The Political Methodologist

Newsletter of the Political Methodology Section
American Political Science Association
Volume 3, Number 2, Fall 1990

Editor: Charles H. Franklin, Washington University

Associate Editor: Larry Bartels, University of Rochester

Contents

Notes From the Editor

Charles H. Franklin
Washington University

Model specification is the fundamental problem of empirical political science. Most critically, our models must capture our substantive hypotheses. Unless we can connect model parameters with substantive implications our empirical efforts are pointless. But this is not always easy. Nor is it always done as well as it might be. The best papers make the connection between parameter and hypothesis explicit and compelling. The worst papers hopelessly muddle the linkage.

With properly specified models, we can develop estimators which have desirable properties such as consistency and efficiency. Without proper specification, our estimators will lack these properties. The problem is knowing what is a proper specification.

The "true model" is a bit like God— comforting to believe in but requiring a substantial leap of faith. The classical theory of statistical inference is an elegant, beautiful, subtle construction, which still has much to offer. But we are reaching our age of doubt. We are becoming aware of how difficult model specification can be and how consequential the specification is in determining the answers we get.

If we agree with Bartels (p. 2) that what we care about is estimating parameters of theoretical interest, then how can we do this when we do not know the proper model specification? One approach is to think of our estimates as conditional on the maintained hypothesis embodied in the model specification. This is useful because it reminds us

In addition to the section members, this issue of *TPM* is being sent to members of APSA who have listed methodology as a field of interest but who have not yet joined the organized section. The subfield is growing rapidly and is extending the limits on our ability to address substantive problems with sophisticated methods. I want to take this opportunity to encourage you to join the organized section. —*CHF*

that our estimates depend upon the model we have chosen to represent the data. However, it immediately raises the question of how our estimates vary with the maintained hypotheses we might choose. In other words, as we change the specification of the model what will be the distribution of parameter estimates? Is this distribution narrow or very wide? The answer is critical.

Bartels has a number of useful suggestions for how we should proceed in the face of uncertainty as to the proper model specification and I will not appropriate his arguments here. What I will do is point to the tension between an interest in parameter estimation and a recognition of the uncertainties of model specification. If the model specification is uncertain, then claims about properties of estimators are exceedingly hard to make within the classical framework. Yet our interest remains in the parameters of an incompletely known model. How shall we go about making inferences when the model itself is uncertain? While Bayesian methods offer one alternative, they are not a panacea. The development of solutions to this problem should be a prominent item on the agenda of political methodology.

Five Approaches to Model Specification

Larry M. Bartels[1]
University of Rochester

> ". . . conventional statistical methods require strong and precise assumptions about the functional relationship among the variables and the behavior of unmeasured causes. In social science applications, these postulates are not supplied by theory. The ensuing logical gap is the principal obstacle to social data analysis and the most challenging intellectual problem facing the social science methodologist." — Achen 1982

When I took over a second-semester graduate course in applied regression from Eric Hanushek, he advised me that the most important thing to teach students in such a course is that regression parameter estimates are random variables. The well known textbook by Hanushek and Jackson (1977) makes this point very effectively by using Monte Carlo experiments to illustrate the statistical properties of various estimators.

Since taking over Hanushek's course (and his book), I have come to believe that the second most important thing to teach students in such a course is that the distributions of our regression parameter estimates in practice are not the ones suggested either by statistical theory or by elegant Monte Carlo experiments. Real data analysis is a very different animal.

A sobering first step in the direction of reality can be taken by simply replicating the Monte Carlo setup, minus the assumption that the "true" (data-generating) model is known in advance. Specification experiments of the sort described by Donald Green elsewhere in this newsletter often suggest that the stochastic error addressed by formal statistical theory is dominated, in practice, by error stemming from uncertainty about the correct specification of a statistical model. My own experience with experiments of this sort is that the observed error for a typical regression coefficient in a relatively simple problem may be several times as large as the idealized error generated from repeated estimation of the "true" model. Although no quantitatively precise generalization is possible, these results suggest that Achen's claim that uncertainty about model specification "is the principal obstacle to social data analysis" is something more than a rhetorical flourish.

How should we proceed in the face of this obstacle? The present outline identifies five distinct approaches to model specification. (Thus, it is similar in spirit to, but more elaborate than, that of Sims (1988).) The five approaches span the range from fundamentally misguided to potentially quite useful; but none is a panacea. One of my aims in setting them out here is to cast some brickbats at the worst aspects of current practice. But more constructively, I hope to provide a clearer understanding of the nature and extent of statistical uncertainty—and of the various more or less useful tactics we have adopted in our struggle to reduce that uncertainty.

It may be helpful to begin with a note on the scope of our topic. In textbooks, "model specification" is sometimes taken quite narrowly to refer to the problem of choosing an appropriate list of explanatory variables in a regression model. Sometimes the scope of the topic is broadened to include questions of functional form and assumptions about the distribution of the disturbance term. But even this broader definition makes it easy to gloss over some important practical problems that are fundamentally similar in their causes and consequences to those encompassed in the usual formulation. The problem of choosing an appropriate population, for example, is essentially identical to the problem of choosing appropriate explanatory variables (since two or more distinct populations can be represented in a pooled model by appropriate use of indicator variables and interaction terms). Problems that arise in the use of aggregated data can be addressed using much the same framework used to address seemingly unrelated problems of model specification (Theil 1971, chap. 11; Hanushek, Jackson, and Kain 1974; Erbring 1989). Time series (Beck 1985) and simultaneous equation models (Bartels 1985, 1991) generate fun-

[1]My views on model specification have been significantly influenced by Christopher H. Achen, who also provided some helpful specific comments on an earlier draft of this essay. Remaining errors and infelicities are, as always, my own.

damentally similar problems. These issues and many others are encompassed more or less directly in the discussion below.

It may also be helpful to say a bit about my own philosophical prejudices. First, I do not believe in "true models" (except for those invented by analysts to generate artificial data). Reality is sufficiently complex (or at least so it seems to me) to make the search for a metaphysically "correct" specification fruitless. Statistical models are, and should be, more or less useful approximations. Nevertheless, the difference between more and less useful approximations to some underlying (and unknown) reality is a crucially important one.

Second, I care fundamentally not about "explaining variance" or making forecasts but about estimating parameters of some theoretical interest. Thus, I exclude from consideration approaches such as vector autoregression (Freeman, Williams, and Lin 1989) and Box-Jenkins techniques (Box and Jenkins 1976), which seem useful for exploratory analysis and forecasting but evade the real problems of model specification by abdicating the estimation of structural parameters.

Finally, I care more about getting answers than about fairness as an abstract virtue. Thus, unbiased or consistent parameter estimates, even if we could have them, would not be sufficient to solve my problem if they came with standard errors a mile wide. The inherent tension between bias and imprecision (or between parsimony and realistic complexity) is, it seems to me, a basic tradeoff that must be managed in any empirical research.

1: Model Selection by the Numbers

Apparently, even the modest amount of work required of an academic is too much for some people. So they strive to remove any element of mental strain from their research by replacing their own judgment with one or another mechanical procedure for specifying models and analyzing data. The most notorious of these procedures is stepwise regression, which rifles through a specified set of potential explanatory variables to find some subset that produces a big R-squared statistic. Given the power of modern computers, it is even possible to do stepwise regression with unobserved variables all around (using the "automatic model modification" feature in LISREL).

