Statistical Models and Shoe Leather
Author(s): David A. Freedman
Source: *Sociological Methodology*, Vol. 21 (1991), pp. 291-313
Published by: American Sociological Association
Stable URL: http://www.jstor.org/stable/270939
Accessed: 20/06/2013 12:13

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Sociological Association is collaborating with JSTOR to digitize, preserve and extend access to *Sociological Methodology*.

http://www.jstor.org

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

10

STATISTICAL MODELS AND SHOE LEATHER

*David A. Freedman**

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 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 nonexperimental data. Failure stories are also discussed, and comparisons may provide some insight. In particular, this paper suggests that statistical technique can seldom be an adequate substitute for good design, relevant data, and testing predictions against reality in a variety of settings.

1. INTRODUCTION

Regression models have been used in the social sciences at least since 1899, when Yule published his paper on changes in "out-relief" as a cause of pauperism: He argued that providing income support outside the poorhouse increased the number of people on relief. At present, regression models are used to make causal argu-

This research was 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. Special thanks go to Peter Marsden.

*University of California, Berkeley

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

ments in a wide variety of social science applications, and it is perhaps time to evaluate the results.

A crude four-point scale may be useful:

1. Regression usually works, although it is (like anything else) imperfect and may sometimes go wrong.
2. Regression sometimes works in the hands of skillful practitioners, but it isn't suitable for routine use.
3. Regression might work, but it hasn't yet.
4. Regression can't work.

Textbooks, courtroom testimony, and newspaper interviews seem to put regression into category 1. Category 4 seems too pessimistic. My own view is bracketed by categories 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 nonexperimental data. (Experimental confirmation is always welcome, although some experiments have problems of their own.) Statistics can play a useful role. With multidimensional 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 do 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 that specifies the variables in the system, their causal interconnections, the functional form of the relationships,

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

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 to decide whether or not these are zero. 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.

2. 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

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

> effect, however, is much less than has been commonly believed. (Louis [1835] 1986, p. 48)

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 [1835] 1986, p. 63)

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 waterborne infectious disease (Snow [1855] 1965). A few years later, Ignaz Semmelweiss discovered how to prevent puerperal fever (Semmelweiss [1861] 1941). 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 (Lombard and Doering 1928; Mueller 1939; Cornfield et al. 1959; U.S. Public Health Service 1964). In epidemiology, careful reasoning on observational data has led to considerable progress. (For failure stories in that subject, 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 and forget that germs cause disease. Microscopes are available but their resolution is poor. Most human pathogens cannot be seen. The isolation of such microorganisms lies 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. (Belief that disease-causing poisons are in the ground comes later.)

Snow was studying cholera, which had arrived in Europe in the early 1800s. Cholera came in epidemic waves, attacked its victims

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

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 that got into the alimentary canal with food or drink, multiplied in the body, and generated some poison that caused the body to expel water. The organism passed out of the body with these evacuations, got back into the water supply, 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 with 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" (p. 3). Snow also found the second case, a man who had taken the room in which Harnold had stayed. More evidence for the infection theory.

Snow found even better evidence in later epidemics. For example, he studied two adjacent apartment buildings, one heavily hit by cholera, the other not. He found that the water supply in the first building was contaminated by runoff from privies and that the water supply in the second building was much cleaner. He also made several "ecological" studies to demonstrate the influence of water supply on the incidence of cholera. In the London of the 1800s, there were many different water companies serving different areas of the city, and some areas were served by more than one company. Several companies took their water from the Thames, which was heavily polluted by sewage. The service areas of such companies had much higher rates of cholera. 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 found that 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

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

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 1853–54 is even more interesting than the fable. For example, there was a large poorhouse in the Broad Street area with few cholera cases. Why? Snow found that the poorhouse had its own well and that 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 water company, however, left its intake point downstream from the sewage discharges. An ecological analysis of the data for the epidemic of 1853–54 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.

> 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

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

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 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. (pp. 74–75)

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

TABLE 1
Snow's Table IX

	Number of Houses	Deaths from Cholera	Deaths Per 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 identified the companies supplying water to the houses of cholera victims in his study area. This gave him the numerators in Table 1. (The denominators were taken from parliamentary records.)

Snow concluded that *if* 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 differences between ecological correlations and individual correlations. And his counterfactual inference is compelling.

As a piece of statistical technology, Table 1 is by no means remarkable. 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 to get 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 1892, there was an epidemic in Hamburg. The leaders of Hamburg rejected Snow's arguments. They followed Max von Pettenkofer, who taught the miasma theory: Contamination of the ground caused cholera. Thus, Hamburg paid little attention to its water supply but spent a great deal of effort digging up and carting

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

away carcasses buried by slaughterhouses. 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 microorganism. 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 that invades the human gut and causes cholera. Today, the molecular biology of cholera is reasonably well understood (Finlay, Heffron, and Falkow 1989; Miller, Mekalanos, and Falkow 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 nonexperimental data. What is impressive is not the statistical technique but the handling of the scientific issues. He made 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 (Snow [1855] 1965, pp. 125–27).

