The Role of Statistical Heuristics in Public Policy Analysis

Gregory G. Brunk

Since policy prescriptions frequently are contradictory, people commonly suspect that many researchers manipulate their statistical findings to justify preconceived notions. An increasing awareness of divergent policy prescriptions undermines public, professional, and academic confidence in all areas of policy analysis (see Beardsley 1980, Lovell 1983, Diesing 1985, Leamer 1983, McKown 1986, and particularly McAdams 1984). Similar criticisms routinely are leveled against econometric models, many of which are not very reliable, rest on unstated—and sometimes unrecognized—assumptions, and suffer from other serious difficulties (see Cooper 1972, Porter et al. 1981, and Ascher 1982).

Since the behavioral revolution, applied social scientists have put much effort into searching for mechanistic solutions to society's problems. Many thought such solutions would emerge through simple analysis of empirical evidence, and they expected the mechanistic approach to bring a consensus, not dissension, among policy experts. Neither expectation has proven true. As a result of the failure of naive empiricism, statistically minded critics of the current state of public policy analysis face great frustration. We are unable to propose any formula-like solutions to this malaise because many problems that plague policy analysis result from our largely nontheoretical examination of empirical evidence.

Cato Journal, Vol. 9, No. 1 (Spring/Summer 1989). Copyright © Cato Institute. All rights reserved.

The author is Associate Professor of Political Science at the University of Oklahoma. He wishes to thank Howard Tamashiro for his helpful comments and particularly for his initial suggestion that it would be worthwhile to gather empirical evidence on regression modeling, as done in this paper. The author also thanks his colleagues and graduate students with whom he had a number of heated debates about the role of statistics and theory in policy analysis. Finally, he expresses his appreciation to the graduate students who participated in his advanced regression course and who helped him gather the empirical evidence used in this study.

Faced with the current situation, applied policy specialists have developed new warrants to justify their craft. More and more, the experts' divergent policy prescriptions are claimed as a virtue rather than a vexation. Today, policy analysis is often justified on the basis of its potential to clarify opposing positions and to develop sets of feasible alternatives for politicians and bureaucrats. Such a position is ironic since clarification of positions is the traditional virtue of classical political philosophy, and such a defense of policy analysis puts the subfield much closer to social sciences traditionalists than to behavioralists.

This paper has three major purposes: first, to discuss a few commonly encountered sinister explanations for the wide diversity of policy prescriptions; second, to offer an alternative statistical explanation and discuss the very important, but usually overlooked, role that research heuristics play in modeling social processes; and finally, to present the results of an experiment in econometric modeling, in which a series of "well-intentioned" individuals developed a multitude of different models of bank failures for the American Southwest. My arguments suggest that practitioners adopt a different approach to prescriptive analysis—an approach based less on the empirical evidence of single studies and more on principles derived from general theoretical models of behavior that have been verified in previous analyses of related policy areas.

The Traditional Arguments

The sources of divergent policy prescriptions are little understood, but three sinister causes are commonly cited in anecdotal accounts of public policy formation: money, fame, and ideology.

Money

Few of us doubt that the influence of money will be seen in public policy research if analysts think that reaching the "correct" decision will affect their income. "Hired guns" are always available and are willing to offer justifications for almost any conclusion that a powerful politician, interest group, or bureaucrat desires. The problems arising from this process have been intensified by the existence of numerous policy institutes whose research associates must survive from day to day on contract work; these problems multiply when the results from the initial stages of research have an impact on the probability of additional funding.

Fame

A second, seemingly common, source of distortion is subconscious deception caused by a desire to achieve professional attention. History is littered with examples of prominent scientists who deceived themselves in their research. One of the most interesting stories concerns Rene Blondlot, the distinguished French physicist, who later won the Nobel Prize. In 1903 he announced the discovery of "N-rays," a new form of radiation that could be diffracted into visible light by passing it through an aluminum prism (Klotz 1980). The observation of N-rays took a particularly well-trained eye, and soon the existence of this new radiation was confirmed by many observers, who took great pride in their acute scientific perception. The fad began with Blondlot's criterion for hiring lab assistants: He employed only those who could "see" N-rays during their job interviews!

Ideology

Ideology is such a clear factor affecting research that occasionally it is possible to predict what prescriptions will be advocated by various individuals in particularly contentious policy areas, such as the efficacy of the death penalty (Leamer 1983). More generally, it seems likely that an individual's attitude toward "big government" plays a role in how one feels about problems caused by policy mistakes. Conservatives and liberals within government differ mostly about policy areas in which government should have a greater role.

Suppose "Star Wars" will not work and the administration's analysis is wrong. How would a commentator, who wants to expand the size of the American military, view such a possibility? The commentator might be more accepting of the consequences of the administration's forecasting mistakes since such errors tend to give the military more funding and power. This error would be dismissed as an evil that has desirable consequences. Such policy mistakes have probably enlarged the size of government, particularly when students of the craft are told they should propose "policy relevant" models.

An Alternative Explanation for Diverse Policy Prescriptions

This paper explores a less-sinister explanation for the many instances in which professionals disagree on what is correct policy. Much effort has been expended in the social sciences to find algorithmic solutions to policy problems (Tamashiro 1984, p. 205), and social scientists have diligently pursued the El Dorado of methodology, hoping to find some statistical procedure for all situations. My argument is that many policy disagreements arise from the heuristics, or rules of thumb, that are used—often unknowingly—by researchers. These disagreements are compounded by common data and computational errors (Dewald, Thurdby, and Anderson 1986). The issue of statistical heuristics has received little attention in the scholarly literature and is generally ignored in most statistics and econometrics textbooks, which concentrate on classical hypothesis testing and give readers a false sense of security about statistical results. These heuristics have seen even less attention in the applied literature (Hennessy 1982).