The only problem is that the parameter estimates produced by procedures of this sort may bear little relationship to the structural parameter values of theoretical interest. Substantively important variables may be dropped from the model if they happen to be strongly correlated with other important variables; different variables may pick up the same structural effect in different samples; indeed, different stepwise procedures may produce different specifications from the same data, and none of these may correspond to the specification that would actually maximize the stipulated statistical criterion (for example, an adjusted R- squared statistic).

The inferential pitfalls of stepwise regression seem to be more widely recognized now than they were a few years ago. What we need to recognize next is that people can be almost as unthinking as computers. Having learned not to let their stepwise regression programs mechanically select a set of explanatory variables to produce big R-squared statistics, analysts sometimes do their (somewhat less efficient) best to mechanically select a set of explanatory variables to produce big R-squared statistics (or big t-ratios, or few "insignificant" parameter estimates). If the efficient stupidity of stepwise regression is replaced by the inefficient stupidity of human beings adopting the same misguided "by-the-numbers" approach to model specification, then little will have been gained.

2: Specification Tests

Restrictive assumptions are necessary elements of any model specification. But they are often problematic: how restrictive should they be? Specification tests are intended to check the restrictive assumptions in a given model specification by treating them explicitly as statistical hypotheses to be tested. For example, Wald, likelihood ratio, or Lagrange multiplier tests may be used to decide whether a restriction "significantly" decreases the goodness of fit of a given model (Judge et al., chap 21).

One crucial problem in the use of specification tests is that they tend, like most hypothesis tests, to be applied arbitrarily. In the absence of any meaningful substantive criterion for trading off "Type 1" and "Type 2" errors, what is the right principle for choosing an appropriate level of significance? Suppose we can reject (at, say, the .05 level) the null hypothesis that the data satisfy some stipulated restriction. Should we relax that restriction? Doing so may make it impossible to relax some other, even more egregious restriction. Suppose, on the other hand, that we cannot reject some stipulated restriction. With enough data we could. Should our model specification depend upon how many observations we happen to have? (The answer is yes; but there is no reason to expect more than a coincidental correspondence between the dependence produced by reasonable scientific criteria and the dependence produced by a .05 rejection rule.)

In any event, tests of this sort can, at best, be only partial solutions to the problem of model specification. We can never test all of our assumptions at once. Whenever we can test a specific restrictive specification against a more general alternative, we could also have adopted that more general alternative in the first place. The knottiest problems come when we get to alternatives so general that our degrees of freedom are small (producing parameter estimates too

imprecise to be of real use) or negative (in which case the model is underidentified and some or all of the parameters cannot be estimated at all).

One useful byproduct of the specification testing approach has been to focus attention on the statistical properties of conditional estimation strategies, or "pretest estimators" (Judge and Bock 1978; Judge et al., 1985, chap 3). Suppose we decide in advance to specify a restrictive model, test its adequacy against a more general alternative using some specification test, and then adopt the parameter estimates from the restricted model (if the specification test fails to reject the null hypothesis at some prespecified confidence level) or the more general alternative (if the specification test results in rejection of the null hypothesis). For simple strategies of this sort, it is feasible to examine the statistical properties of the resulting parameter estimates much as we would those of the parameter estimates produced by either of the two alternative specifications themselves. Formally specified conditional estimation strategies are only a rough approximation to the complexity of actual data analysis. But in the absence of any fully worked out theory of "ad hoc statistical inference" (Leamer 1978), rigorous analysis of pretest estimators is a useful first step toward accommodating statistical theory to statistical practice.

3: Out-of-Sample Validation

One important practical criticism of stepwise regression and other similar mechanical model selection procedures is that they tend to overfit the data, producing impressive results in the sample at hand that do not stand up—even by their own criterion of "explaining variance"—when applied to new data. The same problem, capitalizing on chance, occurs in an attenuated form when the data have been ransacked casually rather than systematically. Thus, accurate forecasting reappears as a useful measure of statistical validity—but only in the sense, and to the extent, that, in the long run, the right parameter values will tend to outperform the wrong parameter values when applied to new data. Indeed, the deterioration in the fit of a model (or in the pattern of parameter estimates it produces) when it is applied to new data is a useful measure of the extent to which its original success was based on "data dredging" or capitalizing on chance.

Consider Kramer's pioneering work on the effect of economic conditions on American election outcomes. Kramer's model fit aggregate vote results for congressional elections from 1896 to 1964 with a standard error of about 3 percentage points (Kramer 1971, 140). Post-sample forecasts for the period from 1966 to 1980 by Atesoglu and Congleton (1982) produced a root mean squared forecasting error of about 4 percentage points (omitting the anomalous Watergate election of 1974—about 6 percentage points for the whole period). Atesoglu and Congleton (1982, 873, 875) concluded that "Kramer's equations have relatively good post- sample predictive ability" and discounted the allegation of "data mining" made by at least one of Kramer's critics.

Political scientists unused to validating their models with independent data may be surprised that a "relatively good" result would produce out-of-sample prediction errors on the order of 30 to 100 percent larger than those produced in the original sample. What should be more surprising—and embarrassing—is how few important statistical relationships in political science have ever been validated, even to this degree of accuracy, on samples other than those used to generate them in the first place. How many researchers, given the availability of new data, would be willing to put cash on the line in support of the confidence intervals published with their parameter estimates? For the most part, we have either preferred to delude ourselves with pseudo-success rather than acknowledge the extent of our uncertainty about important political processes, or else embraced that uncertainty with unscientific enthusiasm by treating statistical analyses as descriptions of particular samples—as case studies without the charm.

4: Sensitivity Analysis

Given the prevailing degree of uncertainty about model specification in typical political science problems, it is unreasonable to expect even very large amounts of data to resolve that uncertainty entirely. Fortunately, we typically care not about the precise specification of a statistical model, but about a few of the quantitative results—key parameter estimates— produced by that model. One very useful way to explore the robustness of key parameter estimates is to estimate various alternative models and compare the results they produce. When several plausible specifications produce similar results, we can have considerably more confidence in the validity of those results than when apparently minor changes in model specification produce very different conclusions.

No competent empirical researcher would be so oblivious to this fact as to run a single regression and trust the results it produced. Nevertheless, a strong tendency persists to cover our tracks when we describe our analyses to the outside world. Often a single regression is reported as though it was the only one run. If any sensitivity analysis is reported, it tends to appear in a footnote to the effect that "I tried something different and the results came out the same." It is striking how seldom, in footnotes of this sort, trying something different makes the results come out different.

A better approach would be to keep trying something different until the results do change significantly; then we would get some feel not only for the robustness of the anal-

ysis but also for the specific dimensions of certainty and uncertainty in the data at hand. It is natural to be embarrassed when a plausible change in model specification erases or reverses a hard-won empirical result. But that sort of embarrassment is a fruitful indicator of what we don't yet know—and thus, an important source of further scientific progress. We can only hope to reach the level of scientific maturity at which the misleading certainty of a single, implicitly unproblematic model specification is even more embarrassing.

5: Using Prior Information

To learn from data we need to begin with a perspective, techniques, and assumptions. In one way or another, any statistical analysis must depend at every turn on information beyond that provided directly by the data. But what kind of information, and how?

Classical statistical theory is based on the rather curious assumption that we know everything about some parts of our problem (the model specification) and nothing at all about other parts (the parameter values). If you were really satisfied with that assumption you would have stopped reading this little essay long ago, because the issues of model specification addressed here would never arise; you would choose a model, run a regression, and report the results. But here you are.

