Let's Take the Con Out of Econometrics

Edward E. Leamer

The American Economic Review, Volume 73, Issue 1 (Mar., 1983), 31-43.

Stable URL:
http://links.jstor.org/sici?sici=0002-8282%28198303%2973%3A1%3C31%3ALTTCOO%3E2.0.CO%3B2-R

Let's Take the Con out of Econometrics

By EDWARD E. LEAMER*

Econometricians would like to project the image of agricultural experimenters who divide a farm into a set of smaller plots of land and who select randomly the level of fertilizer to be used on each plot. If some plots are assigned a certain amount of fertilizer while others are assigned none, then the difference between the mean yield of the fertilized plots and the mean yield of the unfertilized plots is a measure of the effect of fertilizer on agricultural yields. The econometrician's humble job is only to determine if that difference is large enough to suggest a real effect of fertilizer, or is so small that it is more likely due to random variation.

This image of the applied econometrician's art is grossly misleading. I would like to suggest a more accurate one. The applied econometrician is like a farmer who notices that the yield is somewhat higher under trees where birds roost, and he uses this as evidence that bird droppings increase yields. However, when he presents this finding at the annual meeting of the American Ecological Association, another farmer in the audience objects that he used the same data but came up with the conclusion that moderate amounts of shade increase yields. A bright chap in the back of the room then observes that these two hypotheses are indistinguishable, given the available data. He mentions the phrase "identification problem," which, though no one knows quite what he means, is said with such authority that it is totally convincing. The meeting reconvenes in the halls and in the bars, with heated discussion whether this is the kind of work that merits promotion from Associate to Full Farmer; the Luminists strongly opposed to promotion and the Aviophiles equally strong in favor.

*Professor of economics, University of California-Los Angeles. This paper was a public lecture presented at the University of Toronto, January 1982. I acknowledge partial support by NSF grant SOC78-09479.

One should not jump to the conclusion that there is necessarily a substantive difference between drawing inferences from experimental as opposed to nonexperimental data. The images I have drawn are deliberately prejudicial. First, we had the experimental scientist with hair neatly combed, wide eyes peering out of horn-rimmed glasses, a white coat, and an electronic calculator for generating the random assignment of fertilizer treatment to plots of land. This seems to contrast sharply with the nonexperimental farmer with overalls, unkempt hair, and bird droppings on his boots. Another image, drawn by Orcutt, is even more damaging: "Doing econometrics is like trying to learn the laws of electricity by playing the radio." However, we need not now submit to the tyranny of images, as many of us have in the past.

I. Is Randomization Essential?

What is the real difference between these two settings? Randomization seems to be the answer. In the experimental setting, the fertilizer treatment is "randomly" assigned to plots of land, whereas in the other case nature did the assignment. Now it is the tyranny of words that we must resist. "Random" does not mean adequately mixed in *every* sample. It only means that on the average, the fertilizer treatments are adequately mixed. Randomization implies that the least squares estimator is "unbiased," but that definitely does not mean that for each sample the estimate is correct. Sometimes the estimate is too high, sometimes too low. I am reminded of the lawyer who remarked that "when I was a young man I lost many cases that I should have won, but when I grew older I won many that I should have lost, so on the average justice was done."

In particular, it is possible for the randomized assignment to lead to exactly the same allocation as the nonrandom assignment,

namely, with treated plots of land all being under trees and with nontreated plots of land all being away from trees. I submit that, if this is the outcome of the randomization, then the randomized experiment and the nonrandomized experiment are exactly the same. Many econometricians would insist that there is a difference, because the randomized experiment generates "unbiased" estimates. But all this means is that, if this particular experiment yields a gross overestimate, some other experiment yields a gross underestimate.

Randomization thus does not assure that each and every experiment is "adequately mixed," but randomization does make "adequate mixing" probable. In order to make clear what I believe to be the true value of randomization, let me refer to the model

$$(1) \qquad Y_i = \alpha + \beta F_i + \gamma L_i + U_i,$$

where Y_i is the yield of plot i; F_i is the fertilizer assigned to plot i; L_i is the light falling on plot i; U_i is the unspecified influence on the yield of plot i, and where β, the fertilizer effect, is the object of the inferential exercise. We may suppose to begin the argument that the light level is expensive to measure and that it is decided to base an estimate of β initially only on measurement of Y_i and F_i. We may assume also that the natural experiment produces values for F_i, L_i, and U_i with expected values $E(U_i|F_i) = 0$ and $E(L_i|F_i) = r_0 + r_1 F_i$. In the more familiar parlance, it is assumed that the fertilizer level and the residual effects are uncorrelated, but the fertilizer level and the light level are possibly correlated. As every beginning econometrics student knows, if you omit from a model a variable which is correlated with included variables, bad things happen. These bad things are revealed to the econometrician by computing the conditional mean of Y given F but not L:

$$(2) \quad \begin{aligned} E(Y|F) &= \alpha + \beta F + \gamma E(L|F) \\ &= \alpha + \beta F + \gamma(r_0 + r_1 F) \\ &\equiv (\alpha + \alpha^*) + (\beta + \beta^*) F, \end{aligned}$$

where $\alpha^* = \gamma r_0$ and $\beta^* = \gamma r_1$. The linear regression of Y on F provides estimates of the parameters of the conditional distribution of Y given F, and in this case the regression coefficients are estimates not of α and β, but rather of $\alpha + \alpha^*$ and $\beta + \beta^*$. The parameters α^* and β^* measure the bias in the least squares estimates. This bias could be due to left-out variables, or to measurement errors in F, or to simultaneity.

