July 1990

STATISTICAL MODELS AND SHOE LEATHER

by

D. A. Freedman
Statistics Department
University of California
Berkeley, Calif 94720
415-642-2781

with discussion by

R. Bodkin, M. Intriligator, D. Rothman, W. Mason

Technical Report No. 217
Statistics Department
University of California
Berkeley, Ca 94720

Research partially supported by
NSF Grant DMS 86-01634 and by the
Miller Institute for Basic Research

July 1990

STATISTICAL MODELS AND SHOE LEATHER

by

D. A. Freedman[1]
Statistics Department
University of California
Berkeley, Calif 94720

To appear in Sociological Methodology 1991
Peter Marsden, editor

Abstract

Regression models have been used in the social sciences at least since 1899, when Yule published a paper on the causes of pauperism. Regression models are now used to make causal arguments in a wide variety of applications, and it is perhaps time to try to evaluate the results. No definitive answers can be given, but this paper takes a rather negative view. Snow's work on cholera is presented as a success story for scientific reasoning based on non-experimental data. Failure stories are also discussed, and comparisons may provide some insight. In particular, the suggestion is made that statistical technique can seldom be an adequate substitute for good design, collection of relevant data, and testing predictions against reality in a variety of settings.

---

[1]Research partially supported by NSF Grant DMS 86-01634; and by the Miller Institute for Basic Research. Much help was provided by Richard Berk, John Cairns, David Collier, Persi Diaconis, Sander Greenland, Steve Klein, Jan de Leeuw, Thomas Rothenberg and Amos Tversky. Particular thanks go to Peter Marsden.

STATISTICAL MODELS AND SHOE LEATHER

Introduction

Regression models have been used in the social sciences at least since 1899, when Yule published a paper in the Journal of the Royal Statistical Society on changes in "out-relief" as a cause of pauperism: he argued that providing income support outside the poor-house increased the number of people on relief. At present, regression models are used to make causal arguments in a wide variety of social-science applications, and it is perhaps time to try to evaluate the results.

A crude four-point scale may be useful. Regression is a method which:

1) usually works, although it is (like anything else) imperfect and may sometimes go wrong;

2) sometimes works in the hands of skillful practitioners, but isn't suitable for routine use;

3) might work, but hasn't yet;

4) can't work.

Textbooks, courtroom testimony, and newspaper interviews often seem to put regression into category 1). Option 4) seems too pessimistic. My own view is bracketed by 2) and 3), although good examples are quite hard to find.

Regression modeling is a dominant paradigm; and many investigators seem to consider that any piece of empirical research has to be equivalent to a regression model. Questioning the value of regression is then tantamount to denying the value of data. Some declarations of faith may therefore be necessary. Social science is possible, and sound conclusions can be drawn from non-experimental data. (Experimental confirmation is always welcome, although some experiments have problems of their own.) Statistics can play a useful role. With multi-dimensional data sets, regression may provide helpful summaries of the data.

However, I do not think that regression can carry much of the burden in a causal argument. Nor do regression equations, by themselves, give much help in controlling for confounding variables. Arguments based on "statistical significance" of coefficients seem generally suspect; so do causal interpretations of coefficients. More recent developments, like two-stage least squares, latent-variable modeling, and specification tests, may be quite interesting. However, technical fixes will not solve the problems, which are at a deeper level. In the end, I see many illustrations of technique but few real examples with validation of the modeling assumptions.

Indeed, causal arguments based on significance tests and regression are almost necessarily circular. To derive a regression model, we need an elaborate theory which specifies the variables in the system, their causal inter-connections, the functional form of the relationships, and the statistical properties of the error terms-- independence, exogeneity, etc. (The stochastics may not matter for descriptive purposes, but they are crucial for significance tests.) Given the model, least squares and its variants can be used to estimate parameters and decide whether these are zero or not. However, the model cannot in general be regarded as given, because current social science theory does not provide the requisite level of technical detail for deriving specifications.

There is an alternative validation strategy, which is less dependent on prior theory: Take the model as a black box, and test it against empirical reality. Does the model predict new phenomena? does it predict the results of interventions? are the predictions right? The usual statistical tests are poor substitutes, because they rely on strong maintained hypotheses. Without the right kind of theory, or reasonable empirical validation, the conclusions drawn from the models must be quite suspect.

At this point, it may be natural to ask for some real examples of good empirical work, and strategies for research that do not involve regression. Illustrations from epidemiology may be useful; the problems in that field are quite similar to those faced by contemporary workers in the social sciences. Snow's work on cholera will be reviewed, as an example of real science based on observational data; regression is not involved.

A comparison will be made with some current regression studies in epidemiology and social science. This may give some insight into the weaknesses of regression methods. The possibility of technical fixes for the models will be discussed, other literature will be reviewed, and then some tentative conclusions will be drawn.

Some examples from epidemiology

Quantitative methods in the study of disease precede Yule-- and regression. In 1835, Pierre Louis published a landmark study on bleeding as a cure for pneumonia. He compared outcomes for groups of pneumonia patients who had been bled at different times, and found

> "That bloodletting has a happy effect on the progress of pneumonitis; that it shortens its duration; and this effect, however, is much less than has been commonly believed...." [Louis, p48, 1986 edition]

The finding, and the statistical method, were roundly denounced by contemporary physicians:

> "By invoking the inflexibility of arithmetic in order to escape the encroachments of the imagination, one commits an outrage upon good sense...."
> [Louis, p63 footnote, 1986 edition]

Louis may have started a revolution in our thinking about empirical research in medicine, or his book may only provide a convenient line of demarcation. But there is no doubt that within a few decades, the "inflexibility of arithmetic" had helped identify the causes of some major diseases and the means for their prevention; statistical modeling played almost no role in these developments.

In the 1850s, John Snow demonstrated that cholera was a water-borne infectious disease (Snow, 1855). A few years later, Ignaz Semmelweiss discovered how to prevent puerperal fever (Semmelweiss, 1861). Around 1914, Joseph Goldberger found the cause of pellagra (Carpenter, 1981; Terris, 1964). Later epidemiologists have shown, at least on balance of argument, that most lung cancer is caused by smoking; two early papers are Lombard & Doering (1928), and Mueller (1939); Cornfield et al. (1959) and U.S. Public Health Service (1964) review the evidence. In epidemiology, careful reasoning on observational data has lead to considerable progress. (For failure stories in epidemiology, see below.)

An explicit definition of good research methodology seems elusive; but an implicit definition is possible, by pointing to examples. In that spirit, I give a brief account of Snow's work. To see his achievement, I ask you to go back in time: forget that germs cause disease. Microscopes are available but their resolving power is poor; most human pathogens cannot be seen. Clear ideas about isolating such micro-organisms lie decades into the future. The infection theory has some supporters, but the dominant idea is that disease results from "miasmas": minute, inanimate poison particles in the air. (Poison in the ground is a later variant.)

Snow was studying cholera, which had arrived in Europe in the early 1800s. Cholera came in epidemic waves, attacked its victims suddenly, and was often fatal. Early symptoms were vomiting and acute diarrhea. Based on the clinical course of the disease, Snow conjectured that the active agent was a living organism, which got into the alimentary canal with food or drink, multiplied in the body, and generated some poison which caused the body to expel water. The organism passed out of the body with these evacuations, got back into the water, and infected new victims.

Snow marshalled a series of persuasive arguments for this conjecture. For example, cholera spreads along the tracks of human commerce. If a ship goes from a cholera-free country to a cholera-stricken port, the sailors get the disease only after they land or take on supplies. The disease strikes hardest at the poor, who live in the most crowded housing with the worst hygiene. These facts are consistent with the infection theory and hard to explain on the miasma theory.

Snow also did a lot of scientific detective work. In one of the earliest epidemics in England, he was able to identify the first case, "a seaman named John Harnold, who had newly arrived by the Elbe steamer from Hamburgh, where the disease was prevailing." Snow also found the second case-- who had taken the room where Harnold had stayed (No. 8, New Lane, Gainsford Street, Horsleydown). More evidence for the infection theory.

In later epidemics, Snow went on to develop even better evidence. For example, he found two adjacent apartment buildings; one was heavily hit by cholera, the other not. He could show that the first building had a water supply contaminated by run-off from its privies; the second building had much cleaner drinking water. He also made several "ecological" studies to demonstrate the influence of water supply on cholera incidence. In the London of the 1800s, there were many different water companies serving different areas of the city; some areas were served by more than one company. Several companies took their water from the Thames, which was heavily polluted by sewage. Such companies had much higher rates of cholera in their service areas. The Chelsea water company was an exception-- but it had an exceptionally good filtration system.

In the epidemic of 1853-54, Snow made a "spot map" showing where the cases occurred, and they clustered around the Broad Street pump. He identified the pump as a source of contaminated water and persuaded the public authorities to remove the handle. As the story goes, removing the handle stopped the epidemic and proved Snow's theory. In fact, he did get the handle removed and the epidemic did stop. However, as he demonstrated with some clarity, the epidemic was stopping anyway; and he attached little weight to the episode.

For our purposes, what Snow actually did in 1853-4 is even more interesting than the fable. For example, there was a large poor-house in the Broad Street area, with few cholera cases. Why? Snow's answer was that the poor-house had its own well and the inmates did not take water from the pump. There was also a large brewery, with no cases. The reason is obvious; the workers drank beer, not water. (But if any wanted water, there was a well on these premises too.)