The Bayesian theory of statistical inference is much more flexible and self-conscious about incorporating prior information. (A useful comparative analysis of the role of prior information in Bayesian and alternative approaches to statistical inference is provided by Barnett (1982).) Personally, I do not see how anyone familiar with the Bayesian approach could prefer to think in classical statistical terms. Nevertheless, Bayesian theory cannot provide any real solution to our problems either. So here I am, too.

In Bayesian statistics, the import of data is represented by a mapping from subjective prior beliefs (characterized as probability distributions over a set of unknown parameters) into corresponding posterior beliefs. But it is not enough to report any single posterior distribution, since readers will not agree upon any single subjective prior belief. The difficulty of conveying any general mapping of prior beliefs into posterior beliefs is an important practical problem that has limited the impact of Bayesian techniques in actual research.

It is possible to ignore this difficulty by simply choosing a single prior distribution and reporting the corresponding posterior beliefs. At best, this approach shortchanges the inferential implications of the data in much the same way as when conventional sensitivity analyses are curtailed for lack of time or space. At worst, the use of a single prior distribution may amount to active abuse of the Bayesian approach, if a prior distribution is chosen simply to "improve" the apparent precision of an otherwise useless data analysis. (Consumers should be especially suspicious when the prior distribution is introduced furtively in a footnote or appendix rather than being described and defended forthrightly.)

Although a thoroughgoing Bayesian approach can easily be criticized as unwieldy, its more pressing limitation in the present context is less practical than existential. Bayesian theory offers no solution to the really intractable problems of model specification simply because it is not itself a source of prior information. Methodology by itself cannot make something out of nothing.

The real value of the Bayesian perspective, in my view, is that it provides an unusually rigorous system for keeping track of what we know and what we don't know. By policing our uncertainty, it can prevent us from pretending that we know more than we really do. Indeed, in the broadest, most sophisticated (and most entertaining) treatment of model specification issues in the econometric literature, Leamer (1978) has suggested that many familiar ad hoc approaches to model specification can be interpreted from a Bayesian perspective either as (essentially legitimate) approximations to full-scale Bayesian analysis or as (essentially illegitimate) post hoc fiddling with prior beliefs to produce pleasing posterior results.

In addition to its value as a conceptual framework, the Bayesian perspective seems to offer some as yet untapped value as a concrete tool in practical problems of model specification—even those for which a fully specified Bayesian analysis would be unwieldy. For example, in what amounts to an especially systematic version of sensitivity testing, incomplete prior information can be used to map out a range of parameter values supported by the data (Leamer 1978, 182-187).

Authors of methods texts typically exhort their readers to "incorporate relevant theory" or to "model the process that generated the data." But we haven't learned how to pull back the curtain and see the little man working the levers that generate our data. And even at Rochester, nobody really believes that theory is going to come to our rescue any time soon. Economists have a lot of theory, by social science standards, but still run such a substantial deficit in degrees of freedom that one wonders, "Can our profession use data to make progress?" (Leamer 1983, 286). His answer (1983, 325) may be worth reporting here:

> There is little question that the absence of completely defined models impinges seriously on the usefulness of data in economics. On pessimistic days I doubt that economists have learned anything from the mountains of computer printouts that fill their offices. On especially pessimistic days, I doubt that they ever will. But there are optimistic days as well.

We have all had our good and bad days. But it is a significant intellectual embarrassment that we can account

for the pessimistic days more successfully than for the optimistic ones. The practical problems of model specification are obviously intractable from the perspective of any rigorous theory of statistical inference; but as data analysts we sometimes seem to be making progress nonetheless. Perhaps we had best not push too hard on this gap between theory and practice, but simply ask ourselves: how can we make more progress than we have so far?

For one thing, I believe we can learn to make better use of whatever prior information we do have, and to understand better the inferential implications both of the prior information we have and of the prior information we lack. How important must a variable be for us to want to include it in our model? How endogenous must it be for us to want to replace it with an instrumental variable? What happens to our parameter estimates if we use an instrumental variable that is itself (at least slightly) endogenous? Elsewhere I have suggested some rules of thumb for managing these sorts of inferential dilemmas by choosing among simple alternative misspecified models on the basis of particular, identifiable bits of prior information (Bartels 1985, 1991). Obviously, such rules of thumb do not obviate the need for prior information; but they may be helpful in pinpointing the inferential implications of particular sorts of prior information in a given situation, and in working back and forth between data and assumptions in the search for useful conclusions. In that sense they fall somewhere between the thoroughgoing but inordinately complex incorporation of prior information in the formal Bayesian approach and the ad hoc, unselfconscious practice common in empirical research.

Journal editors could help by reserving more space for reporting sensitivity analyses, and for replications—successful or unsuccessful—of previous analyses using new data. Current editorial practices too often encourage researchers, individually and collectively, to enshrine anomalous "statistically significant" results (Weimer 1986).

Finally, those of us who teach political methodology could help by spending more time emphasizing practical problems of statistical inference with uncertain and approximate model specifications. The specification experiments reported by Green elsewhere in this newsletter may be depressing, but we should take some heart from the fact that what and how we teach does seem to make a significant difference in the way our students negotiate the pitfalls of practical data analysis. To that extent there is at least some hope for progress.

References

Achen, Christopher (1982). *Interpreting and Using Regression.* Beverly Hills: Sage Publications.

Atesoglu, H. Sonmez, and Roger Congleton (1982). "Economic Conditions and National Elections, Post-Sample Forecasts of the Kramer Equations." *American Political Science Review* 76:873-875.

Barnett, Vic (1982). *Comparative Statistical Inference,* Second Edition. Chichester: John Wiley & Sons.

Bartels, Larry M. (1985). "Alternative Misspecifications in Simultaneous-Equation Models." *Political Methodology* 11:181-199.

Bartels, Larry M. (1991). "Instrumental and 'Quasi-Instrumental' Variables." *American Journal of Political Science* (forthcoming).

Beck, Nathaniel (1985). "Estimating Dynamic Models is Not Merely a Matter of Technique." *Political Methodology* 11:71-89.

Box, G. E. P., and G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day.

Erbring, Lutz (1989). "Individuals Writ Large: An Epilogue on the 'Ecological Fallacy'." *Political Analysis* 1:235-269.

Freeman, John R., John T. Williams, and Tse-min Lin (1989). "Vector Autoregression and the Study of Politics." *American Journal of Political Science* 33:842-877.

Hanushek, Eric A., and John E. Jackson (1977). *Statistical Methods for Social Scientists.* New York: Academic Press.

Hanushek, Eric A., John E. Jackson, and J. F. Kain (1974). "Model Specification, Use of Aggregate Data, and the Ecological Correlation Fallacy." *Political Methodology* 1:87-106.

Judge, George G., and M. E. Bock (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics.* Amsterdam: North-Holland.

Judge, George G., W. E. Griffiths, R. Carter Hill, Helmut Lutkepohl, and Tsoung-Chao Lee (1985). *The Theory and Practice of Econometrics,* Second Edition. New York: John Wiley and Sons.

Leamer, Edward E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: Wiley.

Leamer, Edward E. (1983). "Model Choice and Specification Analysis." Chapter 5 in Zvi Griliches and Michael D. Intriligator, eds., *Handbook of Econometrics,* Volume 1. Amsterdam: North- Holland.

Sims, Christopher A. (1988). "Uncertainty Across Models." *American Economic Review* 78:163-167.

Theil, Henri (1971). *Principles of Econometrics.* New York: John Wiley & Sons.

Weimer, David L. (1986). "Collective Delusion in the Social Sciences: Publishing Incentives for Empirical Abuse." *Policy Studies Review* 5:705-708.

