoretical research included the analysis of martingale inequalities, Markov processes, de Finetti's theorem, consistency of Bayes estimators, sampling, the bootstrap, and procedures for testing and evaluating models of methods for causal inference. Freedman published widely on the application—and misapplication—of statistics in works within a variety of social sciences, including epidemiology, demography, public policy, and law. He emphasized exposing and checking the assumptions that underlie standard methods, as well as understanding how those methods behave when the assumptions are false—for example, how regression models behave when fitted to data from randomized experiments. He had a remarkable talent for integrating carefully honed statistical arguments with compelling empirical applications and illustrations. Freedman was a member of the American Academy of Arts and Sciences, and in 2003 he received the National Academy of Science's John J. Carty Award for his "profound contributions to the theory and practice of statistics."

David Collier is Robson Professor of Political Science at the University of California, Berkeley. His current research focuses on conceptualization and measurement and on causal inference in qualitative and multi-method research. He is co-author of *Rethinking Social Inquiry: Diverse Tools, Shared Standards* and co-editor of *The Oxford Handbook of Political Methodology* and *Concepts and Method in Social Science*. His articles on concepts and on causal inference have appeared in the *American Political Science Review*, *World Politics*, *Annual Review of Political Science*, *Political Science: State of the Discipline*, *Journal of Political Ideologies*, and *Political Analysis*.

Jasjeet S. Sekhon is Associate Professor of Political Science at the University of California, Berkeley. His research interests include elections, applied and computational statistics, causal inference in observational and experimental studies, voting behavior, public opinion, and the philosophy and history of science. Professor Sekhon received his Ph.D. in 1999 from Cornell University and was a professor at Harvard University in the Department of Government from 1999 to 2005.

Philip B. Stark is Professor of Statistics at the University of California, Berkeley. His research centers on inference (inverse) problems, primarily in physical science. He is especially interested in confidence procedures tailored for specific goals and in quantifying the uncertainty in inferences that rely on simulations of complex physical systems. Professor Stark has done research on the Big Bang, causal inference, the U.S. Census, earthquake prediction, election auditing, the geomagnetic field, geriatric hearing loss, information retrieval, Internet content filters, nonparametrics (confidence procedures for function and probability density estimates with constraints), the seismic structure of the Sun and Earth, spectroscopy, and spectrum estimation.

Contents

Preface xi

Editors' Introduction: Inference and Shoe Leather xiii

Part I
Statistical Modeling: Foundations and Limitations

1. Issues in the Foundations of Statistics:
Probability and Statistical Models 3

Bayesians and frequentists disagree on the meaning of probability and other foundational issues, but both schools face the problem of model validation. Statistical models have been used successfully in the physical and life sciences. However, they have not advanced the study of social phenomena. How do models connect with reality? When are they likely to deepen understanding? When are they likely to be sterile or misleading?

2. Statistical Assumptions as Empirical Commitments 23

Statistical inference with convenience samples is risky. Real progress depends on a deep understanding of how the data were generated. No amount of statistical maneuvering will get very far without recognizing that statistical issues and substantive issues overlap.

3. Statistical Models and Shoe Leather 45

Regression models are used to make causal arguments in a wide variety of applications, and it is time to evaluate the results. Snow's work on cholera is a success story for causal inference based on nonexperimental data, which was collected through great expenditure of effort and shoe leather. Failures are also discussed. Statistical technique is seldom an adequate substitute for substantive knowledge of the topic, good research design, relevant data, and empirical tests in diverse settings.

Part II
Studies in Political Science, Public Policy, and Epidemiology

4. Methods for Census 2000 and Statistical Adjustments 65

The U.S. Census is a sophisticated, complex undertaking, carried out on a vast scale. It is remarkably accurate. Statistical adjustments are likely to introduce more error than they remove. This issue was litigated all the way to the Supreme Court, which in 1999 unanimously supported the Secretary of Commerce's decision not to adjust the 2000 Census.