To set up Snow's main argument, I have to back up just a bit. In 1849, the Lambeth water company had moved its intake point upstream along the Thames, above the main sewage discharge points, so that its water was fairly pure. The Southwark and Vauxhall company, however, left its intake point downstream from the sewage discharges. An ecological analysis of the data for the epidemic of 1853-4 showed that cholera hit harder in the Southwark and Vauxhall service areas, and largely spared the Lambeth areas. Now let Snow finish in his own words (pp.74-5):

> "Although the facts shown in the above table [the ecological data] afford very strong evidence of the powerful influence which the drinking of water containing the sewage of a town exerts over the spread of cholera, when that disease is present, yet the question does not end here; for the intermixing of the water supply of the Southwark and Vauxhall Company with that of the Lambeth Company, over an extensive part of London, admitted of the subject being sifted in such a way as to yield the most incontrovertible proof on one side or the other. In the subdistricts enumerated in the above table as being supplied by both Companies, the mixing of the supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. A few houses are supplied by one Company and a few by the other, according to the decision of the owner or occupier at that time when the Water Companies were in active competition. In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies. Now it must be evident that, if the diminution of cholera, in the districts partly supplied with improved water, depended on this supply, the houses receiving it would be the houses enjoying the whole benefit of the diminution of the malady, whilst the houses supplied with the water from Battersea Fields would suffer the same mortality as they would if the improved supply did not exist at all. As there is no difference whatever either in the houses or the people receiving the supply of the two Water Companies, or in any of the physical conditions with which they are surrounded, it is obvious that no experiment could have been devised which would more thoroughly test the effect of water supply on the progress of cholera than this, which circumstances placed ready made before the observer.

> "The experiment, too, was on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and in most cases, without their knowledge; one group being supplied with water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity.
>
> "To turn this grand experiment to account, all that was required was to learn the supply of water to each individual house where a fatal attack of cholera might occur...."

Snow identified the companies supplying water to the houses of cholera victims in his study area. This gave him the numerators in the table below. (The denominators were available from parliamentary records.)

Snow's Table IX

	Number of houses	Deaths from cholera	Deaths in each 10,000 houses
Southwark and Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

Snow concluded that <u>if</u> the Southwark and Vauxhall company had moved their intake point as Lambeth did, about 1,000 lives would have been saved. He was very clear about quasi-randomization as the control for potential confounding variables. He was equally clear about the difference between ecological correlations and individual correlations. And his counterfactual inference is compelling.

The table is by no means remarkable, as a piece of statistical technology. But the story it tells is very persuasive. The force of the argument results from the clarity of the prior reasoning, the bringing together of many different lines of evidence, and the amount of shoe leather Snow was willing to use up while getting the data.

Later, there was to be more confirmation of Snow's conclusions. For example, the cholera epidemics of 1832 and 1849 in New York were handled by traditional methods: exhorting the population to temperance, bringing in pure water to wash the streets, treating the sick by bleeding and mercury. After the publication of Snow's book, the epidemic of 1866 was dealt with using the methods suggested by his theory: boiling the drinking water; isolating sick individuals and disinfecting their evacuations. The death rate was cut by a factor of 10 or more (Rosenberg, 1962).

In Hamburg, there was an epidemic in 1892. The leaders of Hamburg rejected Snow's arguments; they followed Max von Pettenkofer, who taught the miasma theory-- contamination of the ground caused cholera. As a result, Hamburg paid little attention to its water supply, but spent a great deal of effort digging up and carting away carcasses buried by slaughter houses. The results were disastrous (Evans, 1987).

What about evidence from microbiology? In 1880, Pasteur created a sensation by showing that the cause of rabies was a micro-organism. In 1884, Koch isolated the cholera vibrio, confirming all the essential features of Snow's account; Filipo Pacini may have discovered this organism even earlier; see Howard-Jones (1975). The vibrio is a water-borne bacterium, which invades the human gut and causes cholera. Today, the molecular biology of cholera is reasonably well understood (Finlay et al., 1989; Miller et al., 1989). The vibrio makes protein enterotoxin, which affects the metabolism of human cells, and causes them to expel water. The interaction of enterotoxin with the cell has been worked out, and so has the genetic mechanism used by the vibrio to manufacture this protein.

Snow did some brilliant detective work on non-experimental data. What is impressive is not the statistical technique, but the handling of the scientific issues. There is steady progress from shrewd observation through case studies to analysis of ecological data. In the end, he found and analyzed a natural experiment. (Of course, he also made his share of mistakes: for example, based on rather flimsy analogies, he concluded that plague and yellow fever were also propagated through the water; pp125-7.)

The next example is from modern epidemiology, which has adopted regression methods. The example shows how modeling can go off the rails. In 1980, Kanarek et al. published an article in the American Journal of Epidemiology-- perhaps the leading journal in the field-- arguing that asbestos fibers in the drinking water caused lung cancer. The study was based on 722 census tracts in the San Francisco Bay Area. There were huge variations in fiber concentrations from one tract to another; factors of 10 or more were commonplace.

This study examined cancer rates at 35 sites, for blacks and whites, men and women. It controlled for age by standardization, for sex and race by cross-tabulation. But the main tool was log-linear regression, to control for other covariates (marital status, educational level, income, occupation). Causation was inferred, as usual, if a coefficient was statistically significant after controlling for covariates.

The paper has no discussion of its stochastic assumptions, that outcomes are independent and identically distributed given covariates. The argument for the functional form was only that "theoretical construction of the probability of developing cancer by a certain time yields a function of the log form." However, this model of cancer causation is open to serious objections (Freedman & Navidi, 1989).

For lung cancer in white males, the asbestos fiber coefficient was highly significant ($P<.001$), so the effect was described as "strong". Actually, the model only predicts a risk multiplier of about 1.05 for a 100-fold increase in fiber concentrations. There was no effect in women or blacks. Moreover, Kanarek et al. had no data on cigarette smoking, which affects lung cancer rates by factors of 10 or more. So imperfect control over smoking could easily account for the observed effect, as could even minor errors in functional form. Finally, Kanarek et al. ran upwards of 200 equations; only one of the P-values was below .001. So the real significance level may be closer to $200 \times .001 = .20$. The model-based argument is not a good one.

What is the difference between Kanarek et al. and Snow? Kanarek et al. ignore the ecological fallacy; Snow dealt with it. Kanarek et al. try to control for covariates by modeling, with socio-economic status as a proxy for smoking; Snow found a natural experiment and collected the data he needed. Kanarek et al.'s argument for causation rides on the statistical significance of a coefficient; Snow's argument used logic and shoe leather. Regression models make it all too easy to substitute technique for work.

Some examples from social science

Now, some contemporary social science applications. If regression is a successful methodology, the routine paper in a good journal should be a modest success story. However, the situation is quite otherwise. I recently spent some time looking through leading American journals in quantitative social science: American Journal of Sociology, American Sociological Review, and The American Political Science Review. These refereed journals accept perhaps 10% of the submissions. For analysis, I selected papers which were published in 1987-88, which posed reasonably clear research questions, and which used regression to answer them. I will discuss three of these papers here. These papers may not be the best of their kind, but they are far from the worst; indeed, one was later awarded a prize "For the best article published in The American Political Science Review during 1988." In sum, I believe these papers are quite typical of good current research practice.

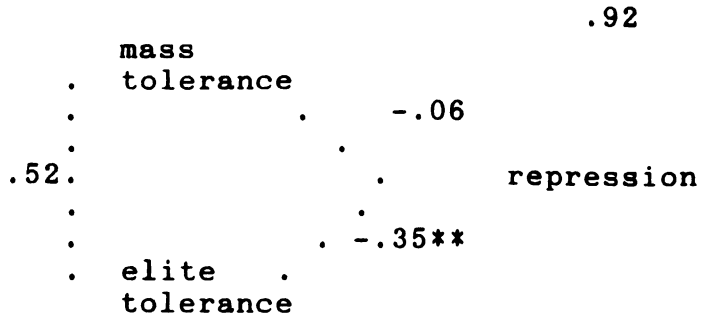
Example 1. Bahry & Silver (1987) hypothesized that in Russia, perception of the KGB as efficient deterred political activism. The study was based on questionnaires filled out by Russian emigres in New York. There was a lot of missing data, and perhaps some confusion between response variables and control variables. Leave all that aside. In the end, the argument was that after adjustment for covariates, subjects who viewed the KGB as efficient were less likely to describe themselves as activists. And this negative correlation was statistically significant.

Of course, that could be evidence to support the research hypothesis of the paper: if you think the KGB is efficient, you don't demonstrate. Or the line of causality could run the other way: if you're an activist, you find out the KGB is inefficient. Or the association could be driven by a third variable: people of certain personality types are more likely to describe themselves as activists, and also more likely to describe the KGB as inefficient. Correlation is not the same as causation; statistical technique, alone, does not make the connection. The familiarity of this point should not be allowed to obscure its force.

Example 2. Erikson et al. (1987) argued that in the U.S., different states really do have different political cultures. After controlling for demographics and geographical regions, adding state dummies increased R^2 for predicting party identification from .0898 to .0953. The F to enter the state dummies was about 8. The data base consisted of 55,000 CBS/New York Times questionnaires. With 40 degrees of freedom in the numerator and 55,000 in the denominator, P is spectacular.

On the other hand, at the risk of the obvious, the R^2s are trivial-- never mind the increase. The authors argued that the state dummies are not proxies for omitted variables. As proof, they put in trade union membership and the estimated state effects did not change much. This is an argument in support of the specification, but a weak one.

Example 3. Gibson (1988) asked whether political intolerance during the McCarthy era was driven by mass opinion or elite opinion. The unit of analysis was the state. Legislation was coded on a tolerance/intolerance scale; there were questionnaire surveys of elite opinion and mass opinion. Then comes a path model; one coefficient is significant, one is not.



Gibson concludes, "Generally it seems that elites, not masses, were responsible for the repression of the era."

Of the three papers, I thought this one had the clearest question and the best summary data. However, the path diagram seems to be an extremely weak causal model. Moreover, even granting the model, the difference between the two path coefficients is not significant. The paper's conclusion does not follow from its data.

