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.