Why has there been so little study of rules of thumb used in practical research situations? Perhaps it is because we think these heuristics are unscientific, but our perception of their role in developing knowledge has changed with the increased attention given to artificial intelligence (for some policy examples, see Tamashiro 1981, 1984; Tamashiro and Brunk 1985).

Many standard research techniques were developed by statisticians, who were interested primarily in the abstract properties of their mathematical discoveries, although some early statisticians, including R. A. Fisher and Karl Pearson, were more applications oriented. While econometricians largely have been interested in the practical characteristics of statistics, they too have confined their research primarily to technical issues like robustness, biasedness, and efficiency. Such studies offer little practical help in assessing the validity and reliability of our collective research efforts. We have almost no evidence concerning how statistics actually are used by practitioners. Because the issue is complex, no one yet has analyzed most of the biases that may be induced by higher-order modeling heuristics (Klitgaard, Dadabhoy, and Litkouchi 1981, p. 103).

Statistical Heuristics

An algorithm is "any procedure that can be carried out in a 'mechanical' fashion (i.e., according to fixed rules) without the great need for intellectual judgment or initiative" (Tamashiro 1984, p. 204). In the social sciences, algorithms often are mathematical formulas that guarantee an optimal solution to a problem when certain conditions are met. Algorithms are the heart of all computer statistics programs. Heuristics, on the other hand, evolve from practical experience in certain types of problem-solving activities and are informed guesses about efficient rules that can be used for problem solving (Tamashiro and Brunk 1985). These two approaches are not total opposites. Depending on the context, an algorithm often can be used as a heuristic. Algorithms are most effective in analyzing well-ordered, well-defined, and well-understood processes, few of which yet exist in the social sciences. Heuristics are more useful in less well-charted areas.

The statistical algorithm that is used in stepwise regression guarantees that we will "explain" the most variance for a given number of variables. But maximizing explained variance is never our ultimate goal—rather it is only a means to an end (see King 1986). Instead, social scientists are interested in explaining, understanding, predicting, or controlling social phenomena. Any statistical algorithm is only a *method* that we use in achieving one of these goals. A particular algorithm may, or may not, be of much use to us, and sophisticated researchers treat all algorithms—usually unconsciously—as heuristics.

The nature of the algorithm problem in theory building is easy to demonstrate using stepwise regression. Suppose our goal is to understand a process, and we decide to use stepwise regression as a mechanism to help us. The technique has serious limitations when there is extensive correlation among independent variables. Let us assume that two variables X_1 and X_2 are both very strong and theoretically meaningful predictors of the dependent variable, but are very highly correlated with each other. Further, the correlation between X_1 and the dependent variable Y is .94, and the correlation between X_2 and Y is .93. In this case, only X_1 will enter the equation as an explanatory variable, but X_2 will not, because X_2 will not explain enough of the residual variance to meet any reasonable cutoff criteria. As a result, we will conclude that X_1 is a significant predictor of Y, but that X_2 is not—a false theoretical conclusion. As an algorithm, stepwise regression has worked perfectly. The technique has identified the variable that explains the most variance, but substantively it has been a total failure.

Let us elaborate on this example. Suppose that Y is the number of missile attacks against merchant shipping in the Persian Gulf each month during 1987, while X_1 is the intelligence estimate of the number of anti-ship missiles held by Iraq and X_2 is the number held by Iran. Since the two countries had been at war for six years, we would expect a high correlation between the level of armaments of the belligerents. What does our stepwise regression model advise us to offer as a prescription to merchant shippers? Since only the number of anti-ship missiles held by Iraq is a significant predictor of attacks, captains should feel free to ignore information concerning Iranian supplies. This seems an odd prescription. Suppose an accident occurs in Iraq, similar to the one that destroyed all armaments of the Soviet Union's Northern Fleet in May 1984. Our model would advise merchant shippers to relax because the number of attacks on vessels is only a function of the number of Iraqi armaments. The missiles held by Iran are unimportant!

The above example begs for a policy specialist who has more than simply statistical expertise. A major policy-related assumption of regression is that we have eliminated "specification error," which means that all the correct independent variables are found in an equation and that no independent variables are included that are not causal. The economists' phrase "all things being equal" means that any variable not included in an equation is uncorrelated with those that are; thus the omitted variables will not bias the coefficients. All algorithms that have been proposed to help researchers in regression modeling are an attempt to solve the specification problem.

In response to realizing that each newly proposed algorithm is flawed from the standpoint of theory building, researchers have proposed a variety of further problem-solving devices. These devices always begin as a rule of thumb and eventually some become algorithms, but by the time that they reach the statistics books, most of their substantive content has been lost and they *implicitly* are presented as devices that guarantee solutions to problems. No algorithm has had total success in building or discovering theory, because associated with each algorithm is some commonly occurring instance when it will not yield the correct theoretical model.

Once there was general agreement that a basic criterion necessary to define a discipline as a science was its ability to produce hypotheses that could empirically be disproven (Friedman 1953, Popper 1959). Beginning with Hanson (1958), the problem began to be viewed quite differently. To quote Diesing (1985, p. 63):

> Philosophers of science used to assert that the testing problem was to state a hypothesis or prediction in clear, precise terms so that facts could clearly contradict it. Now we realize that the problem is also to produce facts robust enough to stand up to a passionately held hypothesis and knock it down. Science used to be described as a process of asking clear, specific questions of Nature so Nature could answer Yes or No. Now we realize that when we ask a question, Nature mumbles or speaks in riddles. Consequently, when hypotheses and data seem to disagree, we can either declare the hypothesis refuted or declare the data misleading and in need of adjustment or reinterpretation. Conversely when hypothesis and data seem to agree, we can either declare the hypothesis confirmed or reject the data as atypical, badly adjusted, or contaminated by hidden variables, so that the agreement is a coincidence.