On the Value of Not Teaching Students to Be Dangerous

Donald P. Green
Yale University

Is there a relationship between what graduate students hear in their statistics courses and how they go about specifying regression models? I recently conducted a study patterned after a small-scale experiment conducted by Larry Bartels at the University of Rochester. Bartels asked Ph.D. and M.A. students in political science to develop a regression model and estimate the parameters using data he had simulated for the assignment. Each student's data set was generated by the same underlying parameters, but the disturbances varied randomly from one data set to the next. Bartels reported at the 1985 Political Methodology meeting that the empirically derived MSE for any particular parameter estimate far exceeded the MSE that would be expected had each student estimated the true model.

The present study differs from the Bartels study in three ways. First, Bartels had roughly 40 participants in his study, and the small number of cases makes it difficult to identify clear patterns in the way in which students misspecified their models. The present study overcomes this problem by using 199 subjects. Second, Bartels offered students a great deal of rope with which to hang themselves. Students were given a data set containing 6 potential independent variables, three of which actually influenced the dependent variable. The present study offers students only three potential independent variables, two of which are actual causes. Finally, the present study extends Bartels' analysis by examining the effect of course content on the empirically derived MSE.

In brief, we find that the observed variability of the parameter estimates greatly exceeds what one would expect based on sampling error alone because misspecified models far outnumber correctly specified ones. In addition, we find that in comparison to students who were warned against atheoretical model specification, those whose training in regression emphasized classical F-tests and goodness-of-fit were more likely to misspecify their models by including endogenous regressors and excluding causally relevant, but statistically insignificant variables.

Research Design

Subjects. Participants in the study were first year students in the Yale School of Management who were enrolled in a core course in probability and regression analysis during fall semester of 1989. The course was required of first year students, who were randomly assigned to one of three meeting times and were prevented from transferring from one class to another. All students were instructed by one of two faculty members (hereafter designated Professor A and Professor B): one-third attended Professor A's afternoon class; the remainder attended one of Professor B's two morning classes.

Task. During the final week of classes, all students were required to complete a problem set involving regression analysis. Each student was given a unique data set and instructed to work independently. After the problem sets were handed in, selective checks were made to assess whether pairs of students known to study together had produced similar answers; no evidence of collaboration was detected. Neither instructor discussed the problem set with students until after it was handed in.

The assignment asked each student to "develop a regression model of charitable contributions" made by chemical corporations. Unbeknownst to students, the data were fictitious. Students were presented with four vectors of data, (contributions, profits, age of firm, and chamber of commerce ratings of each firm's service to the community) three of which represented potential regressors.[1] An attempt was made in the introductory paragraph to portray the chamber of commerce ratings as potentially endogenous. Hints that the ratings might be the consequence rather than the cause of contributions included (1) the fact that the ratings were compiled during the same year as the contributions and (2) the fact that the ratings were based on service to the community, which presumably included charitable donations.

Had the objective been prediction of contributions, the endogeneity of the chamber of commerce ratings would have been moot. However, students were told explicitly that the objective was to "understand the factors which determine philanthropic contributions," not merely to predict contributions. The principal objective, therefore, was parameter estimation, not maximization of R^2.

Specification of the model. Each student was given his/her own data set of 25 observations. All data sets contained the same values for two variables, profits and longevity. The variance of PROFITS was 810; the variance of AGE was 3039; the correlation between PROFITS and AGE, .26.

Charitable contributions and chamber of commerce ratings were generated by the following formulae:

$$\text{CONTRIBUTIONS} = 15 + .15\ \text{PROFITS} + .01\ \text{AGE} + u_1$$

$$\text{CHAMBER RATINGS} = 50 + 1.0\ \text{CONTRIBUTIONS} + u_2$$

The values for u_1 and u_2, the disturbance terms, were drawn at random from a normal distribution with zero mean and variance 16. From these specified values, it follows that

[1] Students were asked not to use interactions or transform the variables. These instructions allowed us to maintain the comparability of the results. Presumably, had we allowed students to be creative in their treatment of the regressors, we would have observed even wider discrepancies between the true model and the student's OLS estimates.

regressions using PROFITS and AGE as independent variables would be expected to produce statistically significant estimates ($p < .05$) 99% of the time for PROFITS but only 20% of the time for AGE.

Note that a constant of 50 was added to the ratings variable so that if students added RATINGS to the model they would have at least one direct indication that something was wrong. With all three regressors in the equation, the estimated intercept becomes negative, which makes no sense substantively.

The premise of the simulation is relatively simple. Tempt students (1) to include an endogenous variable, RATINGS, which would greatly enhance their R^2 and (2) to dump a regressor of marginal significance, AGE, so as to improve the value of the F or the adjusted R^2. The first represents a clear violation of the assumptions upon which the unbiasedness of the OLS procedure rests. The second sort of mistake is less reprehensible and, given the relatively low correlation between AGE and PROFITS, less problematic. While it seems clear that AGE is exogenous (it strains credulity to think that chemical firms die off because of tightfistedness), the theoretical basis for suspecting age to be a cause of charity is a bit thin. Doubtless, a certain number of students would have sound theoretical reasons for omitting AGE. Others, however, would run a series of models and then elect to drop AGE when it proved to be statistically insignificant — a procedure which alters the sampling distribution of the parameter estimates.

Response rate. Although the assignment was required of all 211 students, only 204 turned in solutions. Of these, 199 gave answers that were usable for the present analysis. Two were illegible; one used the natural logarithm of PROFITS; two others modeled a dependent variable other than charitable contributions.

Results

For the sake of brevity, let us focus on the estimated coefficient for PROFITS, which we will refer to as b_1. If the model is correctly specified (with PROFITS and AGE as predictors), then b_1 is unbiased, and its MSE is .00085. Table 1 suggests that more often than not, students ran something other than the true model. The specifications that students came up with suggest that a heavy emphasis was placed on achieving a high adjusted R^2, as 43.2% chose to put PROFITS and RATINGS but not AGE in the equation. This is the specification one would expect if students fished for a high R^2 and dropped regressors that failed to achieve statistical significance. Another 13.5% either put all of the independent variables available to them in the equation, used only the RATINGS score, or used AGE and RATINGS but not PROFITS. These approaches, too, seem to have been aimed at maximizing the significance of the F value and/or adjusted R^2.

Table 1: Students' Choice of Model Specification

Model Specification	Percent Choosing Specification
PROFITS, AGE	14.1
PROFITS	29.1
PROFITS, AGE, RATINGS	5.5
PROFITS, RATINGS	43.2
RATINGS	6.5
RATINGS, AGE	1.5
AGE	0.0
Number of cases	199

Distribution of Regression Coefficient for Profits

Mean	.106075
SD	.052633
MSE	.004700

A fair number of students (29.1%) elected to include only PROFITS in the model. For many, no doubt, this was a decision based on the statistical insignificance of the AGE coefficient. Some students, however, justified their decision on theoretical grounds, arguing for example that younger and older firms face roughly the same incentives when deciding whether to make a good impression on the community. Thus, while some students may have fished for this specification, others did show some restraint in leaving RATINGS out of the model. Finally, 14.1% of the students ran the "true" model, regressing CONTRIBUTIONS on PROFITS and AGE.

Given that only 1 out of 7 students ran the true model it comes as no surprise that the MSE of the students' b_1 estimates (.0047) far exceeds the analytic MSE we derived earlier (.00085). As Table 1 indicates, student estimates vary more than random sampling error would predict and tend to be biased downward.