5. On "Solutions" to the Ecological Inference Problem 83

Gary King's book, *A Solution to the Ecological Inference Problem*, claims to offer "realistic estimates of the uncertainty of ecological estimates." Applying King's method and three of his main diagnostics to data sets where the truth is known shows that his diagnostics cannot distinguish between cases where estimates are accurate and those where estimates are far off the mark. King's claim to have arrived at a solution to this problem is premature.

6. Rejoinder to King 97

King's method works with some data sets but not others. As a theoretical matter, inferring the behavior of subgroups from aggregate data is generally impossible: The relevant parameters are not identifiable. King's diagnostics do not discriminate between probable successes and probable failures.

7. Black Ravens, White Shoes, and Case Selection: Inference with Categorical Variables 105

Statistical ideas can clarify issues in qualitative analysis such as case selection. In political science, an important argument about case selection evokes Hempel's Paradox of the Ravens. This paradox can be resolved by distinguishing between population and sample inferences.

8. What is the Chance of an Earthquake? 115

Making sense of earthquake forecasts is surprisingly difficult. In part, this is because the forecasts are based on a complicated mixture of geological maps, rules of thumb, expert opinion, physical models, stochastic models, and numerical simulations, as well as geodetic, seismic, and paleoseismic data. Even the concept of probability is hard to define in this

context. Other models of risk for emergency preparedness, as well as models of economic risk, face similar difficulties.

9. Salt and Blood Pressure:
Conventional Wisdom Reconsidered 131

Experimental evidence suggests that the effect of a large reduction in salt intake on blood pressure is modest and that health consequences remain to be determined. Funding agencies and medical journals have taken a stronger position favoring the salt hypothesis than is warranted, demonstrating how misleading scientific findings can influence public policy.

10. The Swine Flu Vaccine and Guillain-Barré Syndrome:
A Case Study in Relative Risk and Specific Causation 151

Epidemiologic methods were developed to prove general causation: identifying exposures that increase the risk of particular diseases. Courts of law often are more interested in specific causation: On balance of probabilities, was the plaintiff's disease caused by exposure to the agent in question? There is a considerable gap between relative risks and proof of specific causation because individual differences affect the interpretation of relative risk for a given person. This makes specific causation especially hard to establish.

11. Survival Analysis: An Epidemiological Hazard? 169

Proportional-hazards models are frequently used to analyze data from randomized controlled trials. This is a mistake. Randomization does not justify the models, which are rarely informative. Simpler methods work better. This discussion matters because survival analysis has introduced a new hazard: It can lead to serious mistakes in medical treatment. Survival analysis is, unfortunately, thriving in other disciplines as well.

Part III
New Developments: Progress or Regress?

12. On Regression Adjustments in Experiments
with Several Treatments 195

Regression adjustments are often made to experimental data to address confounders that may not be balanced by randomization. Since randomization does not justify the models, bias is likely. Neither are the usual variance calculations to be trusted. Neyman's non-parametric model

serves to evaluate regression adjustments. A bias term is isolated, and conditions are given for unbiased estimation in finite samples.

13. Randomization Does Not Justify Logistic Regression 219

The logit model is often used to analyze experimental data. Theory and simulation show that randomization does not justify the model, so the usual estimators can be inconsistent. Neyman's non-parametric setup is used as a benchmark: Each subject has two potential responses, one if treated and the other if untreated; only one of the two responses can be observed. A consistent estimator is proposed.

14. The Grand Leap 243

A number of algorithms purport to discover causal structure from empirical data with no need for specific subject-matter knowledge. Advocates have no real success stories to report. These algorithms solve problems quite removed from the challenge of causal inference from imperfect data. Nor do they resolve long-standing philosophical questions about the meaning of causation.