In looking for theories of processes, our efforts often shift from study to study, making it impossible to develop an all-purpose algorithm that can produce reasonable solutions for all our practical problems. This situation is similar to one often encountered in strategic planning (Tamashiro 1984). Take the following set of arguments by Leamer (1983, pp. 36–37):

> 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 advise 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.

Heuristics in Regression Models

A heuristic is a problem-solving rule of thumb. Heuristics do not guarantee solutions to problems, but they are methods that have been time tested and found to be useful. A major part of our problem is that an infinite number of models can be proposed to explain each data set, and we need some mechanism to reduce the number of potential models to a manageable group (Leamer 1983). One of the first statistical heuristics proposed was the critical significance level. It is important to note the difference between statistical and substantive significance. Statistical significance merely indicates that the probability of an observed relationship being a chance occurrence is very low. When we examine a process, however, we are interested in the variables that have a substantive or practical effect. Although statistical significance often is used as a heuristic to identify substantively significant variables, these two concepts are not the same.

Heuristics are particularly useful in modeling complex political phenomena where flexibility is required, but so too is a consistency in presenting evidence. One source of these heuristics derives from the fact that statistical evidence is often contradictory and hard to evaluate, but it is illegitimate to use wildly different methods of evaluation in the same study. Using different standards idiosyncratically is in bad taste, giving the impression that you are trying to prove your pet theory, and most people find gross inconsistency objectionable. Even in politics, consistency is recognized as a time-saving device that simplifies argumentation, which is why the strategy is adopted by many politicians (Asher and Weisberg 1978).

While researchers rarely make their heuristics explicit—often because they do not realize that they are doing anything novel—it is possible to catalog some rules of thumb in statistical arguments. Some come from the tradition of exploratory data analysis (Tukey 1977, Mosteller and Tukey 1971), while others have evolved independently. Many heuristics also are algorithms that do not focus on theory building; these nontheoretical algorithms emphasize other criteria, such as the maximization of explained variance. Most have numerous permutations, and a few will take researchers in opposing directions (see "Number of Statistically Significant Coefficients" and compare "Occam's Razor" with "Policy Relevance"). To simplify the discussion of these statistical heuristics, the following sections will be restricted to rules of thumb used with one of the most popular modeling approaches: regression analysis.

Heuristics for the Construction of Models

Linearity

The most commonly encountered rule of thumb is the assumption that most social processes are linear additive functions. This heuristic seems to have developed from two observations. First, the more assumptions about a process when modeling, the more the possibility of error; and a curve often requires at least one more parameter. Second, linear equations can be handled conveniently in mathematics, unlike nonlinear equation systems.

All Inclusiveness

Collect all the available independent variables and use them in one giant equation. If a smaller number of variables eventually are used in the final equation, claim that the significant coefficients represent theoretically important relationships. If the potential number of variables used is so great that you might be ridiculed as a "barefoot empiricist," apply a second step. Take only those variables that are statistically significant (as opposed to substantively significant) in the first equation and run another regression. Iterate this second step as many times as necessary until the number of variables becomes reasonable or all remaining coefficients are significant. This step is sometimes referred to as model pruning, a method that also has been recommended to obtain stable factor scores. Leamer (1983, p. 36) argues that "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." He writes (p. 37) that "The consuming public is hardly fooled by this chicanery. The econometrician's shabby art is humorously and disparagingly labeled 'data mining,' 'fishing,' 'grubbing,' 'number crunching.' "

Maximize Explained Variance

Use *all* the potential independent variables that can be gathered to maximize explained variance or the significance level of the overall F statistic. R^2 is maximized using ordinary least squares, but if one thinks about explained variance, it is unclear why a researcher would care much about maximizing R^2 in developing most simple, predictive models (King 1986). By necessity, R^2 increases monotonically as the number of independent variables increases and reaches a maximum of 1.00 by the time that N-*1* independent variables are used as predictors. Corrected R^2, on the other hand, takes into account the number of predictor variables and subtracts from R^2 the amount of variance that one would expect to explain by chance. Corrected R^2 is used more commonly in economics than the other social sciences because the level of statistically explained variance in economics generally is quite high. The various R^2 approaches can be defended as maximum likelihood procedures, but stepwise regression, maximization of R^2, and maximization of t scores may produce radically different models if the independent variables are correlated. (For more on various interpretations that can be given to R^2 in complex models, see Luskin 1984.)

Stepwise Regression

Another series of heuristics is based on maximization of explained variance in some sequential order of steps. In forward stepwise regression, variables are added one at a time until a cutoff criterion is reached. The most common criteria that are used to determine inclusion are an additional increment of explained variance or the significance level of a coefficient. In backward stepwise regression, all variables initially are included in an equation and discarded one at a time. Various criteria have been proposed for inclusion, and these sometimes can produce radically different equations (Bendel and Afifi 1977, Hocking 1976, Berk 1978).

Simple Correlations as a Screening Device

Since stepwise regression procedures can produce radically different results, some researchers have preferred a two-step process.