When observing a nonexperiment, the bias parameters α^* and β^* can be thought to be small, but they cannot sensibly be treated as exact zeroes. The notion that the bias parameters are small can be captured by the assumption that α^* and β^* are drawn from a normal distribution with zero means and covariance matrix M. The model can then be written as $Y = \alpha + \beta F + \varepsilon$, where ε is the sum of three random variables: $U + \alpha^* + \beta^* F$. Because the error term ε is not spherical, the proper way to estimate α and β is generalized least squares. My 1974 article demonstrates that if (a, b) represent the least squares estimates of (α, β), then the generalized least squares estimates $(\hat{\alpha}, \hat{\beta})$ are also equal to (a, b):

$$(3) \qquad \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix},$$

and if S represents the sample covariance matrix for the least squares estimates, then the sample covariance matrix for $(\hat{\alpha}, \hat{\beta})$ is

$$(4) \qquad Var(\hat{\alpha}, \hat{\beta}) = S + M,$$

where M is the covariance matrix of (α^*, β^*).

The meaning of equation (3) is that unless one knows the direction of the bias, the possibility of bias does not call for any adjustment to the estimates. The possibility of bias does require an adjustment to the covariance matrix (4). The uncertainty is composed of two parts: the usual sampling uncertainty S plus the misspecification uncertainty M. As sample size grows, the sampling uncertainty S ever decreases, but the misspecification uncertainty M remains ever constant. The misspecification matrix M that we must add to the least squares variance

matrix is just the (prior) variance of the bias coefficients (α^*, β^*). If this variance matrix is small, the least squares bias is likely to be small. If M is large, it is correspondingly probable that (α^*, β^*) is large.

It would be a remarkable bootstrap if we could determine the extent of the misspecification from the data. The data in fact contain no information about the size of the bias, a point which is revealed by studying the likelihood function. The misspecification matrix M is therefore a pure prior concept. One must decide independent of the data how good the nonexperiment is.

The formal difference between a randomized experiment and a natural experiment is measured by the matrix M. If the treatment is randomized, the bias parameters (α^*, β^*) are exactly zero, or, equivalently, the matrix M is a zero matrix. If M is zero, the least squares estimates are consistent. If M is not zero, as in the natural experiment, there remains a fixed amount of specification uncertainty, independent of sample size.

There is therefore a sharp difference between inference from randomized experiments and inference from natural experiments. This seems to draw a sharp distinction between economics where randomized experiments are rare and "science" where experiments are routinely done. But the fact of the matter is that no one has ever designed an experiment that is free of bias, and no one can. As it turns out, the technician who was assigning fertilizer levels to plots of land, took his calculator into the fields, and when he was out in the sun, the calculator got heated up and generated large "random" numbers, which the technician took to mean no fertilizer; and when he stood under the shade of the trees, his cool calculator produced small numbers, and these plots received fertilizer.

You may object that this story is rather fanciful, but I need only make you think it is possible, to force you to set $M \neq 0$. Or if you think a computer can really produce random numbers (calculated by a mathematical formula and therefore perfectly predictable!), I will bring up mismeasurement of the fertilizer level, or human error in carrying out the computer instructions. Thus, the attempt to randomize and the attempt to measure accurately ensures that M is small, but not zero, and the difference between scientific experiments and natural experiments is difference in degree, but not in kind. Admittedly however, the misspecification uncertainty in many experimental settings may be so small that it is well approximated by zero. This can very rarely be said in nonexperimental settings.

Examples may be ultimately convincing. There is a great deal of empirical knowledge in the science of astronomy, yet there are no experiments. Medical knowledge is another good example. I was struck by a headline in the January 5, 1982 *New York Times*: "Life Saving Benefits of Low-Cholesterol Diet Affirmed in *Rigorous* Study." The article describes a randomized experiment with a control group and a treated group. "Rigorous" is therefore interpreted as "randomized." As a matter of fact, there was a great deal of evidence suggesting a link between heart disease and diet before any experiments were performed on humans. There were cross-cultural comparisons and there were animal studies. Actually, the only reason for performing the randomized experiment was that someone believed there was pretty clear nonexperimental evidence to begin with. The nonexperimental evidence was, of course, inconclusive, which in my language means that the misspecification uncertainty M remained uncomfortably large. The fact that the Japanese have both less incidence of heart disease and also diets lower in cholesterol compared to Americans is not convincing evidence, because there are so many other factors that remain unaccounted for. The fact that pigs on a high cholesterol diet develop occluded arteries is also not convincing, because the similarity in physiology in pigs and humans can be questioned.

When the sampling uncertainty S gets small compared to the misspecification uncertainty M, it is time to look for other forms of evidence, experiments or nonexperiments. Suppose I am interested in measuring the width of a coin, and I provide rulers to a room of volunteers. After each volunteer has reported a measurement, I compute the mean and standard deviation, and I conclude that

the coin has width 1.325 millimeters with a standard error of .013. Since this amount of uncertainty is not to my liking, I propose to find three other rooms full of volunteers, thereby multiplying the sample size by four, and dividing the standard error in half. That is a silly way to get a more accurate measurement, because I have already reached the point where the sampling uncertainty S is very small compared with the misspecification uncertainty M. If I want to increase the true accuracy of my estimate, it is time for me to consider using a micrometer. So too in the case of diet and heart disease. Medical researchers had more or less exhausted the vein of nonexperimental evidence, and it became time to switch to the more expensive but richer vein of experimental evidence.