A Quasi-Experiment

What effect, if any, does the instructor have on the way students model charitable contributions? Until shortly after the midterm examination, the three sections of the course were taught virtually identically. Both instructors followed the same syllabus, assigned the same readings[2] and problem sets, and told the same jokes. On the midterm and final exams, the two sets of students performed similarly, with Professor A's section scoring slightly below the two sections taught by Professor B.

[2]The textbook for the course was Morris Hamburg's *Statistical Analysis for Decision Making* (4th Edition), published by Harcourt, Brace, and Jovanovich.

Table 2: Choice of Model Specification by Instructor

	Percent Choosing Specification by Instructor	
Model Specification	Professor A	Professor B
PROFITS, AGE	24.2	9.5
PROFITS	37.1	26.5
PROFITS, AGE, RATINGS	12.9	2.2
PROFITS, RATINGS	24.2	51.8
RATINGS	0.0	9.5
RATINGS, AGE	1.6	1.5
AGE	0.0	0.0
Number of cases	62	137

Distribution of Regression Coefficient for PROFITS, by Instructor		
	Professor A	Professor B
mean	.121613	.099044
SD	.045991	.054079
MSE	.002921	.005521

Following the midterm, however, the three sections were taught somewhat different approaches to regression analysis. Although both courses covered the same computational basics, Professor B's sections received instruction on F-testing and fit statistics, while Professor A made no mention of F-tests, stressed the limitations of R^2 as a criterion by which to choose among model specifications, and focused primarily on the conditions under which OLS yields biased estimates. Professor B, in short, instructed students to use a combination of theory-driven and data-driven approaches to model specification; Professor A gave only a cursory overview of regression diagnostics and advised students to evaluate competing specifications on theoretical grounds.

Table 2 suggests that differences in teaching emphasis manifest themselves in students' answers. Fully 61.3% of Professor A's students resisted the temptation to include an endogenous regressor, as compared to 35% of Professors B's students. Similarly, just 24.2% of Professor A's students fell for both traps – dropping AGE and including RATINGS – as opposed to 61.3% of Professor B's students. In sum, while neither set of students reproduced the analytic standard error, students in Professor A's class produced estimates which were considerably less biased and less variable.

Conclusion

However commonplace it may be for methodologists to warn that maximizing R^2 is an untrustworthy way to specify structural models, the practice is pervasive in social science and policy studies. Equally prevalent and no less pernicious is the tendency to drop variables that are deemed "statistically insignificant." The present research underscores the idea that such mechanical approaches to regression analysis may produce biased and unreliable parameter estimates. With just a small handful of independent variables from which to choose, students in our study very often managed to select the least optimal combination of regressors. All indications point to preoccupation with goodness-of-fit as the culprit.

Although students in our study tended to "data-dredge," our results suggest that data-dredgers are made, not born. Those who received little instruction on the topic of fit statistics were much less likely to choose their independent variables in an atheoretical manner. By contrast, Professor B's students, who were told that one should specify models in a theoretically sensible fashion but who also heard a good deal about classical F-tests and goodness-of-fit, were quick to abandon common sense when the opportunity arose. Evidently, when students are presented with both deductive and inductive approaches to model specification, they tend to gravitate toward the latter, perhaps because there are clearly defined rules and procedures. They then leave graduate school and go on to produce journal articles and policy reports.

(The GAUSS 2.0 code used to generate the problem set discussed in this piece is available from the author.)

When to Use R-Squared

Michael S. Lewis-Beck
University of Iowa

Andrew Skalaban
University of California, Davis

Some political science methodologists find the R-squared statistic of little utility. Others, in contrast, find it of great use. This debate is ongoing, and can be traced by consulting the current exchange in *Political Analysis* (see Lewis-Beck and Skalaban, forthcoming, and the responses by Achen and King, same issue). In the few pages at hand, we state the case for R-squared as directly as possible. Generally speaking, when the researcher wishes to evaluate model performance, the R-squared is indispensable.

Below, we explore this point first with the example of the single model. Then, we look at the simple two model comparison (within the same sample). Finally, we examine the more complex two model comparison (across samples from different populations). Throughout, the assumption is that the researcher has formulated a single-equation regression model, met the regression assumptions, and calculated ordinary least squares estimates from an appropriate probability sample.

Scenario I: The Single Model

In this situation, which is by far the most common, the researcher seeks to evaluate the performance of a multiple regression model, such as

$$\begin{aligned} \widehat{Y} &= b_0 + b_1X_1 + b_2X_2 + b_3X_3 \qquad (1) \\ & (\mathrm{SE}b_1),\ (\mathrm{SE}b_2),\ (\mathrm{SE}b_3) \\ & R^2 = \ldots\ \mathrm{SEE} = \ldots\ \mathrm{N} = \ldots \end{aligned}$$

where $\widehat{Y}$ is the dependent variable (predicted); X_1, X_2 and X_3 are independent variables (whose values were observed in the sample); b_0 is the estimated population intercept; b_1, b_2, b_3 are the estimated population partial slopes; $\mathrm{SE}b_i$ is the standard error of estimate for a partial slope; R^2 is the coefficient of multiple determination; SEE is the standard error of estimate of Y; and N is the sample size.

For any meaningful performance evaluation, a baseline is necessary. The R-squared provides two: linearity and proportional reduction in prediction error. To what extent can the relationship be described as linear? Perfectly so, if R-squared = 1.0. Not at all, if R-squared = .00. Of course, linearity is not the only possible baseline form. However, it often holds in practice, and in any case can serve as a convenient point of departure.

The R-squared also tells how well the model predicts, relative to how much there is to predict. Recall the formula,

$$R^2 = \frac{\sum(\widehat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2}$$

or

$$\frac{\mathrm{RSS}}{\mathrm{TSS}} = \frac{\text{error accounted for by regression}}{\text{error to be accounted for}}$$

where $\widehat{Y}_i$ is the predicted Y, $\overline{Y}$ is the mean of Y; Y_i is the observed Y; RSS is the regression sum of squares; and TSS is the total sum of squares.

The denominator represents the amount of prediction error that would result, not knowing the X values (and knowing only the mean of Y). The numerator represents the reduction in this prediction error, knowing the X values. The larger the numerator relative to the denominator, the closer the fraction moves to 1.0, and the better model performance appears. Say R-squared = .92. Then, the independent variables together account for 92 percent of the prediction error. In other words, it accounts for all but eight percent of the variation in Y (around the mean as the predicted value). Thus, the higher the R-squared, the greater the relative predictive capability of the model.

Scenario II: The Two Model Comparison (Within Sample)

The researcher may want to compare the performance of two (or more) models, estimated on the same sample. Suppose the following model is presented as a challenge to Eq. 1:

$$\widehat{Y} = a_0 + a_1X_1 + a_2X_2 \qquad (2)$$

where the definitions follow Eq. 1.

In this case, the researcher must now evaluate two fit statistics, R-squared from Eq. 1 and R-squared from Eq. 2. The task is straightforward, essentially involving an F-test to see if the two values differ significantly (see Kmenta, 1971, p.371). Other procedures might be explored as well; i.e., comparison of adjusted R-squared, or application of a "substantive" significance test. These various comparisons all help the analyst decide whether the model with the larger R-squared does or does not represent the preferred specification.

Scenario III: The Two Model Comparison (Across Samples)

Researchers may want to compare two (or more) models across samples taken from different populations; e.g., nations, states, counties, organizations. This goal, rare in certain subfields of political science, is common in comparative politics. Suppose the student of Western European elections estimates Eq. 1 on a probability sample of French voters and on a probability sample of British voters. The student asks, "Which model performs better, the French or the British?" A careful comparison of the two R-squared statistics may answer this question.