15. On Specifying Graphical Models for Causation, and the Identification Problem 255

Causal relationships cannot be inferred from data by fitting graphical models without prior substantive knowledge of how the data were generated. Successful applications are rare because few causal pathways can be excluded a priori.

16. Weighting Regressions by Propensity Scores 279

The use of propensity scores to reduce bias in regression analysis is increasingly common in the social sciences. Yet weighting is likely to increase random error in the estimates and to bias the estimated standard errors downward, even when selection mechanisms are well understood. If investigators have a good causal model, it seems better just to fit the model without weights. If the causal model is improperly specified, weighting is unlikely to help.

17. On the So-Called "Huber Sandwich Estimator" and "Robust Standard Errors" 295

In applications where the statistical model is nearly correct, the Huber Sandwich Estimator makes little difference. On the other hand, if the model is seriously in error, the parameters being estimated are likely to be meaningless, except perhaps as descriptive statistics.

18. Endogeneity in Probit Response Models 305

The usual Heckman two-step procedure should not be used for removing endogeneity bias in probit regression. From a theoretical perspective this procedure is unsatisfactory, and likelihood methods are superior. Unfortunately, standard software packages do a poor job of maximizing the biprobit likelihood function, even if the number of covariates is small.

19. Diagnostics Cannot Have Much Power Against General Alternatives 323

Model diagnostics cannot have much power against omnibus alternatives. For instance, the hypothesis that observations are independent cannot be tested against the general alternative that they are dependent with power that exceeds the level of the test. Thus, the basic assumptions of regression cannot be validated from data.

Part IV
Shoe Leather Revisited

20. On Types of Scientific Inquiry: The Role of Qualitative Reasoning 337

Causal inference can be strengthened in fields ranging from epidemiology to political science by linking statistical analysis to qualitative knowledge. Examples from epidemiology show that substantial progress can derive from informal reasoning, qualitative insights, and the creation of novel data sets that require deep substantive understanding and a great expenditure of effort and shoe leather. Scientific progress depends on refuting conventional ideas if they are wrong, developing new ideas that are better, and testing the new ideas as well as the old ones. Qualitative evidence can play a key role in all three tasks.

References and Further Reading 357

Index 393

Preface

David A. Freedman presents in this book the foundations of statistical models and their limitations for causal inference. Examples, drawn from political science, public policy, law, and epidemiology, are real and important.

A statistical model is a set of equations that relate observable data to underlying parameters. The parameters are supposed to characterize the real world. Formulating a statistical model requires assumptions that are routinely untested. Indeed, some are untestable in principle, as Freedman shows in this volume. Assumptions are involved in choosing which parameters to include, the functional relationship between the data and the parameters, and how chance enters the model. It is common to assume that the data are a simple function of one or more parameters, plus random error. Linear regression is often used to estimate those parameters. More complicated models are increasingly common, but all models are limited by the validity of the assumptions on which they ride.

Freedman's observation that statistical models are fragile pervades this volume. Modeling assumptions—rarely examined or even enunciated —fail in ways that undermine model-based causal inference. Because of their unrealistic assumptions, many new techniques constitute not progress but regress. Freedman advocates instead "shoe leather" methods, which identify and exploit natural variation to mitigate confounding and which require intimate subject-matter knowledge to develop appropriate research designs and eliminate rival explanations.

Freedman assembled much of this book in the fall of 2008, shortly before his death. His goal was to offer an integrated presentation of his views on applied statistics, with case studies from the social and health sciences, and to encourage discussion of those views. We made some changes to Freedman's initial selection of topics to reduce length and broaden coverage. The text has been lightly edited; in a few cases chapter titles have been altered. Citations to the original publications are given on the first page of each chapter and in the reference list, which has been consolidated at the end. When available, references to unpublished articles have been updated with the published versions. To alert the reader, chapter numbers have been added for citations to Freedman's works that appear in this book.