In economics, too, we are switching to experimental evidence. There are the laboratory experiments of Charles Plott and Vernon Smith (1978) and Smith (1980), and there are the field experiments such as the Seattle/Denver income maintenance experiment. Another way to limit the misspecification error M is to gather different kinds of nonexperiments. Formally speaking, we will say that experiment 1 is qualitatively different from experiment 2 if the bias parameters (α_1^*, β_1^*) are distributed independently of the bias parameters (α_2^*, β_2^*). In that event, simple averaging of the data from the two experiments yields average bias parameters $(\alpha_1^* + \alpha_2^*,\ \beta_1^* + \beta_2^*)/2$ with misspecification variance matrix $M/2$, half as large as the (common) individual variances. Milton Friedman's study of the permanent income hypothesis is the best example of this that I know. Other examples are hard to come by. I believe we need to put much more effort into identifying qualitatively different and convincing kinds of evidence.

Parenthetically, I note that traditional econometric theory, which does not admit experimental bias, as a consequence also admits no "hard core" propositions. Demand curves can be shown to be positively sloped. Utility can be shown not to be maximized. Econometric evidence of a positively sloped demand curve would, as a matter of fact, be routinely explained in terms of simultaneity bias. If utility seems not to have been maximized, it is only that the econometrician has misspecified the utility function. The misspecification matrix M thus forms Imre Lakatos' "protective belt" which protects certain hard core propositions from falsification.

II. Is Control Essential?

The experimental scientist who notices that the fertilizer treatment is correlated with the light level can correct his experimental design. He can control the light level, or he can allocate the fertilizer treatment in such a way that the fertilizer level and the light level are not perfectly correlated.

The nonexperimental scientist by definition cannot control the levels of extraneous influences such as light. But he can control for the variable light level by including light in the estimating equation. Provided nature does not select values for light and values for fertilizer levels that are perfectly correlated, the effect of fertilizer on yields can be estimated with a multiple regression. The collinearity in naturally selected treatment variables may mean that the data evidence is weak, but it does not invalidate in any way the usual least squares estimates. Here, again, there is no essential difference between experimental and nonexperimental inference.

III. Are the Degrees of Freedom Inadequate with Nonexperimental Data?

As a substitute for experimental control, the nonexperimental researcher is obligated to include in the regression equation all variables that might have an important effect. The NBER data banks contain time-series data on 2,000 macroeconomic variables. A model explaining gross national product in terms of all these variables would face a severe degrees-of-freedom deficit since the number of annual observations is less than thirty. Though the number of observations of any phenomenon is clearly limited, the number of explanatory variables is logically unlimited. If a polynomial could have a degree as high as k, it would usually be admitted that the degree could be $k+1$ as well. A theory that allows k lagged explanatory vari-

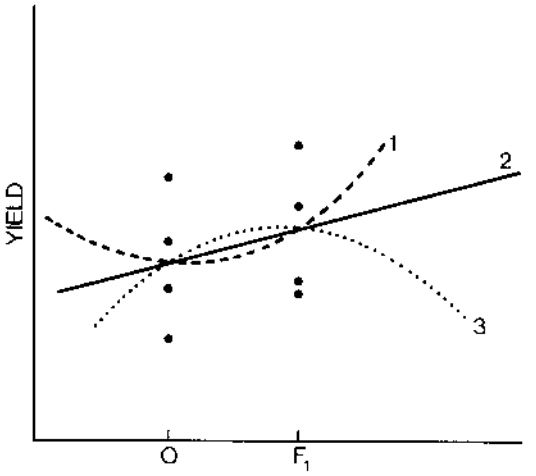

FIGURE 1. HYPOTHETICAL DATA AND THREE ESTIMATED QUADRATIC FUNCTIONS

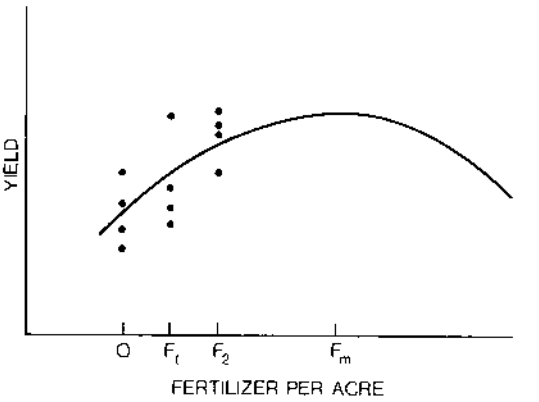

FIGURE 2. HYPOTHETICAL DATA AND ESTIMATED QUADRATIC FUNCTION

ables would ordinarily allow $k+1$. If the level of money might affect *GNP*, then why not the number of presidential sneezes, or the size of the polar ice cap?

The number of explanatory variables is unlimited in a nonexperimental setting, but it is also unlimited in an experimental setting. Consider again the fertilizer example in which the farmer randomly decides either to apply F_1 pounds of fertilizer per acre or zero pounds, and obtains the data illustrated in Figure 1. These data admit the inference that fertilizer level F_1 produces higher yields than no fertilizer. But the farmer is interested in selecting the fertilizer level that maximizes profits. If it is hypothesized that yield is a linear function of the fertilizer intensity $Y = \alpha + \beta F + U$, then profits are

$$Profits = pA(\alpha + \beta F + U) - p_F AF,$$

where A is total acreage, p is the product price, and p_F is the price per pound of fertilizer. This profit function is linear in F with slope $A(\beta p - p_F)$. The farmer maximizes profits therefore by using no fertilizer if the price of fertilizer is high, $\beta p < p_F$, and using an unlimited amount of fertilizer if the price is low, $\beta p > p_F$. It is to be expected that you will find this answer unacceptable for one of several reasons:

1) When the farmer tries to buy an unlimited amount of fertilizer, he will drive up its price, and the problem should be reformulated to make p_F a function of F.

2) Uncertainty in the fertilizer effect β causes uncertainty in profits, $Variance(profits) = p^2A^2F^2Var(\beta)$, and risk aversion will limit the level of fertilizer applied.

3) The yield function is nonlinear.