First, examine the simple correlations between each independent variable and the dependent variable; then use the correlations as a screening device (Gough and Brunk 1980). According to this heuristic, only those variables that have significant simple correlations are included in the final regression equation. Unfortunately, this heuristic is susceptible to missing important nominal- and ordinal-level control variables in individual-level analyses. These control variables will have significant coefficients only when included in a properly specified model. Generally with contests (Thorngate and Carroll 1987), the more steps introduced in a methodological process, the higher the probability that the true or best result will not be identified.

Automatic Interaction Detection

Another heuristic that is highly susceptible to random data error is automatic interaction detection (AID). This algorithm computes all possible interactions among variables and uses the original variables and the interactions to predict the dependent variable. If one is using a 0.05 level of significance, there is a 5 percent probability of finding a significant relationship by chance for each interaction. Interaction effects rapidly multiply the number of potential explanatory factors, and a substantial increase will occur in both the number of significant variables and the level of explained variance (see Lovell 1983).

Heuristics for the Evaluation of Models

Having found a candidate model to explain a relationship, we see a multitude of criteria that can be used to evaluate whether such a model is reasonable. As is true with the first set of heuristics, these criteria often cause the selection of different models of a relationship. In regression, the most commonly applied evaluation criteria include no autocorrelation, white noise, sensitivity tests, and division of the sample.

No Autocorrelation

When errors are correlated over time or space, the use of ordinary least squares is inefficient. The most common problems in time series data are the presence of waves or a portion of a wave. In most instances, it is difficult to tell whether autocorrelation will inflate or deflate the value of a coefficient. The outcome depends (among other things) on whether the data series starts in the peak or valley of a wave.

Autocorrelation can be addressed in two basic ways: first, by autoregressive integrated moving average (ARIMA) or similar style modeling of a process, which admittedly is an "art" rather than a science.

This method often is entirely atheoretical because ARIMA modeling tells us little about the processes that created a data series. Luckily, many economic and political series are only AR(1) and can be corrected easily. Unluckily, some of the most interesting processes seem to have time-dependent or ground-effect parameters. A second way to address autocorrelation is to understand the variables that have generated a process and to include them in an equation. Many times the correct theoretical model will remove autocorrelation without the necessity of including atheoretically estimated AR parameters.

White Noise

Ideally, residuals should be white noise, which means that the errors from an equation should be uncorrelated and without any sort of pattern. If the residuals are not white noise, some process has been left out of the regression that still can be modeled. This means, of necessity, that we have a specification problem and that our estimated coefficients are biased. Examination of residual plots can be quite successful in offering researchers insight for further modeling (Daniel and Woods 1980, Larson and McCleary 1972, Tukey 1977).

Sensitivity Tests

Various sensitivity tests can be applied to the final equation. The most reasonable test is to delete each case from the analysis and see if coefficients remain stable. One occasionally sees the deletion of each variable from an equation as well, but that procedure invites specification problems. Various rules of thumb have been proposed for deleting outliers in regression analysis. The most common are based on the number of standard deviations that an observation is from the mean of a variable. Other criteria are based on joint probability distributions.

Division of the Sample

If enough evidence is available, divide a sample into two portions. Use the first half to estimate a model; then examine the model's validity with the reserved second half of the data.

Heuristics for Comparing Models

Given a series of plausible models of a relationship, a number of criteria can be used to choose among them, such as parsimony or Occam's Razor, theoretically meaningful intercept, number of statistically significant coefficients, policy relevance, theoretical justification, correct signs of coefficients, and the critical experiment.

Parsimony or Occam's Razor

This approach was first proposed by William of Occam. Early science and religion often were treated as two sides of the same coin, and religious men, such as Gregor Mendel, gave us many important scientific theories. Occam's rule was quickly accepted by the scientists of his day, who wanted to remove supernatural factors, such as the intervention of angels, from explanations of common physical events. Occam argued that it was preferential to say, "A causes B," rather than "A is implemented by one of God's angels to cause B." If the angel is busy one day or you have offended God, B will not occur. Occam's Razor helped remove religious influences from modern science but was equally successful in solving more complex problems of causation, which is why it is cited today. If one theory of behavior has four steps and another theory has three, the theory with only three steps is most likely correct because nature prefers simplicity.

Statistically, an F test can be used to see if one model explains significantly more variance than its competitors. If there is no significant difference in the levels of explained variance, choose the model with the smallest number of parameters. It is reasonable to think that simple effects are more common than interactions and to argue for finite causation, disregarding those effects for which a strong case cannot be made (Campbell and Stanley 1966). Unless a good theoretical argument supports a complex model, the model probably will fall apart when replicated.

Theoretically Meaningful Intercept

If two models are otherwise equal, choose the alternative in which the prediction of "the value of Y when all X's equal zero" is most reasonable from a substantive standpoint.

Number of Statistically Significant Coefficients

Given two models with an equal number of estimated coefficients that explain about the same amount of variance, choose the model with the largest number of significant coefficients (if you prefer a complex system) and with the least number of significant coefficients (if you prefer Newtonian simplicity). This choice really is a matter of personal taste, although it seems that sociologists and causal modelers prefer complexity, while political scientists and regression modelers prefer simplicity.

Policy Relevance

An example of a "good" reason for more elaborate models comes from a rule of thumb called "policy relevance." According to this heuristic, models of social processes are not useful to policymakers unless they contain variables that can be manipulated by government. Applied policy analysts choose this method because their most important goal is to offer public officials policy alternatives for the potential resolution of society's problems (Kash and Ballard 1987, pp. 601–2).