Summary of the position

In this set of papers, and in many papers outside the set, the adjustment for covariates is by regression; the argument for causality rides on the significance of a coefficient. But significance levels depend on specifications, especially of error structure. For example, if the errors are correlated or heteroscedastic, the conventional formulas will give the wrong answers. And the stochastic specification is never argued in any detail. (Nor does modeling the covariances fix the problem, unless the model for the covariances can be validated; more about technical fixes, below.)

To sum up, in each of the examples:

- There is an interesting research question, which may or may not be sharp enough to be empirically testable.
- Relevant data are collected, although there may be considerable difficulty in quantifying some of the concepts, and important data may be missing.
- The research hypothesis is quickly translated into a regression equation; more specifically, into an assertion that certain coefficients are (or are not) statistically significant.

- Some attention is paid to getting the right variables into the equation, although the choice of covariates is usually not compelling.

- Little attention is paid to functional form or stochastic specification; textbook linear models are just taken for granted.

Clearly, evaluating the use of regression models in a whole field is a difficult business; there are no well-beaten paths to follow. Here, I have selected for review three papers which, in my opinion, are good of their kind and fairly represent a large (but poorly delineated) class. These papers illustrate some basic obstacles in applying regression technology to make causal inferences.

In Freedman (1987), I took a different approach and reviewed a modern version of the classic model for status attainment. I tried to state the technical assumptions needed for drawing causal inferences from path diagrams-- assumptions which seem to be very difficult to validate in applications. I also summarized previous work on these issues. Modelers had an extended opportunity to answer. The technical analysis was not in dispute and serious examples were not forthcoming.

If the assumptions of a model are not derived from theory, and predictions are not tested against reality, then deductions from the model must be quite shaky. However, without the model, the data cannot be used to answer the research question. Indeed, the research hypothesis may not really be translatable into an empirical claim except as a statement about nominal significance levels of coefficients in a model.

Two authorities may be worth quoting in this regard; of course, both of them have said other things in other places:

> "The aim...is to provide a clear and rigorous basis for determining when a causal ordering can be said to hold between two variables or groups of variables in a model.... The concepts...all refer to a model-- a system of equations-- and not to the 'real' world the model purports to describe."
> --Simon (1957, p12, emphasis added)

> "If...we choose a group of social phenomena with no antecedent knowledge of the causation or absence of causation among them, then the calculation of correlation coefficients, total or partial, will not advance us a step toward evaluating the importance of the causes at work." --Fisher (1958, p190)

In my view, regression models are not a particularly good way of doing empirical work in the social sciences today, because the technique depends on knowledge that we do not have. Investigators who use the technique are not paying adequate attention to the connection-- if any-- between the models and the phenomena they are studying. Their conclusions may be valid for the computer code they have created, but the claims are hard to transfer from that microcosm to the larger world.

For me, Snow's work exemplifies one point on a continuum of research styles; the regression examples mark another. My judgment on the relative merits of the two styles will be clear; and with it, some implicit recommendations. Comparisons may be invidious, but I think Snow's research stayed much closer to reality than do the modeling exercises. He was not interested in the properties of systems of equations, but in ways of preventing a real disease. He formulated sharp, empirical questions which could be answered using data that could, with effort, be collected. At every turn, he anchored his argument in stubborn fact. And he exposed his theory to harsh tests in a variety of settings. That may explain how he could discover something extraordinarily important about cholera; and why his book is still worth reading more than a century later.

Can technical fixes rescue the models?

Regression models often seem to be used to compensate for problems in measurement, data collection, and study design. By the time the models are deployed, the scientific position is nearly hopeless. Reliance on models in such cases is Panglossian. At any rate, that is my view. By contrast, some readers may be concerned to defend the technique of regression modeling: according to them, the technique is sound and only the applications are flawed. Other readers may think the criticisms of regression modeling are merely technical, so that technical fixes-- e.g., robust estimators, generalized least squares, and specification tests-- will make the problems go away.

The mathematical basis for regression is well established. My question is whether the technique applies to present-day social science problems. In other words, are the assumptions valid? Moreover, technical fixes become relevant only when models are nearly right. For instance, robust estimators may be useful if the error terms are independent, identically distributed, and symmetric but long-tailed. If the error terms are neither independent nor identically distributed, and there is no way to find out whether they are symmetric, robust estimators are probably a distraction from the real issues.

This point is so uncongenial that another illustration may be in order. Suppose $y_i = \alpha+\epsilon_i$, the ϵ_i have mean 0, and the ϵ_i are either independent and identically distributed or autoregressive of order 1. Then the well-oiled statistics machine springs into action. However, if the ϵ_i are just a sequence of random variables, the situation is nearly hopeless-- with respect to standard errors and hypothesis testing. So much the worse if the y_i have no stochastic pedigree. The last possibility seems to me the most realistic. Then formal statistical procedures are irrelevant, and we are reduced (or should be) to old-fashioned thinking.

A well-known discussion of technical fixes starts from the evaluation of manpower training programs using non-experimental data. LaLonde (1986) and Fraker & Maynard (1987) compare evaluation results from modeling with results from experiments. The idea is to see whether regression models fitted to observational data can predict the results of experimental interventions. Fraker & Maynard conclude:

> "The results indicate that nonexperimental designs cannot be relied on to estimate the effectiveness of employment programs. Impact estimates tend to be sensitive both to the comparison group construction methodology and to the analytic model used. There is currently no way a priori to ensure that the results of comparison group studies will be valid indicators of the program impacts." [Fraker & Maynard, 1987, p194]

Heckman & Holtz (1989) reply that specification tests can be used to rule out models that give wrong predictions:

> "....a simple testing procedure eliminates the range of nonexperimental estimators at variance with the experimental estimates of program impact.... Thus, while not definitive, our results are certainly encouraging for the use of nonexperimental methods in social-program evaluation."

Heckman & Holtz have in hand a) the experimental data, b) the non-experimental data, and c) the results of LaLonde and Fraker & Maynard. Heckman & Holtz proceed by modeling the selection bias in the non-experimental comparison groups. There are three types of models, each with two main variants. These are fitted to several different time periods, with several sets of control variables. Averages of different models are allowed, and there is a "slight extension" of one model.

By my count, 24 models are fitted to the non-experimental data on female AFDC recipients, and 32 to the data on high-school dropouts. Ex post, Heckman & Holtz have found that models which pass certain specification tests can more or less reproduce the experimental results (up to very large standard errors). However, the real question is what can be done ex ante-- before the right estimate is known. Heckman & Holtz may have an argument, but it is not a strong one; and it may even point us in the wrong direction. Testing one model on 24 different data sets could open a serious enquiry: have we identified an empirical regularity which has some degree of invariance? Testing 24 models on one data set is less serious.

Generally, I think replication and prediction of new results provide a harsher and more useful validation regime than statistical testing of many models on one data set. A partial list of reasons: fewer assumptions are needed; there is less chance of artifact; more kinds of variation can be explored; more alternative explanations can be ruled out. Indeed, taken to the extreme, developing a model by specification tests just comes back to curve fitting-- with a complicated set of constraints on the residuals.

Given the limits to present knowledge, I doubt that models can be rescued by technical fixes. Arguments about the theoretical merit of regression or the asymptotic behavior of specification tests for picking one version of a model rather than another seem like arguments about how to build desalination plants with cold fusion as the energy source. The concept may be admirable, the technical details may be fascinating, but thirsty people should look elsewhere.

Other literature

The issues raised here are hardly new, and this sections reviews some recent literature. No brief summary can do justice to Lieberson (1985), who presents a complicated and subtle critique of current empirical work in the social sciences. A crude paraphrase of one important message: when there are significant differences between comparison groups in an observational study, it is extraordinarily difficult if not impossible to achieve balance by statistical adjustments. Arminger & Bohrnstedt (1987) respond by describing this as a special case of "misspecification of the mean structure caused by the omission of relevant causal variables", and citing literature on that topic.

This trivializes the problem, and almost endorses the idea of fixing misspecification by elaborating the model. However, this idea is unlikely to work. Current specification tests need independent, identically distributed observations, and lots of them; the relevant variables must be identified; some variables must be taken as exogenous; additive errors are needed; and a parametric or semi-parametric form for the mean function is required. These ingredients are rarely found in the social sciences, except by assumption. To model a bias, we need to know what causes it, and how. In practice, this may be even more difficult than the original research question. Some empirical evidence is provided by the discussion of manpower training program evaluations, above; also see Stolzenberg & Relles (1990).

As Arminger & Bohrnstedt concede (1987, p370):

> "There is no doubt that experimental data are to be preferred over nonexperimental data, which practically demand that one knows the mean structure except for the parameters to be estimated."

In the physical or life sciences, there are some situations where the mean function is known, and regression models are correspondingly useful. In the social sciences, I do not see this pre-condition for regression modeling as being met, even to a first approximation.

In commenting on Lieberson (1985), Singer & Marini (1987) emphasize two points:

> a) "it requires rather yeoman assumptions or unusual phenomena to conduct a comparative analysis of an observational study as though it represented conclusions (inferences) from an experiment";
> b) "there seems to be an implicit view in much of social science that any question that might be asked about a society is answerable in principle."

In my view, point a) says that in the current state of knowledge in the social sciences, regression models are seldom if ever reliable for causal inference. With respect to point b), it is exactly the reliance on models which make all questions seem "answerable in principle", and that is a great obstacle to the development of the subject. It is the beginning of scientific wisdom to recognize that not all questions have answers; for some discussion along these lines, see Lieberson (1988).