Suppose the French R-squared = .70 and the British R-squared = .45. Then, the conclusion that the model predicts relatively better in France suggests itself. Such a conclusion remains unqualified, provided the estimated variances on the Xs and/or Y are not meaningfully different. (That is, for example, the estimated standard deviation of, say the independent variable of education, is about the same in France and in Britain). If meaningful variance differences do exist, then the conclusion must be somewhat qualified, as follows: the model predicts relatively better in France than in Britain, given the prevailing dispersion of the variables. A qualification of this type is actually informative, alerting the reader to important cross-national differences. (Moreover, it may be unavoidable, since leading alternative measures are also variance sensitive; i.e., the standard error of the slope responds to variance change in X, and the standard error of estimate of Y responds to variance changes in Y.)

Conclusion

The R-squared, a linear measure of relative predictive capability, is a valuable tool for evaluating the performance of single-equation regression models. The higher the R-squared, the greater the predictive capability of the model relative to the total variation in the dependent variable.

References

Kmenta, Jan. 1971. *Elements of Econometrics.* New York: Macmillian.

Lewis-Beck, Michael S., and Andrew Skalaban. Forthcoming. "The R-squared: Some Straight Talk." *Political Analysis*, Vol. 2, 1990.

When *Not* to Use R-Squared

Gary King
Harvard University

If you are asking political questions in a scientific way (one of the goals of "political science"), then R^2 will not provide any relevant answers. The R^2 statistic contains no useful information beyond that in the regression coefficients, standard errors, and the standard error of the regression. Moreover, the form in which the information exists in R^2 is much less natural than that in these other statistics.

I made these points originally in the *American Journal of Political Science* in 1986 and more recently and in more detail in the forthcoming issue of *Political Analysis*. In this brief note, I follow the structure of Mike Lewis-Beck and Andy Skalaban's article in this newsletter, and continue our discussion.

For clarity, here is a version of the model we are all talking about:

$$Y_i \sim f_n(y_i|\mu_i, \sigma^2).$$

where Y_i and Y_j are independent given μ for all $i \neq j$.[1] In addition, Lewis-Beck and Skalaban's Equation (1) is an estimate of this model with

$$\mu_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3,$$

and their Equation (2) is an estimate of this Normal model with

$$\mu_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

Scenario I: The Single Model

Under their single model scenario, Lewis-Beck and Skalaban argue that R^2 provides two pieces of information about this model and its estimates: (1) "model performance" and (2) "proportional reduction in prediction error."

(1) I begin with Lewis-Beck and Skalaban's *assumption* (their word, my emphasis) "that the researcher has formulated a single-equation regression model, met the regression assumptions, and calculated ordinary least squares estimates from an appropriate probability sample." This is a perfectly reasonable approach for many situations, and these assumptions are very frequently made in practice. However, in this situation, "model performance" need not be evaluated since the model is correct *by assumption.*

If we allow for the possibility that the model is incorrect, then we need specific tests for problems with the model such as non-Normality, autocorrelation, heteroskedasticity, nonlinearity, omitted variable bias, selection bias, endogeneity, or measurement error. Providing evidence for the existence of any one of these problems would mean that the original model is incorrect and some other model should be used. It thus makes considerable sense to conduct some of the numerous available tests for each. However, R^2 is not a test of any of these problems. If R^2 is high, the model might be right or wrong; if R^2 is low, the model might be right or wrong. Obversely, if the model is exactly right R^2 might be zero, or one, or any number in between. Model performance and the value of the R^2 statistic are independent questions.

(2) If R^2 is high, will the model predict new observations better than if R^2 is low? Sometimes yes, sometimes no; these too are separate questions. Indeed, it seems from considerable research in econometrics that models which overfit the data are especially bad at prediction. (It was no small embarrassment when large-scale econometric models with hundreds of parameters, and thus huge R^2 values, that took years to construct and days to estimate were found to predict worse than fairly atheoretical ARIMA models with only one or two parameters.) Thus, R^2 has no necessary relationship to accurate prediction; if anything, practice seems to indicate that *lower* values of R^2 produces *better* predictions.

Finally, If you wish to know how well your regression function maps the points in your one sample, R^2 will tell you this, but $\hat{\sigma}$ will be clearer since it expresses the answer on a scale you have created—that of your dependent variable.

Scenario II: The Two Model Comparison (within Sample)

The comparison between Equations (1) and (2) comes down to a test of whether β_3 is zero. The appropriate way to do this formally is with a t-test of that coefficient (or an F-test, which is just the square of the t in this case), along with a judgment about how substantively important this additional variable is. Although the calculations going into R^2 are used in the formal test, hypothesis tests can only be conducted on parameters, not descriptive statistics like R^2. More informally, whenever you have the same dependent variable in the equations you are comparing, as you do here, it is perfectly reasonable to look at the change in the R^2 statistic. However, if you wish the answer in substantive terms, you need not look farther than the parameter and its standard error.

[1] The assumption of Normality is not required for parameter estimates to be unbiased or consistent in OLS, but OLS will be maximally efficient in this situation. Also, the assumption of independence may be relaxed to uncorrelatedness. I retain the strong version of both for clarity. Any disagreement we might have does not depend on the specific version of these assumptions.

Thus, although you *can* reasonably use R^2 to compare two models that are estimated from the same dependent variable, *statistical significance* is better ascertained with reference to a formal hypothesis test and *substantive importance* is better judged by interpreting the coefficient and its standard error.

Scenario III: The Two Model Comparison (Across Samples)

Lewis-Beck and Skalaban write "The student asks 'Which model performs better, the French or the British?' A careful comparison of the two R-Squared statistics may answer this question." They qualify this considerably in the current version of their argument by focusing on changes in the variance in the explanatory variables, which, as with standardized regression coefficients, can seriously distort model comparisons.

Unfortunately, even with careful variance comparisons, R^2 cannot help determine which model performs better, since we have already established that high or low R^2 values have nothing whatsoever to do with correct model specification or with prediction. In their example with the French R^2 equaling 0.7 and the British equaling 0.45, it may actually be that the British model is better: For example, suppose one of the variables in the model estimated in both countries is irrelevant in France but relevant in Britain. In France, the variable would have a small coefficient, but would still increase the R^2 some due to random variation in the data. Since this variable is just mapping random error in the French data, the French model would actually make *worse* out-of-sample predictions than the British model. Or, even more simply, suppose that the model was applied to France just before the Fifth Republic. In that case, the model might fit very well but have little to do with the future, whereas the same model might forecast quite well in Britain. Higher R^2 values say nothing about whether the current model will apply in the future.

Finally, suppose we have only a regression with a constant term and no explanatory variables. In this case, the parameter estimate will be the mean, and R^2 will be zero. Using R^2 for "model performance" would seem to indicate that we should never be using the mean as a model.

Conclusion

In general, if you have a substantive question to be answered by quantitative analysis with a linear model, the best answer will invariably be found in the coefficient estimates, their standard errors, or the standard error of the regression. R^2 is not usually an answer to a question that a substantively-oriented political scientist should be interested in.

Choose the least statistically tolerant (but empirical) member of your political science department, and try to describe your research results to his or her satisfaction without any statistical language. This is sometimes hard for us, but it is essential in a discipline like political science since substantive concerns should and will dominate statistical questions every time. Fortunately, in the case of the regression model, communicating the substance of an argument is easy: for example, if you have an additional year of education (prior to graduate school!), your expected starting salary will increase by $2000, plus or minus about $400, controlling for the other variables in the equation. Or, if you are a white male who has twelve years of education, rich parents, but did poorly in school, your expected starting salary will be $17,000 per year, plus or minus about $5,000. For learning about substance of politics in substantive language, R^2 is unnecessary and too easily misleading.