There are two important, and undesirable, consequences of this rule of thumb. First, "policy relevance" promotes pro-statist ideologies since it encourages developing a literature that makes it seem that changes in society can easily be affected by governmental action. This situation comes about because the heuristic tells researchers to search in a systematic manner for variables that can be manipulated by government. When this search strategy is combined with statistical errors caused by inclusion of spurious variables because of the level of significance selected by the researcher (for example, a 0.05 level results in 5 percent of tested noncausal variables appearing to be important causal factors), the number of seemingly manipulatable variables affecting society's behavior will be much greater than if researchers did not use the "policy relevance" heuristic when choosing variables. Despite the predominance of such a pro-interventionist literature, a legitimate debate remains over how much impact governmental actions can have on many types of behavior and even whether laws can have an impact on some very fundamental types of behavior (see Lewis-Beck and Alford 1980).

The second consequence of years of applying this heuristic is that we are much more self-confident about the efficacy of public policy analysis than would be justified by applying a tradition of empirical analysis that did not stress including independent variables that can be manipulated by policymakers. This "policy relevance" approach has encouraged the nonrandom production of a particular type of model. While these models tell us that what we public policy analysts do has great value for society, we can be sure that the "policy relevance" rule of thumb has resulted in a substantial overestimation of our potential usefulness.

Theoretical Justification

Use only variables that are explicitly suggested by deductive theory or that appear in the literature. Nothing is more frustrating than having constructed a large data set and having found that your best predictor of Argentine social welfare expenditures is coal production in Chile. While there may be a connection between the variables, it is unlikely, but if you think about them long enough, you surely can come up with something. This version of the "garbage in—garbage out" problem is frequently encountered in factor analysis. You would have been better off to not have collected the atheoretical data in the beginning.

Correct Signs of Coefficients

The signs of all coefficients should be in the correct direction. One of the easiest ways to "cook" data analyses is through an ex post facto determination of the sign of a variable to make it appear to be "policy relevant." This route allows one to find 5 percent more "significant" variables using a 0.05 significance level for one-tailed tests because this is a 0.10 level for a two-tailed tests. Such chance relationships will not be sustained in future studies.

The Critical Experiment

Attempt to conduct a "critical experiment" in which your candidate models predict that radically different outcomes should be observed. Although this strategy has not been used much in the social sciences (for example, Reed and Brunk 1984), it has a long tradition in the physical sciences. With luck, one model will predict correctly and the others will not.

An Experiment in Modeling a Process with Policy Implications

While a number of simulations have investigated the efficiency of particular statistics, only a few simulations exist of research heuristics (see Granger and Newbold 1974, Lovell 1983). Boyne (1985) summarizes over 100 different studies of public policy outputs and finds that many differences between studies seem to depend on the variables or rules of thumb used by researchers. The total number of independent variables also appears to have a major effect on the results of policy studies (Lewis-Beck 1977). Similarly, studies of various sorts of contests indicate that the best person rarely wins any complicated event (Thorngate and Carroll 1987). Unfortunately, the existing quantitative evidence on these issues has reached few practitioners, and we know little systematically about how higher-level heuristics affect research outcomes. Because evidence on these issues is so scattered, some practitioners may not be convinced by a simple recitation of ways that different prescriptive models can evolve from applying statistical techniques. In fact, varied policy prescriptions often will be proposed because of different research heuristics. What is needed is a demonstration of the problem's consequences using a typical data set of the sort often encountered in applied policy situations.

If research heuristics commonly have only small impacts on the conclusions of most studies, then the net impact of research heuristics on policy analysis will be slight, and we will have few problems from this source. But if research heuristics can have a major impact on our results—and anecdotal evidence suggests they often have major impacts—then we will have to recognize that, even in principle, the foundations of policy analysis can never rest primarily on simple mechanical analyses of data. Instead, sound policy analysis will have to be based on a large and explicit body of theory, much of which has yet to be developed.

For my graduate course in advanced regression, the final exam is a modeling exercise that takes about three weeks to complete. Last year's class consisted largely of Ph.D. students in political science, MPA and DPA students in public administration, and people already employed in relatively high positions in the state bureaucracy. A number of class members would soon become professors teaching policy analysis or public administration, or they would take jobs with state, local, or national agencies. The course covered, in part, correlations, partial correlations, assumptions of regression, multiple regression, stepwise regression, analysis of variance, *F* statistics for models, ideas of heteroscedasticity, autocorrelation, white noise, dummy variables, interaction variables, intervention analysis, curvilinear regression, and various transformations, including logarithms. One's grade depended on offering a correct interpretation of each statistic and demonstrating an ability to model social processes. Individuals who completed the course have much more practical training in applied statistics than the majority of academic social scientists and applied policy specialists.

As part of their final exam, students must model the yearly number of bank failures in a southwestern state from 1933 to 1985. They used the type of evidence that might be called on when offering policy prescriptions for the current banking crisis. Modelers were provided with reasonable simulated time series, but not the actual data series, to prevent the argument often encountered at professional conventions: "I have found the only *true* model, and all other candidate models are wrong because I have the *truth*." Individuals were not allowed to introduce new variables, but could assume—for the sake of argument—that all variables in the master data set could be justified on the basis of theory.

The situation was made as realistic as possible, and the potential explanatory variables that were included in the data set were those often found in "political economy" explanations. The potential explanatory variables included a banking law passed in 1960 that could be modeled using intervention analysis, the unemployment rate, the political party of the president, the number of individuals employed by the Federal Deposit Insurance Corporation (FDIC) in the state, the average personal income in the state, the proportion of the state's population over age 60, and whether the United States was at war. Models also could be developed that attempted to capture the relationship using time trends. To give the reader some idea of the relationships among the variables, Table 1 presents the correlation matrix for the experiment, using the dependent variable (bank failures) and some of the potential explanatory variables. There is substantial multicollinearity among some of the independent variables, as is common in actual data analysis.