Economic theorists doubtless find reasons 1) and 2) compelling, but I suspect that the real reason farmers don't use huge amounts of fertilizer is that the marginal increase in the yield eventually decreases. Plants don't grow in fertilizer alone.

So let us suppose that yield is a quadratic function of fertilizer intensity, $Y = \alpha + \beta_1 F + \beta_2 F^2 + U$, and suppose we have only the data illustrated in Figure 1. Unfortunately, there are an infinite number of quadratic functions all of which fit the data equally well, three of which are drawn. If there were no other information available, we could conclude only that the yield is higher at F_1 than at zero. Formally speaking, there is an identification problem, which can be solved by altering the experimental design. The yield must be observed at a third point, as in Figure 2, where I have drawn the least squares estimated quadratic function and have indicated the fertilizer intensity F_m that maximizes the yield. I expect that most people would question whether these data admit the

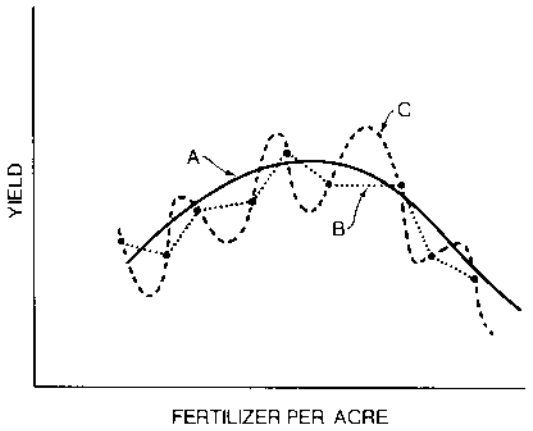

FIGURE 3. HYPOTHETICAL DATA AND THREE ESTIMATED FUNCTIONS

inference that the yield is maximized at F_m. Actually, after inspection of this figure, I don't think anything can be inferred except that the yield at F_2 is higher than at F_1, which in turn is higher than at zero. Thus I don't believe the function is quadratic. If it is allowed to be a cubic then again there is an identification problem.

This kind of logic can be extended indefinitely. One can always find a set of observations that will make the inferences implied by a polynomial of degree p seem silly. This is true regardless of the degree p. Thus no model with a finite number of parameters is actually believed, whether the data are experimental or nonexperimental.

IV. Do We Need Prior Information?

A model with an infinite number of parameters will allow inference from a finite data set only if there is some prior information that effectively constrains the ranges of the parameters. Figure 3 depicts another hypothetical sequence of observations and three estimated relationships between yield and fertilizer. I believe the solid line A is a better representation of the relationship than either of the other two. The piecewise linear form B fits the data better, but I think this peculiar meandering function is highly unlikely on an a priori basis. Though B and C fit the data equally well, I believe that B is much more likely than C. What I am revealing is the a priori opinion that the function is likely to be smooth and single peaked.

What should now be clear is that data alone cannot reveal the relationship between yield and fertilizer intensity. Data can reveal the yield at sampled values of fertilizer intensities, but in order to interpolate between these sampled values, we must resort to subjective prior information.

Economists have inherited from the physical sciences the myth that scientific inference is objective, and free of personal prejudice. This is utter nonsense. All knowledge is human belief; more accurately, human opinion. What often happens in the physical sciences is that there is a high degree of conformity of opinion. When this occurs, the opinion held by most is asserted to be an objective fact, and those who doubt it are labelled "nuts." But history is replete with examples of opinions losing majority status, with once-objective "truths" shrinking into the dark corners of social intercourse. To give a trivial example, coming now from California I am unsure whether fat ties or thin ties are aesthetically more pleasing.

The false idol of objectivity has done great damage to economic science. Theoretical econometricians have interpreted scientific objectivity to mean that an economist must identify exactly the variables in the model, the functional form, and the distribution of the errors. Given these assumptions, and given a data set, the econometric method produces an objective inference from a data set, unencumbered by the subjective opinions of the researcher.

This advice could be treated as ludicrous, except that it fills all the econometric textbooks. Fortunately, it is ignored by applied econometricians. The econometric art as it is practiced at the computer terminal involves fitting many, perhaps thousands, of statistical models. One or several that the researcher finds pleasing are selected for reporting purposes. This searching for a model is often well intentioned, but there can be no doubt that such a specification search invalidates the traditional theories of inference. The concepts of unbiasedness, consistency, efficiency, maximum-likelihood estimation,

in fact, all the concepts of traditional theory, utterly lose their meaning by the time an applied researcher pulls from the bramble of computer output the one thorn of a model he likes best, the one he chooses to portray as a rose. The consuming public is hardly fooled by this chicanery. The econometrician's shabby art is humorously and disparagingly labelled "data mining," "fishing," "grubbing," "number crunching." A joke evokes the Inquisition: "If you torture the data long enough, Nature will confess" (Coase). Another suggests methodological fickleness: "Econometricians, like artists, tend to fall in love with their models" (wag unknown). Or how about: "There are two things you are better off not watching in the making: sausages and econometric estimates."

This is a sad and decidedly unscientific state of affairs we find ourselves in. Hardly anyone takes data analyses seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analyses seriously. Like elaborately plumed birds who have long since lost the ability to procreate but not the desire, we preen and strut and display our t-values.

If we want to make progress, the first step we must take is to discard the counterproductive goal of objective inference. The dictionary defines an inference as a logical conclusion based on a set of facts. The "facts" used for statistical inference about θ are first the data, symbolized by x, second a conditional probability density, known as a sampling distribution, $f(x|\theta)$, and, third, explicitly for a Bayesian and implicitly for "all others," a marginal or prior probability density function $f(\theta)$. Because both the sampling distribution and the prior distribution are actually *opinions* and not *facts*, a statistical inference is and must forever remain an *opinion*.