When to use R^2? Scenario II, if careful. When *not* to use R^2? Scenario I and III.

References

King, Gary. Forthcoming. "Stochastic Variation: A Comment on Lewis-Beck and Skalaban's 'The R-Squared'," *Political Analysis*, Vol. 2, 1990.

King, Gary. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science," *American Journal of Political Science*, 30, 3 (August): 666–687.

Lewis-Beck, Michael S., and Andrew Skalaban. Forthcoming. "The R-squared: Some Straight Talk." *Political Analysis*, Vol. 2, 1990.

Dimensional Analysis in Political Science

William G. Jacoby
University of South Carolina

The term "dimensional analysis" refers to a variety of research strategies and procedures that are intended to provide quantitative and/or geometric representations of data. There are three main reasons for using a dimensional analysis approach to a research problem: (1) Simple data reduction— summarizing a large set of variables with a smaller number of composite measures; (2) examining dimensionality— testing the underlying sources of variation in a dataset; and (3) measurement— obtaining empirical representations of the underlying (and usually unobservable) dimensions, which can be employed as analytic variables in other statistical procedures. All three of the preceding reasons are important for political scientists. And yet, these techniques are not used very frequently in the discipline. Why is this the case?

I believe there are several reasons for the relative paucity of dimensional analysis applications in political science. First, some analysts express concern over the fact that many dimensional analysis models are inherently underidentified— i.e. the rotation problem. Second, many dimensional analysis techniques are entirely descriptive, rather than inferential, strategies. Third, there are some practical difficulties which, although less serious than the preceding problems, still effectively limit the applications of dimensional analysis in political science. For example, advances in the field are reported in specialized journals (e.g. *Psychometrica*); the software required for many of the applications is not widely available; and dimensional analysis is often simply omitted from typical course syllabi in quantitative methodology. Thus, there are problems involved in simply disseminating information about dimensional analysis to political scientists.

But, I contend that all of these obstacles can be overcome. For one thing, I believe that some of the technical objections mentioned earlier are a bit overstated (at least in certain contexts). The rotation problem is really only a major concern when we focus on the point coordinates in a scaling solution. However, that is only one aspect of the results from a factor analysis or a multidimensional scaling (MDS) analysis; it is also important to pay attention to the overall structure of the scaled configuration. At the same time, the lack of inferential strategies in scaling procedures is not a fatal drawback in many cases, precisely because dimensional analysis is often used for inherently exploratory purposes. There are simply many applications where it is more important to search for structure in data, than it is to impose a detailed set of statistical assumptions and evaluate their fit to the data. Furthermore, the development of confirmatory factor analysis strategies, along with maximum likelihood methods for MDS and other scaling models (e.g. Rasch models for cumulative scales) guarantees that inferential uses of dimensional analysis will be more widespread in the future.

The practical difficulties are also being dealt with slowly but (I believe) surely. Perhaps most important, there are now many accessible sources of information on dimensional analysis, ranging from texts for beginners (e.g. the volumes on scaling in the Sage University Papers: Kim and Mueller 1978a; 1978b; Kruskal and Wish 1978; McIver and Carmines 1981; Arabie, Carroll, and DeSarbo 1987) to more advanced treatments of the theory and methods involved in dimensional analysis, such as the volume on factor analysis by Joreskog and Sorbom (1979), and the recent work on multidimensional scaling by Young and Hamer (1987). In a related development, scaling procedures are now available in the commonly used statistical software packages (e.g. both SPSS and SAS contain ALSCAL for MDS analyses, along with other, related routines) and in several varieties of software for PCs (e.g. SYSTAT contains routines for both factor analysis and multidimensional scaling, and an organization called IEC-ProGamma from the University of Groningen in the Netherlands markets a variety of PC programs for scaling analyses. (The address for IEC-ProGamma is: IEC-ProGamma, Kraneweg 8, 9718 JP, Groningen, The Netherlands.)

There are other reasons for optimism as well, because the field of dimensional analysis is progressing in a variety of different ways. And, political scientists are making direct, important contributions to this development. For example, James Enelow and Melvin Hinich (1984), Henry Brady (1989; 1990) and Keith Poole (1984; 1990), proceeding in very different directions, have all made important advances in the estimation of the ideal points (or unfolding) model. Brady has also made seminal contributions to the problem of statistical inference in nonmetric multidimensional scaling (1985). George Rabinowitz (1976) has developed a similarity measure that can be obtained from rating scale responses; this opens up MDS to a wide variety of potential applications, where direct measures of interobject proximity are not available. Finally, van Schuur (1984) and his colleagues in the Netherlands have made important advances in their work on stochastic unfolding and cumulative scales (e.g. Niemoller and van Schuur 1983). Thus, the development of dimensional analysis techniques and strategies is progressing within political science. And, of course, the psychometrics literature (the "traditional home" of dimensional analysis) contains an enormous amount of further material in this field.

I anticipate future work in dimensional analysis to focus on at least three important themes. First, there will be more emphasis on fitting alternative scaling models to data. The choice between, say the factor analysis model and the MDS model is not one of convenience or levels of measurement. The fact that one scaling model fits a dataset better than another always has important substantive implications. This has been recognized by political scientists at least since Weisberg's work (1972) on legislative roll call analyses. However it deserves more attention than it has previously received. Second, it is important that we pursue the implications of measurement models for data analysis. Measurement is, itself, a formal model of observations, and measurement assumptions should be tested, rather than simply taken as fixed, given characteristics of the data. Here, the work by Young and his colleagues on optimal scaling (e.g. Young 1981), and the European advances in correspondence analysis (e.g. Gifi 1981; Greenacre 1984) provide promising directions for future work. And finally, it is important to integrate dimensional analysis models with structural equation models, thereby uniting two very different traditions in political research: approaches based upon psychometric techniques and those derived from econometrics. LISREL models are an important contribution to this integration. And constrained MDS models are another important

starting point for combining scaling solutions with external, multivariate data (e.g. Heiser and Meulman 1983). Future work will undoubtedly emphasize the simultaneous estimation of dimensional parameters along with causal parameters in empirical analyses.

In conclusion, dimensional analysis, while definitely not the predominant research strategy, is certainly alive and well in political science. Researchers have only begun to exploit its capacities. And, there are important developments currently taking place in this field. As information about the various techniques is disseminated within the political science community, dimensional analysis will undoubtedly become a useful, productive tool for empirical research.

References

Arabie, P., J.D. Carroll, W.S. DeSarbo. 1987. *Three-Way Scaling and Clustering.* Beverly Hills, CA: Sage.

Brady, H.E. 1985. "Statistical Consistency and Hypothesis Testing for Nonmetric Multidimensional Scaling." *Psychometrica* 50: 509-537.

Brady, H.E. 1989. "Factor and Ideal Point Analyses for Interpersonally Incomparable Data." *Psychometrica* 54: 181-202.

Brady, H.E. 1990. "Sense and Nonsense in Multidimensional Scaling." Paper presented at the 1990 Annual Meetings of the American Political Science Association, San Francisco, CA.

Enelow, J.M. and M.J. Hinich. 1984. *The Spatial Theory of Voting.* New York: Cambridge University Press.

Gifi, Albert. 1981. *Nonlinear Multivariate Analysis.* Leiden: University of Leiden Press.