A number of factors constrain how an individual honestly may attempt to model a process. These factors include the nature of the process, the variables available to model a process, the data points available, random error, a person's perception of the nature of the world, and the research heuristics that are used.[1] Since all individuals used the same data series, the evidence each individual had available for modeling was identical, as was the element of random error. Further, the same set of variables initially was defined to be of potential theoretical importance.

While individuals were alerted to the problem of autocorrelation and the difficulties it caused in regression, standard corrections for autocorrelation were not discussed in the course, and no one used a computer program that would estimate an autoregression model. A long period sine wave and various other trends were included in the simulated data, making it impossible to construct a parsimonious model with close to white noise for residuals unless one incorporated time trends. This step was deliberate. It forced individuals to choose among alternative models, and no simple model was quite adequate to fully explain the process, as generally is the case in real-life situations.

The Models

A major dispute in the social sciences concerns the worth of statistical modeling techniques vis-à-vis theoretical models. At the extremes, two camps can be discerned. In one group are the extreme inductive practitioners, who claim that if we were just given enough funding for our research, we could solve most of the world's problems by the sheer force of data analysis. This position was forced on the early

[1] Another factor that might affect a person's ability to model such a process is competence. This factor was controlled by the first course in the methods sequence: Only one-half to two-thirds of the graduate students each year manage to pass the first course.

TABLE 1

CORRELATIONS AMONG MAJOR VARIABLES

	Bank Failures	Percent Unemployed	Party of President	Employees of FDIC	Average Income	War	Percent over Age 60
Percent Unemployed	.01						
Party of President	.12	.17					
Employees of FDIC	−.01	−.08	−.39				
Average Income	−.21	−.14	−.42	.92			
War	−.24	−.18	−.14	−.18	−.12		
Percent over Age 60	−.16	−.18	−.39	.69	.77	.08	
Time	−.31	−.23	−.52	.82	.95	−.03	.78

proponents of factor analysis by their critics, who asked them the annoying question of whether the fairly complex law of gravitational attraction could have been discovered through factor analysis, even though it is a nonlinear function. In the other camp are the pure theoreticians, who, in the extreme, contend that social scientists understand so little about social behavior that we should restrain ourselves from giving any policy advice for the foreseeable future. Such scholars further believe that the accumulation of more and more data is not the simple key that will bring the development of better theory.

The present empirical evidence may shed some small light on this issue. If the extreme inductivist position is correct, the models produced by disinterested, well-intentioned, and reasonably competent individuals will tend to be similar. On the other hand, if sound theory motivates most stable statistical analyses, the models produced should be quite different.

The evidence pertaining to this issue is presented in Table 2. At first glance it is clear that these well-intentioned individuals developed many different models. Although encouraged to discuss the process among themselves, which should have minimized variation among the models, 13 different models were developed by 15 individuals. In examining the sources of these various models, one of the two pairs of identical models ("Occam's Razor" and "Theoretically Meaningful Intercept") was found to result from individuals working together, while the other pair ("Sensitivity Tests" and "Division of the Sample") seems to have been produced by using the same automatic stepwise regression routine. The level of explained variance (R^2) ranged from 4 percent to 61 percent, while the corrected variance ($\bar{R}^2$) ranged from zero to 56 percent.[2]

An examination of the written justifications that individuals gave for their models indicates that all the presentations are at least minimally defensible, and some substantive justifications are quite creative. One modeler cited a local newspaper article that claimed the overzealousness of the Federal Deposit Insurance Corporation was responsible for most bank failures—not the actions of the banks themselves. According to this argument, the region's banks have always been reckless. The amount of risk taking has not changed over time; all that has changed is the FDIC's practice of closing almost-bankrupt institutions. The "solution" to the problem is to go

[2] The three individuals who developed models explaining less than 20 percent of the variance will be ignored in the remainder of this discussion.

back to the "good old days" when there was little supervision of local banks.

The first striking feature of the evidence in Table 2 is the remarkable difference in level of variance explained by the modelers. The second striking feature is the substantial difference in the variables that were used to predict bank failures. Three modelers argue that the unemployment rate, as a policy prescription, is significant in forecasting bank failures; thus political manipulation of the economy might be appropriate. But four others contend that economic conditions are not related to bank problems. Only one modeler found that the 1960 banking law had any significant impact, three concluded that increasing state income over time had a significant impact, and only one found that war had a significant impact.

Our discussion of research heuristics and the results of this simple experiment should raise grave doubts about the naive assumption that a group of well-intentioned individuals, who have no ideological axes to grind, will develop similar models of a policy process. In fact, the evidence indicates that *the bare-foot empiricist position is bankrupt.* This argument is similar to the recent conclusion regarding the theory of games published by Thorngate and Carroll (1987). They show that in most types of complicated contests, the most-qualified person rarely wins.[3] Similarly in policy analysis, the more complicated the process and the less the theoretical guidance, the more likely that policy prescriptions will diverge when based on data alone. The problem appears to be so serious that in most complicated situations we can assume that policy prescriptions will diverge most of the time when they are not guided by theoretical considerations.

Conclusion

This paper has examined a number of the heuristics or model-building criteria that researchers use when conducting statistical analysis in the social sciences. While it has been obvious for many years that different scholars often arrive at very different policy prescriptions using similar data sets, an important source of these differences has been neglected. In applied analysis, policy differences often have been ascribed to conscious or unconscious manipulation of data by researchers. In fact, many policy differences seem to come from the research rules of thumb that individuals use when making statistical decisions. While mechanical algorithms are useful in

[3] Besides being embarrassing for policy analysis, this conclusion seriously undermines traditional democratic theory. For a nontraditional logic that justifies traditional voting rules and might have applicability for policy analysis as well, see Riker (1982).