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—which argued 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.

Kanarek et al. examined cancer rates at 35 sites, for blacks and whites, men and women. They controlled for age by standardization and for sex and race by cross-tabulation. But the main tool was loglinear regression, to control for other covariates (marital status, education, income, occupation). Causation was inferred, as usual, if a coefficient was statistically significant after controlling for covariates.

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

Kanarek et al. did not discuss their 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" (1980, p. 62). However, this model of cancer causation is open to serious objections (Freedman and 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 predicts a risk multiplier of only 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. Thus, 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.'s study and Snow's? Kanarek et al. ignored the ecological fallacy. Snow dealt with it. Kanarek et al. tried to control for covariates by modeling, using socioeconomic 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.

3. SOME EXAMPLES FROM THE SOCIAL SCIENCES

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 *American Political Science Review.* These refereed journals accept perhaps 10 percent of their submissions. For analysis, I selected papers that were published in 1987–88, that posed reasonably clear research questions, and that used regression to answer them. I will discuss

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

three of these papers. 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 *American Political Science Review* in 1988. In sum, I believe these papers are quite typical of good current research practice.

Example 1. Bahry and Silver (1987) hypothesized that in Russia, perception of the KGB as efficient deterred political activism. Their 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 that 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, McIver, and Wright (1987) argued that in the U.S., different states really do have different political cultures. After controlling for demographics and geographical region, adding state dummy variables 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 questionnaires from CBS/*New York Times* opinion surveys. With 40 degrees of freedom in the numerator and 55,000 in the denominator, P is spectacular.

On the other hand, the R^2's 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 found that the estimated state effects did not change much. This argument does support the specification, but it is weak.

Example 3. Gibson (1988) asked whether political intolerance during the McCarthy era was driven by mass opinion or elite

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

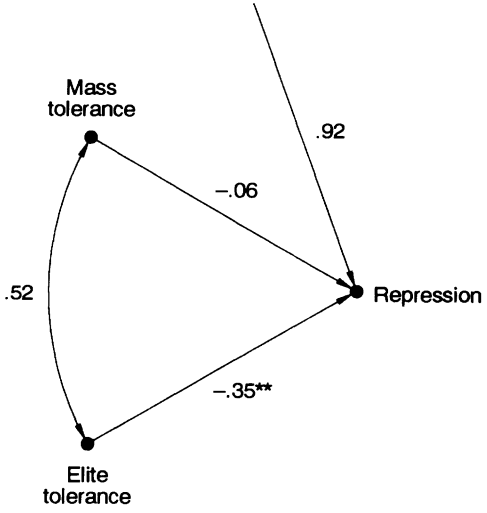

FIGURE 1. Path model of political intolerance. Adapted by permission from Gibson (1988).

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 concluded: "Generally it seems that elites, not masses, were responsible for the repression of the era" (p. 511).

Of the three papers, I thought Gibson's 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 the data.

4. 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 not correlated or heteroscedastic, the conventional formulas will give the wrong answers. And the stochastic specifica-

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

tion 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, each of the examples has these characteristics:

1. There is an interesting research question, which may or may not be sharp enough to be empirically testable.
2. Relevant data are collected, although there may be considerable difficulty in quantifying some of the concepts, and important data may be missing.
3. The research hypothesis is quickly translated into a regression equation, more specifically, into an assertion that certain coefficients are (or are not) statistically significant.
4. Some attention is paid to getting the right variables into the equation, although the choice of covariates is usually not compelling.
5. 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 that, in my opinion, are good of their kind and that 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 that 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 if 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.

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

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, p. 12 [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, p. 190)

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 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 that 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 discovered something extraordinarily important about cholera, and why his book is still worth reading more than a century later.

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

5. 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 that 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 probably distract 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 nonexperimental data. LaLonde (1986) and Fraker and 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 and Maynard conclude:

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

> 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. (p. 194)

Heckman and Hotz (1989, pp. 862, 874) 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 and Hotz have in hand (a) the experimental data, (b) the nonexperimental data, and (c) LaLonde's results as well as Fraker and Maynard's. Heckman and Hotz proceed by modeling the selection bias in the nonexperimental 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 nonexperimental data on female AFDC recipients, and 32 are fitted to the data on high school dropouts. *Ex post facto,* models that 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 and Hotz may have an argument, but it is not a strong one. 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 that has some degree of invariance? Testing 24 models on one data set is less serious.

Generally, replication and prediction of new results provide a

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

harsher and more useful validation regime than statistical testing of many models on one data set. Fewer assumptions are needed, there is less chance of artifact, more kinds of variation can be explored, and 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 over 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.