Many people deserve acknowledgment for their roles in bringing these ideas and this book to life, including the original co-authors and acknowledged reviewers. Above all, we admire David Freedman's tenacity and lucidity during his final days, and we are deeply grateful for his friendship, collaboration, and tutelage. We thank Janet Macher for her assistance in editing the manuscript. Colleagues at Berkeley and elsewhere contributed valuable suggestions. Ed Parsons of Cambridge University Press helped shape the project and moved it to press with amazing speed.

David Collier, Jasjeet Singh Sekhon, and Philip B. Stark
Berkeley, California
July 2009

Editors' Introduction: Inference and Shoe Leather

David Collier, Jasjeet S. Sekhon, and Philip B. Stark

Drawing sound causal inferences from observational data is a central goal in social science. How is controversial. Technical approaches based on statistical models—graphical models, non-parametric structural equation models, instrumental variable estimators, hierarchical Bayesian models, and the like—are proliferating. But David Freedman has long argued that these methods are not reliable. He demonstrated repeatedly that it can be better to rely on subject matter expertise and to exploit natural variation to mitigate confounding and rule out competing explanations.

When Freedman first enunciated this position decades ago, many were skeptical. They found it hard to believe that a probabilist and mathematical statistician of his stature would favor "low-tech" approaches. But the tide is turning. An increasing number of social scientists now agree that statistical technique cannot substitute for good research design and subject-matter knowledge. This view is particularly common among those who understand the mathematics and have on-the-ground experience.

Historically, "shoe-leather epidemiology" is epitomized by intensive, door-to-door canvassing that wears out investigators' shoes. In contrast, advocates of statistical modeling sometimes claim that their methods can salvage poor research design or low-quality data. Some suggest that their algorithms are general-purpose inference engines: Put in data, turn the crank, out come quantitative causal relationships, no knowledge of the subject required.

This is tantamount to pulling a rabbit from a hat. Freedman's conservation of rabbits principle says "to pull a rabbit from a hat, a rabbit must first be placed in the hat."[1] In statistical modeling, assumptions put the rabbit in the hat.

Modeling assumptions are made primarily for mathematical convenience, not for verisimilitude. The assumptions can be true or false—usually false. When the assumptions are true, theorems about the methods hold. When the assumptions are false, the theorems do not apply. How well do the methods behave then? When the assumptions are "just a little wrong," are the results "just a little wrong?" Can the assumptions be tested empirically? Do they violate common sense?

Freedman asked and answered these questions, again and again. He showed that scientific problems cannot be solved by "one-size-fits-all" methods. Rather, they require shoe leather: careful empirical work tailored to the subject and the research question, informed both by subject-matter knowledge and statistical principles. Witness his mature perspective:

> Causal inferences can be drawn from nonexperimental data. However, no mechanical rules can be laid down for the activity. Since Hume, that is almost a truism. Instead, causal inference seems to require an enormous investment of skill, intelligence, and hard work. Many convergent lines of evidence must be developed. Natural variation needs to be identified and exploited. Data must be collected. Confounders need to be considered. Alternative explanations have to be exhaustively tested. Before anything else, the right question needs to be framed.
>
> Naturally, there is a desire to substitute intellectual capital for labor. That is why investigators try to base causal inference on statistical models. The technology is relatively easy to use, and promises to open a wide variety of questions to the research effort. However, the appearance of methodological rigor can be deceptive. The models themselves demand critical scrutiny. Mathematical equations are used to adjust for confounding and other sources of bias. These equations may appear formidably precise, but they typically derive from many somewhat arbitrary choices. Which variables to enter in the regression? What functional form to use? What assumptions to make about parameters and error terms? These choices are seldom dictated either by data or prior scientific knowledge. That is why judgment is so critical, the opportunity for error so large, and the number of successful applications so limited.[2]

Causal inference from randomized controlled experiments using the intention-to-treat principle is not controversial—provided the inference is based on the actual underlying probability model implicit in the randomization. But some scientists ignore the design and instead use regression to analyze data from randomized experiments. Chapters 12 and 13 show that the result is generally unsound.

