Statistical Models and Causal Inference

A Dialogue with the Social Sciences

David A. Freedman

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

Cambridge University Press
2009

Contents

Editors' Introduction: Inference and Shoe Leather vii

Part I
Statistical Modeling: Foundations and Limitations 1

1. Some Issues in the Foundations of Statistics:
Probability and Model Validation 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.

Part II
Studies in Political Science, Public Policy, and Epidemiology 65

4. Methods for Census 2000 and Statistical Adjustments 67
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 unanimously supported the decision by the Secretary of Commerce not to adjust.

5. On "Solutions" to the Ecological Inference Problem 85
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 99
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 107
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? 117
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 133
Experimental evidence suggests that the effect of a large reduction in salt intake on blood pressure is modest, and 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:
Relative Risk and Specific Causation 155
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? 173
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? 197

12. On Regression Adjustments in
Experiments with Several Treatments 199
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 223
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 247
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 259
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 283
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" 299
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 309
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 327

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 339

20. On Types of Scientific Inquiry: The Role of Quantitative Reasoning 341

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 361

Index 383

Inference and Shoe Leather

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

Drawing sound causal inferences is a central goal in social science. How to do that is controversial. Technical approaches to inference 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. Moreover, he demonstrates repeatedly the superiority of "shoe leather" approaches, which exploit natural variation to mitigate confounding, and rely on intimate knowledge of the subject matter to develop meticulous designs and exhaust other explanations.

When Freedman first enunciated this position decades ago, he was met with skepticism, in part because it was 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 with both the mathematical skill to understand the models, and on-the-ground experience.

Historically, "shoe-leather epidemiology" is epitomized by intensive, door-to-door canvassing that wears out the investigators' shoes. In contrast, advocates of statistical modeling sometimes claim that their methods can be a substitute for careful research design and painstaking data collection. Some even claim—explicitly or tacitly—that their algorithms can infer causal structure automatically, without requiring subject-specific knowledge.

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 are solved case by case, with lots of shoe leather and furrowing of the brow. Witness his mature perspective:

> Causal inferences can be drawn from non-experimental 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, the model implicit in the randomization. But some scientists ignore the true probability model, and instead use regression to analyze data from randomized experiments. Chapters 12 and 13 show that the result is generally unsound.

Non-experimental 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 validate the degree to which an observational study is like an experiment requires hard work and subject matter knowledge. Even when nature does not deliver a natural experiment, well chosen case studies and other observational data, combined with substantive expertise and experience, can help rule out possible confounders and lead to 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 without deliberate randomization and intervention. Only shoe leather and substantive wisdom can tell good assumptions from bad ones or rule out confounders without randomization and intervention. These resources are scarce.

Researchers working with observational data need a mix of qualitative and quantitative evidence, including case studies. Researchers need to be alert to anomalies, which can suggest sharp research questions. No single tool is best: researchers 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 Applications*,[4] written at the advanced undergraduate and graduate level, presents standard techniques in statistical modeling and 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 20 articles[5] by David Freedman on 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, with puckish humor. Freedman does not use mathematics when English will do. Two-thirds of the chapters are relatively non-mathematical, readily accessible to most readers. The remaining third are accessible to social science graduate students who have a basic level of 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—techniques for fixing the models when the underlying assumptions fail. As Part III of this volume makes clear, there is no exit. Attempts to rescue models from violations of assumptions 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. 255–6.

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.

5. The articles have been edited a little: Citations to unpublished material have been replaced where possible by citations to equivalent published articles, and references are at the end of the volume.