6. OTHER LITERATURE

The issues raised here are hardly new, and this section 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. I offer 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 and Bohrnstedt (1987, p. 366) respond by describing this as a special case of "misspecification of the mean structure caused by the omission of relevant causal variables" and cite literature on that topic.

This trivializes the problem and almost endorses the idea of fixing misspecification by elaborating the model. However, that 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 semiparametric 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 and Relles 1990).

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

As Arminger and Bohrnstedt concede (1987, p. 370),

> 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 in which the mean function is known, and regression models are correspondingly useful. In the social sciences, I do not see this precondition for regression modeling as being met, even to a first approximation.

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

1. "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." (p. 376)
2. "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." (p. 382)

In my view, point 1 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 2, it is exactly the reliance on models that makes all questions seem "answerable in principle"—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 and 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. (p. 394)

However, much of the analysis in Marini and Singer contradicts this presumed majority view:

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

> 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, 391)

As I read the current empirical research literature, causal arguments depend mainly on the statistical significance of regression coefficients. If so, Marini and Singer are pointing to the fundamental circularity in the regression strategy: The information needed for building regression models comes only from such models. Indeed, Marini and Singer 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. (p. 392)

Holland (1988) raises similar points. Causal inferences from nonexperimental 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 (1988, p. 481).

> The effect . . . cannot be estimated by the usual regression methods of path analysis without making untestable assumptions about the counterfactual regression function. (p. 470)

Berk (1988, p. 161) discusses causal inferences based on path diagrams, including "unobservable disturbances meeting the usual (and sometimes heroic) assumptions." He considers the oft-recited

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

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" (p. 163). 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. (p. 163)

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

> The results may well be misleading if there are *any* other statistical assumptions that are substantially violated. (p. 165)

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 (1978) 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. (p. 817)

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.

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

7. CONCLUSION

One fairly common way to attack a problem involves collecting data and then making a set of statistical assumptions about the process that 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. 363–72 in *Sociological Methodology 1987,* edited by C. C. Clogg. Washington, DC: American Sociological Association.

Bahry, D., and B. D. Silver. 1987. "Intimidation and the Symbolic Uses of Terror in the USSR." *American Political Science Review* 81:1065–98.

Berk, R. A. 1988. "Causal Inference for Sociological Data." Pp. 155–72 in *Handbook of Sociology,* edited by N. J. Smelser. Los Angeles: Sage.

Carpenter, K. J., ed. 1981. *Pellagra.* Stroudsberg, PA: 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

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

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* 243:940–42.

Fisher, R. A. 1958. *Statistical Methods for Research Workers.* 13th ed. Edinburgh: Oliver and Boyd.

Fraker, T., and R. Maynard. 1987. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources* 22:194–227.

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

Freedman, D. A., and W. Navidi. 1989. "Multistage Models for Carcinogenesis." *Environmental Health Perspectives* 81:169–88.

Gibson, J. L. 1988. "Political Intolerance and Political Repression During the McCarthy Red Scare." *American Political Science Review* 82:511–29.

Heckman, J. J., and V. J. Hotz. 1989. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training" (with discussion). *Journal of the American Statistical Association* 84:862–80.

Holland, P. 1988. "Causal Inference, Path Analysis, and Recursive Structural Equations Models." Pp. 449–84 in *Sociological Methodology 1988,* edited by C. C. Clogg. Oxford: Basil Blackwell.

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

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* 76:604–20.

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

———. 1988. "Asking Too Much, Expecting Too Little." *Sociological Perspectives* 31:379–97.

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* 198:481–87.

Louis, Pierre. (1835) 1986. *Researches on the Effects of Bloodletting in Some Inflammatory Diseases, and the Influence of Emetics and Vesication in Pneumonitis.* Translated and reprinted. Birmingham, AL: Classics of Medicine Library.

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions

Marini, M. M., and B. Singer. 1988. "Causality in the Social Sciences." Pp. 347–409 in *Sociological Methodology 1988,* edited by C. C. Clogg. Oxford: Basil 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 Psychology* 46:806–34.

Miller, J. F., J. J. Mekalanos, and S. Falkow. 1989. "Coordinate Regulation and Sensory Transduction in the Control of Bacterial Virulence." *Science* 243: 916–22.

Mueller, F. H. 1939. "Tabakmissbrauch und Lungcarcinom" (Tobacco abuse and lung cancer). *Zeitschrift fur Krebsforsuch* 49:57–84.

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

Semmelweiss, Ignaz. (1861) 1941. "The Etiology, the Concept and the Prophylaxis of Childbed Fever." Translated and reprinted. *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. 373–91 in *Sociological Methodology 1987,* edited by C. C. Clogg. Washington, DC: American Sociological Association.

Snow, John. (1855) 1965. *On the Mode of Communication of Cholera.* Reprint ed. New York: Hafner.

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, DC: 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* 62:249–95.

This content downloaded from 139.78.49.186 on Thu, 20 Jun 2013 12:13:05 PM
All use subject to JSTOR Terms and Conditions