What is a fact? A fact is merely an opinion held by all, or at least held by a set of people you regard to be a close approximation to all.[1] For some that set includes only one person. I myself have the opinion that Andrew Jackson was the sixteenth president of the United States. If many of my friends agree, I may take it to be a fact. Actually, I am most likely to regard it to be a fact if the authors of one or more books say it is so.

The difference between a fact and an opinion for purposes of decision making and inference is that when I use opinions, I get uncomfortable. I am not too uncomfortable with the opinion that error terms are normally distributed because most econometricians make use of that assumption. This observation has deluded me into thinking that the opinion that error terms are normal may be a fact, when I know deep inside that normal distributions are actually used only for convenience. In contrast, I am *quite* uncomfortable using a prior distribution, mostly I suspect because hardly anyone uses them. If convenient prior distributions were used as often as convenient sampling distributions, I suspect that I could be as easily deluded into thinking that prior distributions are facts as I have been into thinking that sampling distributions are facts.

To emphasize this hierarchy of statements, I display them in order: truths; facts; opinions; conventions. Note that I have added to the top of the order, the category truths. This will appeal to those of you who feel compelled to believe in such things. At the bottom are conventions. In practice, it may be difficult to distinguish a fact from a convention, but when facts are clearly unavailable, we must strongly resist the deceit or delusion that conventions can represent.

What troubles me about using opinions is their whimsical nature. Some mornings when I arise, I have the opinion that Raisin Bran is better than eggs. By the time I get to the kitchen, I may well decide on eggs, or oatmeal. I usually do recall that the sixteenth president distinguished himself. Sometimes I think he was Jackson; often I think he was Lincoln.

A data analysis is similar. Sometimes I take the error terms to be correlated, sometimes uncorrelated; sometimes normal and sometimes nonnormal; sometimes I include observations from the decade of the fifties, sometimes I exclude them; sometimes the

[1] This notion of "truth by consensus" is espoused by Thomas Kuhn (1962) and Michael Polanyi (1964). Oscar Wilde agrees by dissent: "A truth ceases to be true when more than one person believes it."

equation is linear and sometimes nonlinear; sometimes I control for variable z, sometimes I don't. Does it depend on what I had for breakfast?

As I see it, the fundamental problem facing econometrics is how adequately to control the whimsical character of inference, how sensibly to base inferences on opinions when facts are unavailable. At least a partial solution to this problem has already been formed by practicing econometricians. A common reporting style is to record the inferences implied by alternative sets of opinions. It is not unusual to find tables that show how an inference changes as variables are added to or deleted from the equation. This kind of sensitivity analysis reports special features of the mapping from the space of assumptions to the space of inferences. The defect of this style is that the coverage of assumptions is infinitesimal, in fact a zero volume set in the space of assumptions. What is needed instead is a more complete, but still economical way to report the mapping of assumptions into inferences. What I propose to do is to develop a correspondence between regions in the assumption space and regions in the inference space. I will report that all assumptions in a certain set lead to essentially the same inference. Or I will report that there are assumptions within the set under consideration that lead to radically different inferences. In the latter case, I will suspend inference and decision, or I will work harder to narrow the set of assumptions.

Thus what I am asserting is that the choice of a particular sampling distribution, or a particular prior distribution, is inherently whimsical. But statements such as "The sampling distribution is symmetric and unimodal" and "My prior is located at the origin" are not necessarily whimsical, and in certain circumstances do not make me uncomfortable.

To put this somewhat differently, an inference is not believable if it is fragile, if it can be reversed by minor changes in assumptions. As consumers of research, we correctly reserve judgment on an inference until it stands up to a study of fragility, usually by other researchers advocating opposite opinions. It is, however, much more efficient for individual researchers to perform their own sensitivity analyses, and we ought to be demanding much more complete and more honest reporting of the fragility of claimed inferences.

The job of a researcher is then to report economically and informatively the mapping from assumptions into inferences. In a slogan, "The mapping is the message." The mapping does not depend on opinions (assumptions), but reporting the mapping economically and informatively does. A researcher has to decide which assumptions or which sets of alternative assumptions are worth reporting. A researcher is therefore forced either to anticipate the opinions of his consuming public, or to recommend his own opinions. It is actually a good idea to do both, and a serious defect of current practice is that it concentrates excessively on convincing one's self and, as a consequence, fails to convince the general professional audience.

The whimsical character of econometric inference has been partially controlled in the past by an incomplete sensitivity analysis. It has also been controlled by the use of conventions. The normal distribution is now so common that there is nothing at all whimsical in its use. In some areas of study, the list of variables is partially conventional, often based on whatever list the first researcher happened to select. Even conventional prior distributions have been proposed and are used with nonnegligible frequency. I am referring to Robert Shiller's (1973) smoothness prior for distributed lag analysis and to Arthur Hoerl and Robert Kennard's (1970) ridge regression prior. It used to aggravate me that these methods seem to find public favor whereas overt and complete Bayesian methods such as my own proposals (1972) for distributed lag priors are generally ignored. However, there is a very good reason for this: the attempt to form a prior distribution from scratch involves an untold number of partly arbitrary decisions. The public is rightfully resistant to the whimsical inferences which result, but at the same time is receptive to the use of priors in ways that control the whimsy. Though the use of conventions does control the whimsy, it can do so at the cost of relevance. Inferences based

on Hoerl and Kennard's conventional "ridge regression" prior are usually irrelevant, because it is rarely sensible to take the prior to be spherical and located at the origin, and because a closer approximation to prior belief can be suspected to lead to substantially different inferences. In contrast, the conventional assumption of normality at least uses a distribution which usually cannot be ruled out altogether. Still, we may properly demand a demonstration that the inferences are insensitive to this distributional assumption.