Marini & Singer (1988) continue the argument:

> "Few would question that the use of 'causal' models has improved our knowledge of causes and is likely to do so increasingly as the models are refined and become more attuned to the phenomena under investigation." [p394]

However, much of the analysis in Marini & Singer (1988) contradicts this presumed majority view. For example:

> "causal analysis.... is not a way of deducing causation but of quantifying already hypothesized relationships.... information external to the model is needed to warrant the use of one specific representation as truly 'structural'. The information must come from the existing body of knowledge relevant to the domain under consideration." [pp 388 and 391]

As I read the current empirical research literature, causal arguments depend mainly on the statistical significance of regression coefficients. If so, Marini & Singer are pointing to the fundamental circularity in the regression strategy--the information needed for building regression models only comes from such models. Indeed, as these authors continue,

> "The relevance of causal models to empirical phenomena is often open to question because assumptions made for the purpose of model identification are arbitrary or patently false. The models take on an importance of their own, and convenience or elegance in the model building overrides faithfulness to the phenomena." [p392]

Holland (1988) raises similar points. Causal inferences from non-experimental data using path models require assumptions that are quite close to the conclusions; so the analysis is driven by the model not the data. In effect, given a set of covariates, the mean response over the 'treatment group' minus the mean over the 'controls' must be assumed to equal the causal effect being estimated [p481]. As Holland says,

> "....the effect...cannot be estimated by the usual regression methods of path analysis without making untestable assumptions about the counterfactual regression function...." [p470]

Berk (1988) discusses causal inferences based on path diagrams, including "unobservable disturbances meeting the usual (and sometimes heroic) assumptions." He considers the oft-recited arguments that biases will be small, or if large will tend to cancel, and concludes: "Unfortunately, it is difficult to find any evidence for these beliefs." He recommends quasi-experimental designs,

> "which are terribly underutilized by sociologists despite their considerable potential. While they are certainly no substitute for random assignment, the stronger quasi-experimental designs can usually produce far more compelling causal inferences than conventional cross-sectional data sets."

He comments on model development by testing, including the use of specification tests:

> "the results may well be misleading if there are <u>any</u> other statistical assumptions that are substantially violated."

I found little to disagree with in Berk's essay. Casual observation suggests that no dramatic change in research practice took place following publication of his essay; further discussion of the issues may be needed.

Of course, Paul Meehl already said most of what needs saying in 1978, in his article, "Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology." In paraphrase, the good knight is Karl Popper, whose motto calls for subjecting scientific theories to grave danger of refutation. The bad knight is Ronald Fisher, whose significance tests are trampled in the dust:

> "the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is... basically unsound...." [p817]

Meehl is an eminent psychologist, and he has one of the best data sets available for demonstrating the predictive power of regression models. His judgment deserves some consideration.

Conclusion

One fairly common way to attack a problem involves collecting data and then making a set of statistical assumptions about the process which generated the data-- for example, linear regression with normal errors, conditional independence of categorical data given covariates, random censoring of observations, independence of competing hazards.

Once the assumptions are in place, the model is fitted to the data, and quite intricate statistical calculations may come into play: three-stage least squares, penalized maximum likelihood, second order efficiency, and so on. The statistical inferences sometimes lead to rather strong empirical claims about structure and causality.

Typically, the assumptions in a statistical model are quite hard to prove or disprove and little effort is spent in that direction. The strength of empirical claims made on the basis of such modeling therefore does not derive from the solidity of the assumptions. Equally, these beliefs cannot be justified by the complexity of the calculations. Success in controlling observable phenomena is a relevant argument, but one that is seldom made.

These observations lead to uncomfortable questions: Are the models helpful? Is it possible to differentiate between successful and unsuccessful uses of the models? How can the models be tested and evaluated? Regression models have been used on social science data since Yule (1899), so it may be time to ask these questions; although definitive answers cannot be expected.

REFERENCES

Arminger, G. and G.W. Bohrnstedt, 1987. "Making it count even more: a review and critique of Stanley Lieberson's Making It Count: The Improvement of Social Theory and Research." Pp. 363-372 in C.C. Clogg (ed.) Sociological Methodology 1987. Washington, D.C.: American Sociological Association.

Bahry, D. and B.D. Silver, 1987. "Intimidation and the symbolic uses of terror in the USSR." American Political Science Review 81:1065-1098.

Berk, R.A., 1988. "Causal inference for sociological data." Pp.155-172 in N.J. Smelser (ed.) Handbook of Sociology. Los Angeles: Sage.

Carpenter, K.J., ed. 1981. Pellagra. Stroudsberg, Penna.: Hutchinson Ross.

Cornfield, J., W. Haenszel, E.C. Hammond, A.M. Lilienfeld, M.B. Shimkin and E.L. Wynder, 1959. "Smoking and lung cancer: recent evidence and a discussion of some questions." Journal of the National Cancer Institute 22:173-203.

Erikson, R.S., J.P. McIver and G.C. Wright, Jr. 1987. "State political culture and public opinion." American Political Science Review 81:797-813

Evans, R.J. 1987. Death in Hamburg: Society and Politics in the Cholera Years, 1830-1910. Oxford: Oxford University Press.

Finlay, B.B., F. Heffron and S. Falkow, 1989. "Epithelial cell surfaces induce Salmonella proteins required for bacterial adherence and invasion." Science 243:940-942.

Fisher, R.A. 1958. Statistical Methods for Research Workers. 13th edition. Edinburgh: Oliver & Boyd.

Fraker, T. and R. Maynard, 1987. "The adequacy of comparison group designs for evaluations of employment-related programs." The Journal of Human Resources 22:194-227.

Freedman, D.A. 1987. "As others see us: a case study in path analysis." Journal of Educational Statistics 12:101-223, with discussion.

Freedman, D.A. and W. Navidi, 1989. "Multistage models for carcinogenesis." Environmental Health Perspectives 81:169-188.

Freedman, D.A. and H. Zeisel, 1988. "Cancer risk assessment: From mouse to man." Statistical Science 3:3-56, with discussion.

Gibson, J.L. 1988. "Political intolerance and political repression during the McCarthy red scare." American Political Science Review 82:511-529.

Heckman, J.J. and V.J. Holtz, 1989. "Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training." Journal of the American Statistical Association 84:862-880, with discussion.

Holland, P. 1988. "Causal inference, path analysis, and recursive structural equations models." Pp.449-484 in C.C. Clogg (ed.) Sociological Methodology 1988. Oxford: Blackwell

Howard-Jones, N. 1975. The Scientific Background of the International Sanitary Conferences 1851-1938. Geneva: World Health Organization.

International Agency for Research on Cancer, 1986. Tobacco Smoking. Monograph 38. Lyon: International Agency for Research on Cancer.

Kanarek, M.S., P.M. Conforti, L.A. Jackson, R.C. Cooper, and J.C. Murchio. 1980. "Asbestos in drinking water and cancer incidence in the San Francisco Bay Area." American Journal of Epidemiology 112:54-72.

LaLonde, R.J. 1986. "Evaluating the econometric evaluations of training programs with experimental data." American Economic Review 76:604-620.

Lieberson, S. 1985. Making It Count: The Improvement of Social Theory and Research. Berkeley: University of California Press.

Lieberson, S. 1988. "Asking too much, expecting too little." Sociological Perspectives 31:379-397

Lombard, H.L. and C.R. Doering, 1928. "Cancer studies in Massachusetts, 2. Habits, characteristics and environment of individuals with and without lung cancer." New England Journal of Medicine 198:481-487.

Pierre Louis, 1835. Recherche sur les effets de la saignee dans quelques maladies inflammatoires: et sur l'action de l'emetique et des vesicatoires dans la pneumonie. Paris: J.B. Bailliere. English edition, 1836. Reprinted in 1986 by Classics of Medicine Library, Birmingham, Alabama.

Marini, M.M. and B. Singer, 1988. "Causality in the social sciences." Pp.347-409 in C.C. Clogg (ed.) Sociological Methodology 1988. Oxford: Blackwell

Meehl, P.E. 1978. "Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology." Journal of Consulting and Clinical Psycholology 46:806-834.

Meehl, P.E. 1954. Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence. Minneapolis: University of Minnesota Press.

Miller, J.F., J.J. Mekalanos and S. Falkow, 1989. "Coordinate regulation and sensory transduction in the control of bacterial virulence." Science 243:916-922.

Mueller, F.H. 1939. "Tabakmissbrauch und Lungcarcinom." Zeitschrift fur Krebsforsuch 49:57-84.

Rosenberg, C.E. 1962. The Cholera Years. Chicago: University of Chicago Press.

Semmelweiss, Ignaz, 1861. Die Aetiologie, der Begriff, und die Prophylaxis des Kindbettfiebers. Pest, Wien u. Leipzig: C.A. Hartleben. Reprinted in 1941 in English. The Etiology, the Concept and the Prophylaxis of Childbed Fever. Medical Classics 5:338-775.

Simon, H. 1957. Models of Man. New York: Wiley.

Singer, B. and M.M. Marini, 1987. "Advancing social research: an essay based on Stanley Lieberson's Making It Count: The Improvement of Social Theory and Research." Pp.373-391 in C.C. Clogg (ed.) Sociological Methodology 1987. Washington, D.C.: American Sociological Association.

Snow, John, 1855. On the Mode of Communication of Cholera. 2nd. ed. London: Churchill. Reprinted in 1965 by Hafner, New York.

Stolzenberg, R.M. and D.A. Relles, 1990. "Theory testing in a world of constrained research design." Sociological Methods and Research 18:395-415.

Terris, M., ed. 1964. Goldberger on Pellagra. Baton Rouge: Louisiana State University Press.

U.S. Public Health Service, 1964. Smoking and Health. Report of the Advisory Committee to the Surgeon General. Washington, D.C.: U.S. Government Printing Office.