TABLE 2

MODELS OF SIMULATED BANK FAILURES, 1933–85

Model[a]	$\bar{R}^2$	Constant[b]	Time[c]	Banking Law Intervention[d]		Unp[e]	Party[f]	FDIC[g]	Income[h]	War[i]	Age[j]
1	.56	54.31	−2.10*	—	—	−1.11*	—	—	−7.74*	—	—
2	.56	23.56	−1.39*	—	—	—	—	−2.73*	—	—	—
3	.54	46.78	−1.72*	—	—	−.93*	—	—	−6.30*	−2.16	—
4	.33	22.67	−0.93*	.62	5.68	−.36*	−4.90*	.59*	—	−2.84	.32
5	.32	34.67	—	.49	9.69*	−.29	−2.35	.56	−13.71*	—	.49
6	.31	12.62	−0.49*	—	—	−.25	−1.79	.60*	—	−3.63	.40
7	.27	13.78	−0.43*	—	—	—	—	.43*	−2.26	—	—
8	.26	14.56	−0.33	—	—	—	—	.58*	−1.13	—	—
9	.26	13.33	−0.45*	—	—	—	—	.47*	—	—	—
10	.26	13.33	−0.45*	—	—	—	—	.47*	—	—	—
11	.22	22.42	—	.85	—	−.15	−0.41	.96*	−5.13	—	−.03
12	.22	22.42	—	.85	—	−.15	−0.41	.96*	−5.13	—	−.03
13	.16	8.87	−0.58*	—	—	—	—	—	2.63	—	—
14	.03	8.11	—	—	—	—	0.25	—	—	−4.13*	.00
15	.00	9.69	—	—	—	—	−0.03	—	−0.53	—	.04
Base	.35	25.66	−0.72*	.76	8.16*	−.36*	−4.50*	.92	−3.09	−2.95	.37

[a]Other variables found in equations and their coefficients:

Model 1: Change in Income = 0.26* Unp (Time) = 0.03* Time = 0.06*
Model 2: FDIC^2 = 0.00* Time^2 = 0.04*
Model 3: Unp (Time) = 0.03* Time^2 = 0.04*
Model 5: Income^2 = 0.71*
Model 6: Income^2 = 0.00*

The “base model (equation) includes all standard variables to provide a baseline for comparison. In each model, a star indicates a coefficient is significant at the 0.05 level for one-tailed tests.

[b] All constants are significantly different from zero.

[c] Time = Time in years.

[d] Banking Law Intervention: First coefficient is change in slope after passage of the law; second coefficient is the first-year effect of the law.

[e] Unp = Percent unemployed.

[f] Party = Dummy variable for president’s party (0 = Republican, 1 = Democrat).

[g] FDIC = Number of people employed by the Federal Deposit Insurance Corporation in the state.

[h] Income = Average individual income in $1,000s.

[i] War = Dummy variable (0 = Peace, 1 = War).

[j] Age = Percent of state’s population over age 60.

maximizing explained variance, they are only of limited utility in theory building or in modeling any but the simplest processes. They cannot provide a foolproof, mechanical solution to most problems. Individuals can use a multitude of research heuristics when modeling political, economic, and social processes. The choice of a particular heuristic is personal, based on previous experience, and often has a major impact on one's final prescriptive model. At a minimum, the realization of this problem casts grave doubts on our ability to offer reasonable policy prescriptions simply on the basis of inductive statistically reasoning.

What then should be done? In the short run, the strategy of policy practitioners in offering politicians and bureaucrats a set of feasible policy options seems defensible. The political process can decide which set of assumptions about society is most reasonable and can use the technical evidence provided by experts in making decisions for society. But we should realize that the probability of our offering reasonable prescriptions to problems based simply on statistical evidence is smaller, perhaps much smaller, than most of us have believed. Particularly as the problem's complexity increases, the probability of offering an optimal or even a reasonable solution based on empirical evidence is very low unless the researcher is guided by explicit and well-developed theory.

Therefore, in the long run, specialists in the academic side of policy analysis should stress theory building. Of equal importance, in teaching future policy practitioners, we should attempt to define as clearly as possible both the areas in which our theories are well developed and those areas in which our theories are soft and fuzzy. The goal of academics should be to develop theories that tell practitioners in advance what variables are important for policymaking, and ideally our theories should tell practitioners this independently of any given statistical analysis. The role of methodology in most applied analysis should become the estimation of the strength of parameters when relationships are well understood, rather than the inductive modeling of processes that are little understood.

Until we have developed better theories of many social processes, we can be sure that ideological debates will continue to flourish in policy analysis. Even when we clearly can demonstrate that researchers have not manipulated their evidence to support their preconceived positions, the methodological heuristics that individuals employ often bring them to advocate radically different inductive models of statistical relationships, and hence to very different public policy prescriptions.

References

Ames, Edward, and Reiter, Stanley. "Distributions of Correlation Coefficients in Economic Time Series." *Journal of the American Statistical Association* 56 (1961): 637–56.

Ascher, William. "Political Forecasting: The Missing Link." *Journal of Forecasting* 1 (1982): 227–39.

Asher, Herbert B., and Weisberg, Herbert F. "Voting Change in Congress: Some Dynamic Perspectives on an Evolutionary Process." *American Journal of Political Science* 22 (1978): 391–425.

Beardsley, Philip L. *Redefining Rigor: Ideology and Statistics in Political Inquiry*. Sage: Beverly Hills, 1980.