Nonexperimental data range from "natural experiments," where Nature provides data as if from a randomized experiment, to observational studies where there is not even a comparison between groups. The epitome of a natural experiment is Snow's study of cholera, discussed in Chapters 3 and 20. Snow was able to show—by expending an enormous amount of shoe leather—that Nature had mixed subjects across "treatments" in a way that was tantamount to a randomized controlled experiment.

To assess how close an observational study is to an experiment requires hard work and subject-matter knowledge. Even without a real or natural experiment, a scientist with sufficient expertise and field experience may be able to combine case studies and other observational data to rule out possible confounders and make sound inferences.

Freedman was convinced by dozens of causal inferences from observational data—but not hundreds. Chapter 20 gives examples, primarily from epidemiology, and considers the implications for social science. In Freedman's view, the number of sound causal inferences from observational data in epidemiology and social sciences is limited by the difficulty of eliminating confounding. Only shoe leather and wisdom can tell good assumptions from bad ones or rule out confounders without deliberate randomization and intervention. These resources are scarce.

Researchers who rely on observational data need qualitative and quantitative evidence, including case studies. They also need to be mindful of statistical principles and alert to anomalies, which can suggest sharp research questions. No single tool is best: They must find a combination suited to the particulars of the problem.

Freedman taught students—and researchers—to evaluate the quality of information and the structure of empirical arguments. He emphasized critical thinking over technical wizardry. This focus shines through two influential textbooks. His widely acclaimed undergraduate text, *Statistics*,[3] transformed statistical pedagogy. *Statistical Models: Theory and Practice*,[4] written at the advanced undergraduate and graduate level, presents standard techniques in statistical modeling and explains their shortcomings. These texts illuminate the sometimes tenuous relationship between statistical theory and scientific applications by taking apart serious examples.

The present volume brings together twenty articles by David Freedman and co-authors on the foundations of statistics, statistical modeling, and causal inference in social science, public policy, law, and epidemiology. They show when, why, and by how much statistical modeling is likely to fail. They show that assumptions are not a good substitute for subject-matter knowledge and relevant data. They show when qualitative, shoe-leather approaches may well succeed where modeling will not. And they point out that in some situations, the only honest answer is, "we can't tell from the data available."

This book is the perfect companion to *Statistical Models*. It covers some of the same topics in greater depth and technical detail and provides more case studies and close analysis of newer and more sophisticated tools for causal inference. Like all of Freedman's writing, this compilation is engaging and a pleasure to read: vivid, clear, and dryly funny. He does not use mathematics when English will do. Two-thirds of the chapters are relatively non-mathematical, readily accessible to most readers. The entire book—except perhaps a few proofs—is within the reach of social science graduate students who have basic methods training.

Freedman sought to get to the bottom of statistical modeling. He showed that sanguine faith in statistical models is largely unfounded. Advocates of modeling have responded by inventing escape routes, attempts to rescue the models when the underlying assumptions fail. As Part III of this volume makes clear, there is no exit: The fixes ride on *other* assumptions that are often harder to think about, justify, and test than those they replace.

This volume will not end the modeling enterprise. As Freedman wrote, there will always be "a desire to substitute intellectual capital for labor" by using statistical models to avoid the hard work of examining problems in their full specificity and complexity. We hope, however, that readers will find themselves better informed, less credulous, and more alert to the moment the rabbit is placed in the hat.

Notes

1. See, e.g., Freedman and Humphreys (1999), p. 102.

2. Freedman (2003), p. 19. See also Freedman (1999), pp. 255–56.

3. David Freedman, Robert Pisani, and Roger Purves (2007). *Statistics*, 4th edn. New York: Norton.

4. David A. Freedman (2009). *Statistical Models: Theory and Practice*, rev. edn. New York: Cambridge.