Yule, G.U. 1899. "An investigation into the causes of changes in pauperism in England, chiefly during the last two intercensal decades." Journal of the Royal Statistical Society LXII:249-295

Comments on David Freedman's talk
by Ronald G. Bodkin (Economics, UCLA)

In commenting on the very sophisticated paper of David Freedman's, I feel as our Chairman has already indicated; we are in the presence of an advanced level of Darrell Huff's How to Lie With Statistics. Certainly, it is possible to misuse the regression technique, as some of Professor Freedman's horror stories show. But is it inevitable that the technique will be misused? I should doubt that Professor Freedman would really want to go that far, although at times his argument appears to verge on this position. In this regard, I believe that I can point up a volume of work in applied economics (applied econometrics) in which the regression technique has been used appropriately and successfully, in order to get new information that could be extracted from the data. I refer to the volume entitled Readings from Econometrica, which appeared around 1970 and which was edited by John W. Hooper and Marc Nerlove. Admittedly, this is hardly a representative sample of average professional work in the discipline of economics: Econometrica is a leading journal and these papers were accepted because they are the crème de la crème. However, if we are attempting to test the proposition that the regression technique can't be applied correctly in the discipline of economics, any counter-example would appear to be legitimate.

Finally, a brief comment on the methodology that Professor Freedman would appear to favor, namely the (virtually) controlled experiment that the British immunologist [after a week, I forget his name] used so successfully in 1858 to isolate bad water as a leading "cause" of cholera. Of course, we can only applaud his scientific genius in making a great stride toward eliminating this scourge of humankind. However, note that he was also extremely lucky to have found almost all of the other relevant factors "controlled" for him in the London of his day. Most of the time researchers in the various fields of social science will not be so lucky, and they could conceivably waste great amounts of their time if they followed Professor Freedman's advice literally. I conclude that a judicious use of the regression technique (which after all represents an artificial "holding constant" of some other factors that the researcher judges relevant), with a full awareness of its limitations, is still our best alternative, in the overwhelming majority of cases. (Of course, in theoretical econometrics, we developed techniques to counter problems that arise when the classical hypotheses of regression analysis do not hold; but this is a development, not a rejection, of the regression technique.)

Comments on David Freedman's talk
by Mike Intriligator

1. Is the method fair? Selecting articles from journals in the social sciences in the way it was done will not yield the most influential or best such articles.

2. Are the results reasonable? Sure, those who use multiple linear regression sometimes exaggerate their results and are not careful enough in providing caveats, discussing assumptions, etc. They should play devil's advocate with their work, being skeptical about conclusions. But why reject all regression studies in the social sciences? This amounts to throwing out the baby with the bath water.

3. An example of a successful social science multiple linear regression is Waldo Tobler's (UCSB/Geography) archaeology study (which was reported to the Marschak Colloquium) on the site of a lost city, based on frequency of its being mentioned on urns of the period. He used an inverse square law and, based on this regression, found the location of the city.

Comments on David Freedman's talk
by David Rothman

1. Even if they are under one hat, a statistician ought to interact with his scientist client on a regular basis while designing his experiment or analyzing his data, rather than retreating into his academic cell and preparing a document useful more for publication in a statistical journal than for his client. In the work which Kanarek the statistician did for Kanarek the epidemiologist, the latter should have used his common sense to question the effect of asbestos fibers on white males only, since sexually or ethnically specific environmental effects are at the very least rare. This criticism applies also to work in the hard sciences. Statisticians should question the reasonableness of models, and should demand checking of (say) the ln(df) worst data for transcription errors and experimental anomalies in personnel, equipment, and ambient conditions. Even if not rejected, the worst data should be identified in any report giving the final models.

2. We question correctness in applying statistical technique, not resourcefulness in getting information out of a database. A "correct" model may omit certain variables or not test for optimal transformations of variables. These sins of omission lower what may be called the Information Extraction Efficiency of the investigation.

3. Instead of being so fussy about the validity of the application, we ought to consider utility. Since regression and other methods have been very useful in all the sciences,

   we can help more by warning users about the most common pitfalls (e.g. overspecification), even if that looks like cookbookery to the guardians of purity.

4. A great example of the utility of regression applied to the smoking problem is the estimate made over two decades ago that, on the average, seven minutes of life are lost for every seven minute cigarette smoked.

August 4, 1989

They might have found a city

by

D A Freedman
Statistics Department
University of California
Berkeley, Calif 94720

Ronald Bodkin defends regression as a good way to control for confounding variables. He goes on to say "we developed techniques to counter problems that arise when the classical hypotheses of regression analysis do not hold"; presumably, the reference is to two-stage least squares and its analogs. However, confounding has been properly controlled by regression only if the assumptions of the regression model-- "the classical hypotheses"-- hold true. My point is that investigators seldom check. If Bodkin will not defend the assumptions in any particular application, how can he recommend the conclusions?

Two-stage least squares is based on its own statistical hypotheses, and is open to similar questions. For more discussion, see Boruch (1971), Daggett & Freedman (1985), Fraker & Maynard (1987), LaLonde (1986), or Vandenbroucke & Pardoel (1989); also see the Summer, 1987, issue of the Journal of Educational Statistics.

Bodkin cites Hooper & Nerlove (1970) for a collection of papers "in which the regression technique has been used appropriately and successfully, in order to get new information that could be extracted from the data." Most of the papers in that collection seemed to be developing statistical theory, very successfully. A few papers illustrated the new techniques on data, and a few papers used regression for empirical work. However, these seem to be open to the same sort of objections as the examples in my talk. In particular, the stochastic assumptions are usually left implicit, and are quite arbitrary.

There is no doubt that regression can be used to extract new numbers from old ones. The question is about the reliability of the product. The main approaches I can see to answer that question are quite old-fashioned: i) checking the assumptions; ii) independently verifying predictions with new data. Neither gets much play in Hooper & Nerlove.

Coming back to empirical matters, John Snow demonstrated in 1855 that cholera was an infectious disease, the infectious agent being carried through the water supply. Bodkin thinks that Snow was a lucky genius; and therefore, by implication, not a good role model. The claim is worth considering. By 1852, Snow had accumulated large amounts of evidence supporting his thesis, which ran contrary to the received opinions of his time. He wanted even stronger proof. To obtain it, he conceived the idea of a natural experiment; he identified a context in which his theory could be put to a decisive test; and he went to extraordinary trouble collecting data, to take advantage of this grand experiment of nature (hence the "shoe leather" in the title of my talk).

Snow was not interested in mathematical descriptions of hypothetical worlds. He wanted to know what caused real cholera epidemics in 19th-century London, and how to prevent them. He figured it out, long before microbiology came on the scene. (At the time, diseases were thought to be caused by bad air and poisons in the ground.) Snow was a perceptive, resourceful, extremely careful investigator, who showed great respect for empirical facts as opposed to received ideas. That is why his book is still worth reading, a century later.

Mike Intriligator was a charming host, and I am grateful for his hospitality. He and I even agree on what a success story should look like: they fit a model, it tells them where to dig, and they find a city. (The collection cited by Bodkin does not have that sort of texture, because the papers do not make specific, verifiable empirical predictions.) The example Mike seems to have in mind is Tobler & Wineburg (1971). Those investigators fitted the "gravity model" to Bilgic's data on the joint mention of pre-Hittite town names in Assyrian tablets. However, the product of the analysis was a map, not an archeological dig. Tobler & Wineburg don't claim to have discovered an ancient city by modeling. Even more: as they are careful to point out, they didn't test their map against reality, because they didn't have enough information--

> "Our experiment resulted in the configuration shown in Fig. 1, which is based on all the joint mentions and the estimated populations, without constraints to fix the positions of any locations. The fit of this figure to the available data is high (>80%) as is usual for the gravity model, but this is mostly a measure of the internal consistency of the data. The more critical test is to compare our results with known sites. Any solution results in relative coordinates and at least two points must be known in absolute coordinates to determine the scale, and a third point to determine the absolute orientation. For statistical stability many positions should be known in advance. In this case there are sixty-two points to be located, and only Kanis and Hattus (perhaps also Akkua) can be considered known, although reasonable speculations are available concerning the locations of several other sites."

Historical novels are not real history, but history as it might have been. Econometric models often seem to me to have similar epistemological status. And now Mike has invented a new genre-- the success story that might have been. After all, in the end, they didn't dig a hole and they didn't find a city.

"Instead of being so fussy about the validity of the application," writes **David Rothman**, "we ought to consider utility." His success story is "the estimate made over two decades ago that, on the average, seven minutes of life are lost for every seven minute cigarette smoked." There is no doubt regression coefficients pack a lot of rhetorical punch. That is one reason they are worth thinking about. Rothman's estimate is certainly useful, if you want to go after the tobacco companies. And it could have been derived from some model, although he gives no reference. But is it right? even approximately? This is not an issue he addresses.

Smoking is bad for the body, but there is still vigorous dispute about the dose-response relationship: does smoking a pack a day for twenty years have the same impact on the risk of lung cancer as two packs a day for ten years? or is the former more risky by an order of magnitude? The idea that risk depends only on "pack-years" seems quite unlikely, but cannot be totally rejected. We are uncertain about dose response partly because the underlying biological mechanisms are still unclear, partly because relationships differ substantially from one cohort to another, and partly because of difficulties with the data. See Freedman & Navidi (1989), Gaffney & Altshuler (1988), Moolgavkar-Dewanji-Luebeck (1989), Whittemore (1988). The seven-minute estimate is perfect, for a two-minute television interview.

To sum up, regression models are widely used in the social sciences. However, these models make quite strong assumptions about the processes generating the data to which the equations are fitted. Unless the models can be validated, inferences based on them are quite shaky--as propositions about the real world. Current empirical work often seems to draw sweeping conclusions from untested, even unarticulated, assumptions. In this research environment, success stories are hard to find. The present exchange should make this clear, as do previous ones (Journal of Educational Statistics, summer, 1987). It may be time to reconsider.

References

R Boruch (1976). On common contentions about randomized field tests. In Evaluation Studies Review Annual, ed GV Glass, 158-94 Sage, Beverly Hills.

RS Daggett & DA Freedman (1985). Econometrics and the law: a case study in the proof of antitrust damages. In LM LeCam and RA Olshen, eds. Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol I, pp123-72. Wadsworth, Belmont, CA.