Bendel, R. B., and Afifi, A. A. "A Comparison of Stopping Rules in Forward 'Stepwise' Regression." *Journal of the American Statistical Association* 72 (1977): 46–53.

Berk, K. N. "Comparing Subset Regression Procedures." *Technometrics* 20 (1978): 1–6.

Boyne, George A. "Review Article: Theory, Methodology, and Results in Political Science—The Case of Output Studies." *British Journal of Political Science* 15 (1985): 473–515.

Campbell, Donald T., and Stanley, Julian C. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand-McNally, 1966.

Cooper, Ronald L. "The Predictive Performance of Quarterly Econometric Models of the United States." In *Econometric Models of Cyclical Behavior*. Edited by Bert G. Hickman. New York: National Bureau of Economic Research, 1972.

Daniel, Cuthbert, and Wood, Fred S. *Fitting Equations to Data*. 2d ed. New York: Wiley, 1980.

Dewald, William G.; Thurdby, Jerry G.; and Anderson, Richard G. "Replication of Empirical Economics: The Journal of Money, Credit, and Banking Project." *American Economic Review* 76 (1986): 587–603.

Diesing, Paul. "Hypothesis Testing and Data Interpretation: The Case of Milton Friedman." *Research in the History of Economic Thought and Methodology* 3 (1985): 61–89.

Friedman, Milton. *Essays in Positive Economics*. Chicago: University of Chicago Press, 1953.

Gough, Paul A., and Brunk, Gregory G. "Are Economic Conditions Really Important for New Zealand Elections?" *Political Science* 33 (1980): 1–9.

Granger, C.W.J., and Newbold, P. "Spurious Regression in Econometrics." *Journal of Econometrics* 2 (1974): 111–20.

Hanson, Norwood. *Patterns of Discovery*. Cambridge: Cambridge University Press, 1958.

Hennessy, Michael. "The End of Methodology? A Review Essay on Evaluation Research Methods." *Western Political Quarterly* 10 (1982): 606–12.

Hocking, R. R. "The Analysis and Selection of Variables in Linear Regression." *Biometrics* 32 (1976): 1–49.

Kash, Don E., and Ballard, Steve. "Academic and Applied Policy Analysis: A Comparison." *American Behavioral Scientist* 30 (1987): 597–611.

King, Gary. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science* 30 (1986): 666–87.

Klitgaard, Robert E.; Dadabhoy, Sadequa; and Litkouchi, Simon. "Regression without a Model." *Policy Sciences* 13 (1981): 99–115.

Klotz, Irving M. "The N-Ray Affair." *Scientific American* 242 (1980): 168–75.

Larson, W. A., and McCleary, S. J. "The Use of Partial Residual Plots in Regression Analysis." *Technometrics* 14 (1972): 781–90.

Leamer, Edward E. "Let's Take the Con out of Econometrics." *American Economic Review* 73 (1983): 31–43.

Lewis-Beck, Michael S. "The Relative Importance of Socio-Economic and Political Variables for Public Policy." *American Political Science Review* 71 (1977): 559–66.

Lewis-Beck, Michael S., and Alford, John R. "Can Government Regulate Safety? The Coal Mine Example." *American Political Science Review* 74 (1980): 745–56.

Litterman, Robert B. "Forecasting with Bayesian Vector Autoregressions—Five Years of Experience." *Journal of Business and Economic Statistics* 4 (1986): 25–38.

Lovell, Michael C. "Data Mining." *Review of Economics and Statistics* 65 (1983): 1–12.

Luskin, Robert C. "Looking for R^2: Measuring Explanation outside of OLS," *Political Methodology* 10 (1984): 513–32.

McAdams, John. "The Anti-Policy Analysis." *Policy Studies Journal* 13 (1984): 91–101.

McKown, Robert. "On the Uses of Econometric Models: A Guide for Policy Makers." *Policy Sciences* 19 (1986): 359–80.

Mosteller, Frederick, and Tukey, John W. *Data Analysis and Regression.* Reading: Addison-Wesley, 1971.

Popper, Karl R. *The Logic of Scientific Discovery.* New York: Basic Books, 1959.

Porter, Alan L.; Connolly, Terry; Heikes, Russell G.; and Park, Choon Y. "Misleading Indicators: The Limitations of Multiple Linear Regression in Formation of Policy Recommendations." *Policy Sciences* 13 (1981): 397–418.

Reed, Steven, and Brunk, Gregory G. "A Test of Two Theories of Economically Motivated Voting: The Case of Japan." *Comparative Politics* 17 (1984): 55–66.

Riker, William H. *Liberalism against Populism: A Confrontation between the Theory of Democracy and the Theory of Social Choice.* San Francisco: W. H. Freeman, 1982.

Tamashiro, Howard. *Problem-Solving Heuristics in International Politics.* Ph.D. Dissertation. Columbus: Ohio State University, 1981.

Tamashiro, Howard. "Algorithms, Heuristics, and the Artificial Intelligence Modelling of Strategic Statecraft." In *Foreign Policy Decision Making,* pp. 197–226. Edited by Donald Sylvan and Steve Chan. New York: Praeger, 1984.

Tamashiro, Howard, and Brunk, Gregory G. "Expert Based Systems as Elite Foreign Policy Advisors." In *Proceeding of the I.E.E.E. Symposium on Expert Systems in Government*, pp. 637–46. Silver Spring: Institute of Electrical and Electronics Engineers Press, 1985.

Thorngate, Warren, and Carroll, Barbara. "Why the Best Person Rarely Wins: Some Embarrassing Facts about Contests." *Simulation and Games* 18 (1987): 299–320.

Tukey, John W. *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley, 1977.