A. *The Horizon Problem: Sherlock Holmes Inference*

Conventions are not to be ruled out altogether, however. One can go mad trying to report completely the mapping from assumptions into inferences since the space of assumptions is infinite dimensional. A formal statistical analysis therefore has to be done within the limits of a reasonable horizon. An informed convention can usefully limit this horizon. If it turned out that sensible neighborhoods of distributions around the normal distribution 99 times out of 100 produced the same inference, then we could all agree that there are other more important things to worry about, and we may properly adopt the convention of normality. The consistency of least squares estimates under wide sets of assumptions is used improperly as support for this convention, since the inferences from a given finite sample may nonetheless be quite sensitive to the normality assumption.[2]

The truly sharp distinction between inference from experimental and inference from nonexperimental data is that experimental inference sensibly admits a conventional horizon in a critical dimension, namely the choice of explanatory variables. If fertilizer is randomly assigned to plots of land, it is conventional to restrict attention to the relationship between yield and fertilizer, and to proceed as if the model were perfectly specified, which in my notation means that the misspecification matrix M is the zero matrix. There is only a small risk that when you present your findings, someone will object that fertilizer and light level are correlated, and there is an even smaller risk that the conventional zero value for M will lead to inappropriate inferences. In contrast, it would be foolhardy to adopt such a limited horizon with nonexperimental data. But if you decide to include light level in your horizon, then why not rainfall; and if rainfall, then why not temperature; and if temperature, then why not soil depth, and if soil depth, then why not the soil grade; ad infinitum. Though this list is never ending, it can be made so long that a nonexperimental researcher can feel as comfortable as an experimental researcher that the risk of having his findings upset by an extension of the horizon is very low. The exact point where the list is terminated must be whimsical, but the inferences can be expected not to be sensitive to the termination point if the horizon is wide enough.

Still, the horizon within which we all do our statistical analyses has to be ultimately troublesome, since there is no formal way to know what inferential monsters lurk beyond our immediate field of vision. "Diagnostic" tests with explicit alternative hypotheses such as the Durbin-Watson test for first-order autocorrelation do not truly ask if the horizon should be extended, since first-order autocorrelation is explicitly identified and clearly in our field of vision. Diagnostic tests such as goodness-of-fit tests, without explicit alternative hypotheses, are useless since, if the sample size is large enough, any maintained hypothesis will be rejected (for example, no observed distribution is exactly normal). Such tests therefore degenerate into elaborate rituals for measuring the effective sample size.

The only way I know to ask the question whether the horizon is wide enough is to study the anomalies of the data. In the words of the physiologist, C. Bernard:

> A great surgeon performs operations for stones by a single method; later he

[2] In particular, least squares estimates are completely sensitive to the independence assumption, since by choice of sample covariance matrix a generalized least squares estimate can be made to assume any value whatsoever (see my 1981 paper).

> makes a statistical summary of deaths and recoveries, and he concludes from these statistics that the mortality law for this operation is two out of five. Well, I say that this ratio means literally nothing scientifically, and gives no certainty in performing the next operation. What really should be done, instead of gathering facts empirically, is to study them more accurately, each in its special determinism... by statistics, we get a conjecture of greater or less probability about a given case, but never any certainty, never any absolute determinism... only basing itself on experimental determinism can medicine become a true science.
>
> [1927, pp. 137–38]

A study of the anomalies of the data is what I have called "Sherlock Holmes" inference, since Holmes turns statistical inference on its head: "It is a capital mistake to theorize before you have all the evidence. It biases the judgements." Statistical theory counsels us to begin with an elicitation of opinions about the sampling process and its parameters; the theory, in other words. After that, data may be studied in a purely mechanical way. Holmes warns that this biases the judgements, meaning that a theory constructed before seeing the facts can be disastrously inappropriate and psychologically difficult to discard. But if theories are constructed after having studied the data, it is difficult to establish by how much, if at all, the data favor the data-instigated hypothesis. For example, suppose I think that a certain coefficient ought to be positive, and my reaction to the anomalous result of a negative estimate is to find another variable to include in the equation so that the estimate is positive. Have I found evidence that the coefficient is positive? It would seem that we should require evidence that is more convincing than the traditional standard. I have proposed a method for discounting such evidence (1974). Initially, when you regress yield on fertilizer as in equation (2), you are required to assess a prior distribution for the experimental bias parameter β^*; that is, you must select the misspecification matrix M. Then, when the least squares estimate of β turns out to be negative, and you decide to include in the equation the light level as well as the fertilizer level, you are obligated to form a prior for the light coefficient γ consistent with the prior for β^*, given that $\beta^* = \gamma r_1$, where r_1 is the regression coefficient of light on fertilizer.[3]

This method for discounting the output of exploratory data analysis requires a discipline that is lacking even in its author. It is consequently important that we reduce the risk of Holmesian discoveries by extending the horizon reasonably far. The degree of a polynomial or the order of a distributed lag need not be data instigated, since the horizon is easily extended to include high degrees and high orders. It is similarly wise to ask yourself before examining the data what you would do if the estimate of your favorite coefficient had the wrong sign. If that makes you think of a specific left-out variable, it is better to include it from the beginning.

Though it is wise to select a wide horizon to reduce the risk of Holmesian discoveries, it is mistaken then to analyze a data set as if the horizon were wide enough. Within the limits of a horizon, no revolutionary inference can be made, since all possible inferences are predicted in advance (admittedly, some with low probabilities). Within the horizon, inference and decision can be turned over completely to a computer. But the great human revolutionary discoveries are made when the horizon is extended for reasons that cannot be predicted in advance and cannot be computerized. If you wish to make such discoveries, you will have to poke at the horizon, and poke again.