RJ Evans (1987). Death in Hamburg: Society & Politics in the Cholera Years, 1830-1910. Oxford University Press.

T Fraker & R Maynard (1987). The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs, Journal of Human Resources 22 194-247.

DA Freedman & W Navidi (1989). Multistage models for carcinogenesis, Environ Health Perspectives 81 169-88

M Gaffney & B Altshuler (1988). Examination of the role of cigarette smoke in lung carcinogenesis using multistage models, J Natl Cancer Inst 80 925-31

J W Hooper & M Nerlove, eds. (1970). Selected Readings in Econometrics. MIT Press.

RJ LaLonde (1986). Evaluating the econometric evaluations of training programs with experimental data, AER 76 604-20

SH Moolgavkar, A Dewanji & G Luebeck (1989). Cigarette smoking and lung cancer: reanalysis of the British Doctors' data, J Natl Cancer Inst 81 415-20.

CE Rosenberg (1962). The Cholera Years. University of Chicago Press.

John Snow (1855). On the Mode of Communication of Cholera. 2nd ed, Churchill, London. Reprinted by Hafner, New York, 1965, in Snow on Cholera

W Tobler & S Wineburg (1971). A Cappadocian speculation, Nature 231 39-41

JP Vandenbroucke & VPAM Pardoel (1989). An autopsy of epidemiologic methods: the case of "poppers" in the early epidemic of the acquired immunodeficiency syndrome (AIDS), Amer J Epidemiol 129 455-7.

A Whittemore (1988). Effect of cigarette smoking in epidemiological studies of lung cancer, Statistics in Medicine 7 223-238.

Comments on the David Freedman Critique of Quantitative Sociology

By

William M. Mason
Sociology Department, University of Michigan
and
Population Studies Center
1225 South University Avenue
Ann Arbor, MI 48104-2590

Presented at
The annual meeting of the American Sociological Association
San Francisco, California
August 12, 1989

I. INTRODUCTION

There is an accumulation of about 10 years of David Freedman's compelling case studies of statistical practice in nonexperimental research, most of it in the social sciences (Freedman 1981; Freedman, Rothenberg, and Sutch 1983; Freedman and Peters 1984a; Freedman and Peters 1984b; Freedman 1985; Daggett and Freedman 1985; Freedman and Navidi 1986; Freedman 1987; Freedman and Zeisel 1988). On a sentence by sentence basis I agree with what David writes the vast majority of the time. This is clearly scholarship and logic of the highest order and integrity, and it deserves our ultimate compliment, which is careful study.

In what follows, I'm not really going to comment on David's work as such. Instead, after I summarize the view that I attribute to David and like minded people, I want to suggest revisions in practice and revisions in our training of sociologists. For good measure, I will also provide a wish list of what I would like from statisticians.

In a set that includes Fred Mosteller, Ed Leamer, Don Rubin, Paul Holland, other excellent statisticians, and of course those whose trade is philosophy, David Freedman is in my view right now the most important philosopher of science to focus on quantitative research in the social sciences. The reason is that David has taken the time to master what we write—in specific instances. His purpose is to assess our implicit framework, the unspoken rules that allow us to think that we are communicating with each other. This is traditionally a philosopher's job. However, I have rarely been satisfied with the results of philosophical discourse on the social sciences, because the philosophers have not convinced me that they understand our trade. David does. He knows the statistics better than we do, he is a quick study when it comes to substance, his range is remarkable, and he carries out his own investigations on our work. His response to it is nonignorable.

Although I read an earlier version of David's comments, and though we talked in advance, I prepared my own remarks without seeing exactly what David presented today, so, here and there, there could be a disjunction between what I am saying and what David has just finished saying. But I suspect this disjunction, if it exists at all, will be slight.

II. THE FREEDMAN POSITION

To begin with, I want to state in my own words the position I believe David advocates. I seem to be able to do this in seven points.

(1) The true experiment is strongly preferred to any other design, if the purpose is to establish causality between X and Y, and strength of relationship.

(2) Absent a true experiment, you should aim for a quasi-experimental design.

(3) Prospective, longitudinal studies can be revealing.

(4) Time-series analysis, whether structural equations modelling, ARIMA, or what have you, is exceedingly difficult to defend.

(5) Cross-sectional designs are perhaps slightly less difficult to defend, but are highly tendentious nevertheless.

(6) Virtually all social science modeling efforts (and here I would include social science experiments, though I'm not sure David would) fail to satisfy reasonable criteria for justification of the stochastic assumptions. Why is Y, conditional on X, Gaussian *iid* or some other distributional form, or why do any other such assumptions hold? Social scientists rarely if ever have theory for the stochastic parts of the specification. In effect, we engage in curve fitting as far as that part of our modeling effort is concerned. Sure, some Y may be normal, or normal under transformation. But so what? We don't have theory that suggests normality, or accounts for normality, or for other functional forms. This is the best case. The worst case doesn't even check to see if normality is satisfied. I am guilty of some of this; most of us are—at least occasionally. Some of us will argue that normality won't matter, so long as we satisfy symmetry and so on. OK, but this means that current practice is indifferent to the stochastic assumptions of our analytic efforts. So there we have it: We tend not to check or satisfy our stochastic assumptions, and even if we do, there is still the problem that we rarely, if ever, have a theoretical rationale for distributional form A, as opposed to forms B and C. I claim this is true for experimental as well as nonexperimental research in the social sciences.

(7) Much reputable work pays inadequate attention to the assumptions and defense of the *deterministic* parts of our models. With regard to structural equation models, this includes the assumptions involving which variables are regressors and which are regressands. This criticism also refers to the nature of the "act" surrounding estimation of a structural equation model. We have *estimated* a "structure." We haven't tested the conception behind that structural concept except, possibly, in the trivial sense that we have used a procedure to restrict some coefficients to zero—either a priori reasoning to identify simultaneity or some kind of test to exclude a variable from a list of regressors. Fundamentally, this enterprise does not distinguish right from wrong—it provides an "account" that may or may not be generalizable or sustain comparisons with other such "accounts" computed for other data. Maybe the whole works is nonlinear. Maybe some kind of threshold model is more appropriate. There is a general failure to try other "accounts" to see if they fit the data better.

These seven points amount to an indictment of *much* current practice: We do relatively *few* experiments. We do *a lot* of analysis of cross-sectional and time-series data. We *rarely* consider the validity of the underlying stochastic assumptions, or their sense in relation to the problem of interest. Our deterministic specifications are *rarely* checked against meaningful alternatives—and here I am not talking about what passes conventionally for sensitivity analysis.

III. MY OWN THOUGHTS

So where do we go from here? A little bit of counterpoint and a prediction:

(1) We can and will do more experiments. However, as Herb Smith (1989) argues in a paper presented at these meetings, as we continue to gain experience with experimentation, some of us will discover that theory is no less necessary. Omitted variables can *interact* with included variables, in which case randomization does not suffice. Further, as Dick Berk and his colleagues have noted (Berk, Lenihan, and Rossi 1980; Rossi, Berk, and Lenihan 1982), uncontrolled intervention can occur *between* experimental

manipulation and experimental outcome. Smith expands upon this point to argue that experimentation don't necessarily yield a single dominant, preferred answer in such a case.

None of this is to say that experimentation is no good, merely to point out that randomization and experimental manipulation don't solve all of our problems, whether these are statistical or philosophical in nature. In any case, we will do more.

(2) I doubt we are going to give up on "observational" data. If anything, we are going to see more of it, and more of it analyzed. Why? For the reasons we know so well already. A lot of what we think about is in the past. It is macro. It is comparative. It is not readily manipulated.

(3) As social scientists, then, we think about lots of problems that we are just not going to do experiments on. And if we do experiments, then there will be problems of external validity. Neither David nor anybody else in relatively free societies seriously argues that sociologists and historians, commentators and journalists, should stop observing, collecting "data," conceptualizing how and why things fit together the way they do, and assessing how well their ideas conform to the reality they attempt to describe. But where do we go from there? We are back to the question of whether our work can cumulate. Like many others before me, I think our work *can* cumulate, in the sense that hypotheses and arguments can be rejected by the marshalling of evidence.

(4) Now, it is perhaps a fair summary to say that, at least since Durkheim and Weber, social scientists have debated whether quantification could be used to assess theories, models, ideas, views. In these debates, the difference between the evidence provided by experiments and nonexperiments that David has concentrated on has played a *minor* role. Much more important have been issues of measurement, conceptualization, comparability, and validity. A major benefit of the debate is that it has sharpened ideas on both "sides," if sides there be. The *attempt* to quantify, even if it is judged inadequate in a given instance, is—or can be—beneficial.

(5) An example of what I am talking about is to be found in Robert Somers's (1971) neglected essay bearing the forbidding title, "Applications of an Expanded Survey Research Model to Comparative Institutional Studies." In this essay, Somers goes to considerable length to "translate" into a quantitative representation. Barrington Moore's masterpiece, *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World.* He does not "test" Moore's thesis. Rather, among other things, he elucidates Moore's argument. This is important. Work of this kind can lead to actual quantification, and to a form of testing. The next steps seem never to be perfect, but they also seem to represent progress.

(6) A related point here is that multivariate, or multivariable, analysis with nonexperimental data may not be able to rule out all *additive* competitors to the argument at hand, but the gain from being able to consider *any* alternative to an argument (even just *one* alternative) is real. This does not always, or perhaps even frequently, happen. But the possibility exists. Moreover, there is emerging work on tests for specification error—in particular, those errors involving the assumption that the omitted variables of the regression are uncorrelated with the included regressors. This line of attack will not solve all the problems of analysis with nonexperimental data, but if it provides any help at all, that is progress.

(7) Everyone here knows that the randomized assignment of a controlled experiment allows us to test a substantive hypothesis about a single effect even if there is an infinity of competing additive hypotheses.

But our universe of discourse is rarely so expansive. Borrowing from Somers's (1971) essay, when Charles Beard argued that it was possible to determine whether support for adoption of the Constitution of the United States came about because of diffuse support for the ideals it embodied, he contrasted that possibility with one other—namely, economic class. Because of his importance to American social thought and history, Beard's work received much scrutiny. His critics have not thought up an infinity of alternative explanations. *One* alternative might be that ethnicity played a role in people's positions. The point here is that alternatives are focussed: If Beard provides us with the usual keys to scrupulous scholarship, and he was thought to do just this, then his readers accept his conclusion *conditionally.* If someone else is able to marshall data for a competing idea, and to demonstrate nonexperimentally that the competing idea "dominates" the prior idea, then we switch allegiance—or we fight back. That is progress. I don't know if it is science, because conceptions of "science" are now quite diverse. I do think that this form of argumentation with data involves a clear element of falsifiability—and that differentiates us from poetry and other humanistic pursuits. It also differentiates us within the field from those who are self-admittedly interested in providing "accounts," yet have relatively implicit or nonexistent rules of evidence.

(8) Now let's turn to assumptions about the stochastic portions of the specifications we use. It is rare that we have theories or knowledge about underlying choice of distribution. We engage in a form of curve fitting, where the curve is rarely seen. As I noted earlier, some would argue that this doesn't much matter. After all, if certain distributional conditions are satisfied, such as symmetry, then OLS or something like it will perform reasonably well. There is an emerging literature on asymptotics, more generally, that provides a kind of escape mechanism to Normality, and so on. There is a big question here: Have statisticians made so much progress that we no longer have to worry so much about the underlying stochastic assumptions? Or, are we simply using inappropriate technical machinery? David would argue the latter. Perhaps he is right. This is a subject that requires directed, formal scrutiny.

When I began graduate school, unit record data processing equipment was still in common use for social science research. That meant that much of what we did involved cross-tabular data, subject to the truly Procrustean Bed of the Hollerith card. In this context, Ed Borgatta was an innovator: He knew how to fake out an IBM collator (which was a device to merge two sets of IBM cards), to obtain a centroid solution factor analysis. That was a stunning technocratic achievement. Computing a regression was a big deal in those days. We didn't worry much about satisfying Normality, and so on, back then. We were thinking about social reality and trying to come up with reasonable quantifications of concepts. A quarter century later, our conceptualizations still do not have much to say in defense of our chosen stochastics; meanwhile statistical technique, and computing, have burgeoned. The microcomputer that sits on my desk probably is faster than the IBM 7094 that occupied a ballroom-sized space and served the entire University of Chicago during much of the 1960s. The software on my PC is enormously powerful. Recognition of this imbalance, this hypertrophy, should be cautionary. Whether it should lead us into alliance with statisticians to develop a new form of quantitative analysis, or lead us to simpler forms of analysis, I just don't know. I do know, however, that I find it harder than ever to read substantive research that is highly statistical. Since some of my own work is of that sort, I have a crisis. Am I the only one?

(9) Much, perhaps most, use of statistical inference in the social sciences is ritualistic and even irrelevant. In many applications, the analysts don't know the universe to which they wish to make inferences, and

don't know how to compute estimates of variability given that they can specify their universe. Moreover, they don't actually *make* inferences for their readers to react to. In addition, in many applications, even if the universe is specified, it is uninteresting. We should get explicit about this. Those asterisks that adorn the tables of our manuscripts are the product of ritual and little if anything more than that. My recommendation is to address this head-on in our writings and in our teachings. Has major progress in the social sciences been made with the aid of the notion of statistical significance, or confidence intervals, or standard errors? Show me if you can. And even if there is some, how much of what we do really depends on that apparatus? Many people agree with what I am saying, but they continue to report standard errors and asterisks—business as usual. In our publications policies, we need to reassess the value of inference. Perhaps we will conclude that we can do with less of it.

(10) I turn next to professional recruitment and socialization. Most professional sociologists come into the field from undergraduate majors in sociology or related fields. Most have little or no mathematics in college—they typically have not had a year of calculus and a semester or quarter of matrix algebra. Nor do they get it as graduate students. As undergraduates, they may have had a course in research methods and an introductory course in statistics. If they are lucky, their instructor will have used Freedman, Pisani, and Purves's (1978) *Statistics*, but they will have forgotten the first author's name, they will probably have been exposed to no more than half to two-thirds of the text, and will have forgotten or never understood most of that.

As graduate students, what will these people experience? The University of Michigan's model is probably fairly intensive and in that sense is a current "best case." Let me state it. Unless students "test out" of the required sequence, they take a year of statistics divided into a semester of introductory topics that we dislike teaching so much (e.g., hypothesis testing) that we try to get to bivariate regression as quickly as possible), followed by a semester on the linear model that somehow manages to squeeze in a section on maximum likelihood estimation of the logistic response model. However, no calculus is used, and maximum likelihood receives no more than a fleeting glance. This does not define the extreme. At least one other program in the United States *also* manages to pack in some work on structural equations modeling.

Additionally, we offer a third semester "topics" course, the substance of which depends on the instructor. It might include survival models or research design, for example.

There is also a substantial commitment to methodology. Students can choose to participate in the Detroit Area Study, which is an annual sample survey of metropolitan Detroit. It is both a teaching survey and a research tool for the principal investigator. It takes three semesters and part of a summer for a student to make it through the DAS. It has been criticized for packing a semester's worth of work into three, but nobody has ever been able to succeed in redesigning this experience—and we know that other Sociology departments look upon it as a model to emulate.

Now, if you don't "like" surveys, and I use the verb advisedly, then you can take a sequence in field work. It takes two semesters and involves a lot of hard work—certainly at Michigan, I think this material has been taught with dedication. Alternatively, if you think of yourself as leaning toward history, you can design your own methods curriculum for learning historiography, as well as work with some truly gifted social historians.

Of course, budding methodologists in sociology don't settle for this curriculum, and they don't have non-mathematical backgrounds. But they are atypical for the field, and their contributions do not make up the bulk of quantitative research.

Instead, we have as students and colleagues doing quantitative work people who say the following kinds of things:

(a) "I have this two-stage least squares *model* that I need you to help me with. . ." This person doesn't understand the difference between model, estimation principle, and computational procedure.

(b) "I want to prove that X causes Y, but not vice versa. . ." This person has a structural equation model in which the simultaneity can not reasonably be identified.

(c) "Why are all these numbers .999?" This person is staring at a page of LISREL output that contains the standardized estimated covariance matrix of the parameter estimates.

I don't want to be thought of as deriding these individuals, or the class of individuals from which they are drawn. They are all trying hard. Besides, I make plenty of mistakes, too.

How do we manage to make fewer mistakes? We can try to modify the curriculum, but this is not going to do the job. You can't pack much more than Michigan does into the graduate curriculum, and the Michigan sociology faculty also tries hard in its own research and in its graduate student mentoring. I have a lot of respect for my colleagues—I can and do learn from all of them. And still we all screw up—we use the incorrect standard errors generated by the usual giant computer packages. We rarely do experiments or discover natural experiments. Our measures are imperfect. We ignore Howard Schuman's (1981) findings on question construction and questionnaire design.

And even if, someday, the computer packages catch up with the statistical profession, and put in all those graphics, and jackknifing, and bootstrapping, and robust estimation, we are still going to have to find time in the curriculum to teach it—and this stuff is hard. Some people will do it well; others will not. A future David Freedman will point this out.

But suppose we all became virtuosic at bootstrapping and graphical analysis of residuals, what then? Our modal design is nonexperimental. Our hope must be that, if we improve the internal logic of our statistical practice, that we will also improve our substantive thinking, and in the process produce more compelling research.

Bengt Muthen (1987) has proposed revisions in the curriculum. Others have and will do so. That is not enough. Sure we want better curriculum. But there is a lot of material out there, and it is pouring out of the statistics departments. Exploratory data analysis á la Tukey, graphics á la Cleveland, the Jackknife, Bootstrapping, frontal attacks on sampling errors for complex sample designs, survival models with or without heterogeneity, and so on. What's best? What do we keep and what do we set aside? We can't all be statisticians, and even if we could, we don't all want to be. What comes next is better cooperation between statisticians and social scientists. And leadership from statisticians, in those spheres where we have a right to expect it.

IV. WHAT I WANT FROM THE STATISTICS PROFESSION

How can this come about? Well, it would help if statisticians got their house in order on the following topics. In mentioning them, I should note that when I ran them past David Freedman, he found them of second-order importance. He's usually right, but he could also be reflecting his own more intimate knowledge of his own field. To a nonstatistician, there is a lot in the world of statisttics that I wish were in better order.

(1) Bayesian vs. non-Bayesian inference: What are we nonstatisticians supposed to do about this fundamental debate? If it is so important for us to be Bayesians, then I want the Bayesians to tell me how to really do Bayesian statistics and give me computer programs that I can use without investing the rest of my life in them. What's out there is quite inadequate. And I'm tired of listening to Bayesians telling me that theirs is the only logically consistent framework.

(2) I'd appreciate recognition from some of our best statisticians that a rejection of and condescension toward so-called "off the shelf statistics" is in fact a mistake. If we have to invent new statistics every time we do substantive research, we are in trouble.

(3) I'd like somebody to tell me how to make meaningful statistical inferences in the social sciences. When do I really have a population? Or what is my superpopulation, and should I care?

(4) In a related vein, please resolve the debate on what to do with sample weights. And when you do, please modify all pertinent computer programs, and please extend the solution beyond regression.