V. An Example

This rhetoric is understandably tiring. Methodology, like sex, is better demonstrated than discussed, though often better anticipated than experienced. Accordingly, let me give you an example of what all this

[3] In a randomized experiment with $r_1 = 0$, the constraint $\beta^* = \gamma r_1$ is irrelevant, and you are free to play these exploratory games without penalty. This is a very critical difference between randomized experiments and nonrandomized nonexperiments.

ranting and raving is about. I trust you will find it even better in the experience than in the anticipation. A problem of considerable policy importance is whether or not to have capital punishment. If capital punishment had no deterrent value, most of us would prefer not to impose such an irreversible punishment, though, for a significant minority, the pure joy of vengeance is reason enough. The deterrent value of capital punishment is, of course, an empirical issue. The unresolved debate over its effectiveness began when evolution was judging the survival value of the vengeance gene. Nature was unable to make a decisive judgment. Possibly econometricians can.

In Table 1, you will find a list of variables that are hypothesized to influence the murder rate.[4] The data to be examined are state-by-state murder rates in 1950. The variables are divided into three sets. There are four deterrent variables that characterize the criminal justice system, or in economic parlance, the expected out-of-pocket cost of crime. There are four economic variables that measure the opportunity cost of crime. And there are four social/environmental variables that possibly condition the taste for crime. This leaves unmeasured only the expected rewards for criminal behavior, though these are possibly related to the economic and social variables and are otherwise assumed not to vary from state to state.

A simple regression of the murder rate on all these variables leads to the conclusion that each additional execution deters thirteen murders, with a standard error of seven. That seems like such a healthy rate of return, we might want just to randomly draft executees from the population at large. This proposal would be unlikely to withstand the scrutiny of any macroeconomists who are skilled at finding rational expectations equlibria.

The issue I would like to address instead is whether this conclusion is fragile or not. Does it hold up if the list of variables in the model is changed? Individuals with different experiences and different training will find different subsets of the variables to be candidates for omission from the equation. Five different lists of doubtful variables are reported in Table 2. A right winger expects

TABLE 1—VARIABLES USED IN THE ANALYSIS

a. Dependent Variable
   M = Murder rate per 100,000, FBI estimate.
b. Independent Deterrent Variables
   PC = (Conditional) Probability of conviction for murder given commission. Defined by $PC = C/Q$, where C = convictions for murder, $Q = M \cdot NS$, NS = state population. This is to correct for the fact that M is an estimate based on a sample from each state.
   PX = (Conditional) Probability of execution given conviction (average number of executions 1946–50 divided by C).
   T = Median time served in months for murder by prisoners released in 1951.
   $XPOS$ = A dummy equal to 1 if $PX > 0$.
c. Independent Economic Variables
   W = Median income of families in 1949.
   X = Percent of families in 1949 with less than one-half W.
   U = Unemployment rate.
   LF = Labor force participation rate.
d. Independent Social and Environmental Variables
   NW = Percent nonwhite.
   AGE = Percent 15–24 years old.
   URB = Percent urban.
   $MALE$ = Percent male.
   $FAMHO$ = Percent of families that are husband and wife both present families.
   $SOUTH$ = A dummy equal to 1 for southern states (Alabama, Arkansas, Delaware, Florida, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, West Virginia).
e. Weighting Variable
   $SQRTNF$ = Square root of the population of the FBI-reporting region. Note that weighting is done by multiplying variables by $SQRTNF$.
f. Level of Observation
   Observations are for 44 states, 35 executing and 9 nonexecuting. The executing states are: Alabama, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Illinois, Indiana, Kansas, Kentucky, Louisiana, Maryland, Massachusetts, Mississippi, Missouri, Nebraska, Nevada, New Jersey, New Mexico, New York, North Carolina, Ohio, Oklahoma, Oregon, Pennsylvania, South Carolina, South Dakota, Tennessee, Texas, Virginia, Washington, West Virginia.
   The nonexecuting states are: Idaho, Maine, Minnesota, Montana, New Hampshire, Rhode Island, Utah, Wisconsin, Wyoming.

[4] This material is taken from a study by a student of mine, Walter McManus (1982).

TABLE 2—ALTERNATIVE PRIOR SPECIFICATIONS

Prior	*PC*	*PX*	*T*	*XPOS*	*W*	*X*	*U*	*LF*	*NW*	*AGE*	*URB*	*MALE*	*FAMHO*	*SOUTH*
Right Winger	*I*	*I*	*I*	*	*D*	*D*	*D*	*D*	*D*	*D*	*D*	*D*	*D*	*D*
Rational Maximizer	*I*	*I*	*I*	*	*I*	*I*	*I*	*I*	*D*	*D*	*D*	*D*	*D*	*D*
Eye-for-an-Eye	*I*	*I*	*D*	*	*D*	*D*	*D*	*D*	*D*	*D*	*D*	*D*	*D*	*D*
Bleeding Heart	*D*	*D*	*D*	*	*I*	*I*	*I*	*I*	*D*	*D*	*D*	*D*	*D*	*D*
Crime of Passion	*D*	*D*	*D*	*	*I*	*I*	*I*	*I*	*I*	*I*	*I*	*I*	*I*	*I*

Notes: 1) *I* indicates variables considered important by a researcher with the respective prior. Thus, every model considered by the researcher will include these variables. *D* indicates variables considered doubtful by the researcher. * indicates *XPOS*, the dummy equal to 1 for executing states. Each prior was pooled with the data two ways: one with *XPOS* treated as important, and one with it as doubtful.

2) With five basic priors and *XPOS* treated as doubtful or important by each, we get ten alternative prior specifications.

the punishment variables to have an effect, but treats all other variables as doubtful. He wants to know whether the data still favor the large deterrent effect, if he omits some of these doubtful variables. The rational maximizer takes the variables that measure the expected economic return of crime as important, but treats the taste variables as doubtful. The eye-for-an-eye prior treats all variables as doubtful except the probability of execution. An individual with the bleeding heart prior sees murder as the result of economic impoverishment. Finally, if murder is thought to be a crime of passion then the punishment variables are doubtful.

In Table 3, I have listed the extreme estimates that could be found by each of these groups of researchers. The right-winger minimum of -22.56 means that a regression of the murder rate data on the three punishment variables and a suitably selected linear combination of the other variables yields an estimate of the deterrent effect equal to 22.56 lives per execution. It is possible also to find an estimate of $-.86$. Anything between these two extremes can be similarly obtained; but no estimate outside this interval can be generated no matter how the doubtful variables are manipulated (linearly). Thus the right winger can report that the inference from this data set that executions deter murders is not fragile. The rational maximizer similarly finds that conclusion insensitive to choice of model, but the other three priors allow execution actually to encourage murder, possibly by a brutalizing effect on society.

TABLE 3—EXTREME ESTIMATES OF THE EFFECT OF EXECUTIONS ON MURDERS

Prior	Minimum Estimate	Maximum Estimate
Right Winger	−22.56	−.86
Rational Maximizer	−15.91	−10.24
Eye-for-an-Eye	−28.66	1.91
Bleeding Heart	−25.59	12.37
Crime of Passion	−17.32	4.10

Note: Least squares is −13.22 with a standard error of 7.2.

I come away from a study of Table 3 with the feeling that any inference from these data about the deterrent effect of capital punishment is too fragile to be believed. It is possible credibly to narrow the set of assumptions, but I do not think that a credibly large set of alternative assumptions will lead to a sharp set of estimates. In another paper (1982), I found a narrower set of priors still leads to inconclusive inferences. And I have ignored the important simultaneity issue (the death penalty may have been imposed in crime ridden states to deter murder) which is often a source of great inferential fragility.

VI. Conclusions

After three decades of churning out estimates, the econometrics club finds itself under critical scrutiny and faces incredulity as never before. Fischer Black writes of "The Trouble with Econometric Models." David

Hendry queries "Econometrics: Alchemy or Science?" John W. Pratt and Robert Schlaifer question our understanding of "The Nature and Discovery of Structure." And Christopher Sims suggests blending "Macroeconomics and Reality."

It is apparent that I too am troubled by the fumes which leak from our computing centers. I believe serious attention to two words would sweeten the atmosphere of econometric discourse. These are whimsy and fragility. In order to draw inferences from data as described by econometric texts, it is necessary to make whimsical assumptions. The professional audience consequently and properly withholds belief until an inference is shown to be adequately insensitive to the choice of assumptions. The haphazard way we individually and collectively study the fragility of inferences leaves most of us unconvinced that any inference is believable. If we are to make effective use of our scarce data resource, it is therefore important that we study fragility in a much more systematic way. If it turns out that almost all inferences from economic data are fragile, I suppose we shall have to revert to our old methods lest we lose our customers in government, business, and on the boardwalk at Atlantic City.

REFERENCES

Bernard, C., *An Introduction to the Study of Experimental Method*, New York: MacMillan, 1927.

Black, Fischer, "The Trouble with Econometric Models," *Financial Analysts Journal*, March/April 1982, *35*, 3–11.

Friedman, Milton, *A Theory of the Consumption Function*, Princeton: Princeton University Press, 1957.

Hendry, David, "Econometrics—Alchemy or Science?," *Economica*, November 1980, *47*, 387–406.

Hoerl, Arthur E. and Kennard, Robert W., "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, February 1970, *12*, 55–67.

Kuhn, Thomas S., *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press, 1962.

Lakatos, Imre, "Falsification and the Methodology of Scientific Research Programmes," in his and A. Musgrave, eds., *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press, 1969.

Leamer, Edward E., "A Class of Prior Distributions and Distributed Lag Analysis," *Econometrica*, November 1972, *40*, 1059–81.

_____, "False Models and Post-data Model Construction," *Journal American Statistical Association*, March 1974, *69*, 122–31.

_____, *Specification Searches: Ad Hoc Inference with Non-experimental Data*, New York: Wiley, 1978.

_____, "Techniques for Estimation with Incomplete Assumptions," *IEEE Conference on Decision and Control*, San Diego, December 1981.

_____, "Sets of Posterior Means with Bounded Variance Priors," *Econometrica*, May 1982, *50*, 725–36.

McManus, Walter, "Bayesian Estimation of the Deterrent Effect of Capital Punishment," mimeo., University of California-Los Angeles, 1981.

Plott, Charles R. and Smith, Vernon L., "An Experimental Examination of Two Exchange Institutions," *Review of Economic Studies*, February 1978, *45*, 133–53.

Polanyi, Michael, *Personal Knowledge*, New York: Harper and Row, 1964.

Pratt, John W. and Schlaifer, Robert, "On the Nature and Discovery of Structure," mimeo., 1979.

Shiller, Robert, "A Distributed Lag Estimator Derived From Smoothness Priors," *Econometrica*, July 1973, *41*, 775–88.

Sims, C. A., "Macroeconomics and Reality," *Econometrica*, January 1980, *48*, 1–48.

_____, "Scientific Standards in Econometric Modeling," mimeo., 1982.

Smith, Vernon L., "Relevance of Laboratory Experiments to Testing Resource Allocation Theory," in J. Kmenta and J. Ramsey, eds., *Evaluation of Econometric Models*, New York: Academic Press, 1980, 345–77.