(5) Now here's one that will raise hackles, if the others haven't already: I'd like statisticians to stop propagating bad *substantive research*, as I think they often do when they work on research projects as "hired guns." You especially see this in the biomedical areas. The doctors do the substance, and the statisticians do the statistics with little understanding of the substance. The result is often disappointing. What is needed is a more genuine interaction between subject matter researcher and statistician. And while we are at it, let's straighten out the Applications Section of JASA, which rarely presents strong *substantive* articles.

(6) Now what about textbooks? Freedman, Pisani, and Purves's (1978) introductory text, titled *Statistics*, is as good as they come. In it you will see much that is fully consistent with what David has been telling the profession for the past decade. It can not and does not go far enough with a program for what researchers should *do* with statistics. I can't find any other books that do. It's no problem finding material that extolls the virtues of experiments, but that's not good enough when I'm trying to work with people who are doing historical analyses, or macro comparative analyses.

(7) The statistics profession needs to recognize that there is a division of labor between statisticians and nonstatisticians. It's OK for a Jay Kadane to write his own computer program for an innovative proposal for adjusting the Census for undercount. It's OK for a David Freedman to write his own programs for doing bootstrapping. It's not so OK for sociologists to do this. We usually don't have the skills or knowledge, and we can't be expected to assess the value of innovative statistical techniques. The statistics profession should be pressuring SAS and SPSS and BMDP and Mintab to put in the features they think we need. There is a substantial lag here, and the lag concerns apparently important stuff.

(8) It's really time for the statistics profession to come to terms with its disdain of the social sciences, especially sociology. Even if the level of practice is not high, the subject itself is difficult. The kind of statistics courses that are routinely offered to those who want to actually use statistics are very narrowly focussed. This is also true for the graduate curriculum for those who would be Ph.D. Statisticians. David Freedman is exceptional in his grasp of the regression model in practice, and in his interest in the social sciences. I bet that most statistician teachers of regression don't hold a candle to him. And isn't it a shame that most Ph.D. statisticians really don't know much at all about structural equation estimation with or without latent variables, though they are ready to express a prejudice? It's very hard for sociologists to find statisticians to talk tp about their problems. In sum, if statisticians made more of an effort to find out what we are up to, and *why* we do what we do, maybe we could make a little more progress. I've had my share of experiences in trying to talk to world famous statisticians who just didn't know enough to be useful to me. Statistics has definitely evolved into a field in which people can do their work without actually seeing and doing applications. Again, David is an exception.

(9) How about comparative evaluation of nonnested models? This is an abstruse topic that has not been brought into the public domain, as it were. Surely we need this kind of procedure if we are to evaluate the adequacy of competing hypotheses.

Now I want to draw from my own research for a little concreteness:

I once spent a lot of time trying to do an analysis of tuberculosis mortality (Mason and Smith 1985). My analysis was based on population counts. I used maximum likelihood estimated logistic regression. There's a problem here. If I've got all the data, why do I need a statistical procedure? If I've got a sample, of what do I have a sample, and how do I figure out what the standard errors ought to be? For that matter, how do I figure out what the right estimation procedure ought to be? My answer at the time was that I had to follow what would have been done if I had had a sample in the usual sense. Not totally satisfying. We need statisticians to give us a worthwhile answer for this kind of problem. It comes up all the time. Don't just tell us "watch out," which is a direct quote from Freedman, Pisani, and Purves.

Here is another case: Not too long ago I took what I thought was a careful look at what I considered to be a fundamental question (Kahn and Mason 1987). To wit, do we need to think of the secular trend in political alienation as a cohort phenomenon, which is complicated, or can we think of it just as well as a period phenomenon, which is relatively simple? I worked with pooled cross-sectional sample surveys to test the Easterlin hypothesis that relative cohort size, especially for young adults, drives a lot of phenomena (including political alienation) that, when aggregated, fluctuate over time. In passing, note the contest here: *one* good idea pitted against another.

What kind of estimation is best for this sort of quasi-historical problem? Is this an estimation problem? Does the answer depend on whether we think we are doing history or whether we think we are doing science? If estimation is appropriate, how do we assess variability? Is this an off the shelf problem? Some critics, including a good statistician said that it was, and that I didn't even know which shelf to look on, though I was standing in front of it. Well, maybe. But show me. I continue to think that I had a sample of *one* cycle, not a sample of N (O'Brien and Gwartney-Gibbs 1989; Mason and Kahn 1989).

V. CONCLUSIONS:

I've been all over the ball park. Where do I end up? In a nut shell, I agree with David Freedman's criticisms. I've tried to tell you where I think this leaves us. My original reaction to David's work was to think that my immediate future lay in a retirement home. On reflection, I think "Later, not now—there is too much to do." For this, I thank David. Before I relax at Halcyon Hills Dormitory for the Nearly Dead, I want to work on, and encourage others to work on, our standards of discourse and training. I want to keep trying to do the perfect piece of substantive research. I know I'll fail, and that if somebody doesn't do a job on me, I may end up doing it on myself. But that is of the essence of our craft: What we do is never proven "right," but it can be shown to be wrong.

Permit me two last parting comments. First, if you take Leamer (1983) seriously, you end up wanting to do a variety of different *kinds* of studies on the same topic—a sort of meta-version of the multitrait-multimethod matrix. We already do this to some extent in sociology. But not enough. Instead, we divide into camps—"hard" vs. "soft," "Marxist–non-Marxist," for example, and we don't talk much across camp boundaries. The political economy of departmental life reinforces this posture and it's not healthy. We need at least a partial truce, so we can get those differing kinds of studies of the same subject in greater abundance. In short, this is a plea for greater catholicity.

And finally, you often hear mentioned the need to go back to the "basics" if we are going to make "real" progress (e.g., Berk 1988). I don't think that is going to happen, because the best scholars already think they are focussing on the "basics." We are struggling, and if progress is slow, it is not for lack of trying.

Thank you.

REFERENCES

Berk, Richard A. 1988. "Causal Inference for Sociological Data." Pp. 155–172 in *Handbook of Sociology*, edited by N. J. Smelser. Beverly Hills, CA: Sage Publications.

Berk, Richard A., Kenneth J. Lenihan, and Peter H. Rossi. 1980. "Crime and Poverty: Some Experimental Evidence from Ex-offenders." *American Sociological Review* 54:447–460.

Daggett, R. S. and D. A. Freedman 1985. "Econometrics and the Law: A Case Study in the Proof of Antitrust Damages." Pp. 123–172 in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol. I, edited by Lucien M. Le Cam and Richard A. Olshen. Belmont, CA: Wadsworth.

Freedman, David A., Robert Pisani, and Roger Purves. 1978. *Statistics*. New York: W. W. Norton.

Freedman, David. 1981. "Some Pitfalls in Large Econometric Models: A Case Study." *Journal of Business* 54:479–500.

Freedman, David A. 1985. "Statistics and the Scientific Method." Pp. 345–390 in *Cohort Analysis in Social Research: Beyond the Identification Problem*, edited by W. Mason and S. Fienberg. New York: Springer.

Freedman, David A. 1987. "As Others See Us: A Case Study in Path Analysis." *Journal of Educational Statistics* 12:101–223 (with discussion).

Freedman, David, Thomas Rothenberg, and Richard Sutch. 1983. "On Energy Policy Models." *The Journal of Business & Economic Statistics* 1:24–36 (with discussion).

Freedman, David A. and Stephen C. Peters. 1984a. "Bootstrapping an Econometric Model: Some Empirical Results." *Journal of Business & Economic Statistics* 2:150–158.

Freedman, David A. and Stephen C. Peters. 1984b. "Bootstrapping a Regression Equation: Some Empirical Results." *Journal of the American Statistical Association* 79:97–106.

Freedman, D. A. and W. C. Navidi. 1986. "Regression Models for Adjusting the 1980 Census." *Statistical Science* 1:3–39 (with discussion).

Freedman, D. A. and H. Zeisel. 1988. "From Mouse-to-Man: The Quantitative Assessment of Cancer Risks." *Statistical Science* 3:3–56 (with discussion).

Kahn, Joan and William M. Mason. 1987. "Political Alienation, Cohort Size, and the Easterlin Hypothesis." *American Sociological Review* 52(April):155–169.

Leamer, Edward. 1983. "Taking the Con out of Econometrics." *American Economic Review* 73:31–43.

Mason, William M. and Herbert L. Smith. 1985. "Age-Period-Cohort Analysis and the Study of Deaths from Pulmonary Tuberculosis." Pp. 151–227 in *Cohort Analysis in Social Research: Beyond the Identification Problem*, edited by W. M. Mason and S. E. Fienberg. New York: Springer-Verlag.

Mason, William M. and Joan R. Kahn. 1989. "Political Alienation and Cohort Size Reconsidered: A Reply to O'Brien and Gwartney-Gibbs." *American Sociological Review* 54(June):480–484.

Muthen, Bengt O. 1987. "Response to Freedman's Critique of Path Analysis: Improve Credibility by Better Methodological Training." *Journal of Educational Statistics* 12:178–184.

O'Brien, Robert M. and Patricia A. Gwartney-Gibbs. 1989. "Relative Cohort Size and Political Alienation: Three Methodological Issues and a Replication Supporting the Easterlin Hypothesis." *American Sociological Review* 54(June):476–480.

Rossi, Peter H., Richard A. Berk, and Kenneth J. Lenihan. 1980. "Saying it Wrong with Figures: A Comment on Zeisel." *American Journal of Sociology* 88:390–393.

Schuman, Howard and Stanley Presser. 1981. *Questions and Answers and Attitude Surveys.* New York: Academic Press.

Smith, Herbert L. 1989. "Problems of Specification Common to Experimental and Nonexperimental Social Research." Manuscript (fourth draft, June). Population Studies Center, University of Pennsylvania.

Somers, Robert H. 1971. "Applications of an Expanded Survey Research Model to Comparative Institutional Studies." Pp. 357–420 in *Comparative Methods in Sociology*, edited by Ivan Vallier. Berkeley, CA: University of California Press.