Greenacre, M.J. 1984. *Theory and Applications of Correspondence Analysis.* London: Academic Press.

Heiser, W.J. and J. Meulman. 1983. "Constrained Multidimensional Scaling, Including Confirmation." *Applied Psychological Measurement* 7: 381-404.

Joreskog, K. and D. Sorbom. 1979. *Advances in Factor Analysis and Structural Equation Models.* Cambridge, MA: Abt.

Kim, J. and C.W. Mueller. 1978a. *Introduction to Factor Analysis.* Beverly Hills, CA: Sage.

Kim, J. and C.W. Mueller. 1978b *Factor Analysis.* Beverly Hills, CA: Sage.

Kruskal, J.B. and M. Wish. 1978. *Multidimensional Scaling.* Beverly Hills, CA: Sage.

McIver, J. and E.G. Carmines. 1981. *Unidimensional Scaling.* Beverly Hills, CA: Sage.

Niemoller, K. and W.H. van Schuur. 1983. "Stochastic Models for Unidimensional Scaling." In D. McKay, N. Schofield, P. Whitely (Eds.) *Data Analysis and the Social Sciences.* London: Frances Pinter.

Poole, K.T. 1984. "Least Squares, Metric, Unidimensional Unfolding." *Psychometrica* 49: 311-323.

Poole, K.T. 1990. "Least Squares Metric, Unidimensional Scaling of Multivariate Linear Models." *Psychometrica* 55: 123-149.

Rabinowitz, G.B. 1976. "A Procedure for Ordering Object Pairs Consistent with the Multidimensional Unfolding Model." *Psychometrica* 41: 349-373.

van Schuur, W.H. 1984. *Structure in Political Beliefs.* Amsterdam: CT Press.

Weisberg, H.F. 1972. "Scaling Models for Legislative Roll-Call Analysis." *American Political Science Review* 66: 1306-1315.

Young, F.W. and R. Hamer. 1987. *Multidimensional Scaling: History, Theory, and Applications.* Hillsdale, NJ: Erlbaum.

Young, F.W. 1981. "Quantitative Analysis of Qualitative Data." *Psychometrica* 46: 357-388.

Teaching the Courses No One Wants to Take

Renée Marlin-Bennett [1]
American University

The two methodology courses required for undergraduate majors in the School of International Service at the American University are the most dreaded but, according to many of my students (anecdotal and highly biased evidence), the most useful courses. I describe here our students, their career goals, and how we view the pedagogical importance of methods courses. Our experience is in international relations, but the goals of methods courses are the same for political science.

The Students

Based on grades and SAT scores, students of international relations at the American University are the best students in the University — but they hate math. Most of these students have been assiduously avoiding anything that has to do with a number since they finished Algebra II. Somewhere along the line, most of them got the impression that their literary and historical talents negated any ability at math and science. It takes a good deal of humor and handholding to get the students over these prejudices.

Also, our students are skeptics: Why do math when the subject is supposed to be international relations? Or the

[1]My thanks to the School of International Service Methodology Task Force, whose members include Lew Howell, Rick Moore, Nanette Levinson, and John Richardson, and to our dean, Louis Goodman.

variant: I'm going to law school, so why do I need this? Our answer is three-fold. First, when we grant a degree in international relations, we imply that the student has some knowledge of the literature as a whole, not just the literature that represents research in a traditional, historical mode. Second, we use the tools of methodology to show the difference between term papers that are literature reviews and term papers that create knowledge. Third, students, as citizens, might need to understand the things we teach to interpret an increasingly quantified world in which statistics are often cited as Truth.

Pedagogical goals

Our courses in methodology are important for the educational program of our school and for providing our students with saleable job skills. Methodology courses teach discipline and rigor to the undergraduates. These are the courses in which they learn that the study of international relations is not just the study of current events, however intriguing journalistic reports may be. Specifically, we want them to learn:

1. interpretation of quantitative international relations literature,
2. conceptual and operational definitions,
3. the relationship between international relations theory and testable hypotheses,
4. primary data sources in international relations,
5. graphical representation of data to describe and summarize information,
6. elementary statistical analysis of IR data, and
7. the breadth of data and methods used by IR scholars.

Students must take one course in research design and one in methods of analysis. Many students will choose to take both courses within the School of International Service. Students must take at least Introduction to International Relations Research or Quantitative Approaches to International Politics. The second course may be taken in a cognate field. We recommend that the students take the sequence in their sophomore year so that they have had some introduction to the field of international relations as a foundation for their substantive courses.

In teaching Quantitative Approaches, I have found it essential both to make the course amusing and to emphasize that it is international relations and not math. On the humorous side, the students' introduction to how to interpret a quantitative article begins with Lee Sigelman's "Toward A Stupidity-Ugliness Theory of Democratic Electoral Debacles" (1990). This article, though obviously not on an international relations topic, is so funny that even my most disaffected students are entertained. In class, I walk them through the article, explaining what those italic p's are, etc.

To relate what they have learned to IR, the students write a critique of a quantitative international relations journal article of their choice, preferably one which will help them with a term paper they are working on for another IR course. Their other paper, a pilot study, should ideally also be tied to a term paper for a substantive course.

Learning to conceptualize and operationalize variables is a skill that students can use in all their courses since proper definitions preclude the fuzziness that afflicts many undergraduate papers. This is a particularly salient issue for international relations: What exactly are power, war, literacy rates?

To test hypotheses, students need to know a little about probability and frequency distributions. Humor gets the point across again. The students are, I inform them, responsible for testing missiles that the US might purchase. They must determine how accurate the missiles are by sampling. The "missiles" are paper airplanes, and we measure how close they come to a target line.

Some of these educational goals translate directly into job skills, particularly the ability to define concepts clearly, the familiarity with data sources for our field, the ability to incorporate appropriate graphics and descriptive statistics into reports, etc. Material that is more specifically job skill oriented includes:

1. familiarity with SPSSx and the more general ability to learn a software package and write and debug programs,
2. familiarity with electronic mail, and
3. ability to complete tasks and solve problems on time.

A two course sequence in international relations methods does not create professional statisticians, but it does produce educated consumers of quantitative material. And sometimes it piques their curiosity to study more.

Reference:

Sigelman, Lee. 1990. "Toward a Stupidity-Ugliness Theory of Democratic Electoral Debacles." *PS: Political Science and Politics* 23: 18-20.

Review of Fenno's *Watching Politicians: Essays on Participant Observation*

Larry M. Bartels
University of Rochester

Richard Fenno. 1990. *Watching Politicians: Essays on Participant Observation*, IGS Press, Institute of Governmental Studies, University of California at Berkeley. $9.95, 133pp.

I cannot pretend to provide a critical review of a new book by my friend and colleague Richard Fenno. However, I do want to call attention to a new Fenno work of considerable interest to political methodologists. *Watching Politicians*, published by Nelson Polsby's IGS Press, makes available five new and old Fenno essays on the pitfalls and payoffs of closely observing working politicians in their own habitats—"soaking and poking"—as a research method.

Readers familiar with the methodological appendix to *Home Style: House Members in their Districts* (1978) will realize that Fenno has always been a vastly more conscious and creative political methodologist than the illiterate mob of cookbook LISREL- crunchers who sometimes appropriate the label. In recent years his methodological preoccupations have come even more clearly to the fore, especially in his 1985 presidential address to the American Political Science Association, "Observation, Context, and Sequence." The present volume reprints these two well-known works, and adds two new essays—Fenno's 1989 Jefferson Lectures at Berkeley—in which he wrestles with issues of validity and bias raised by his more recent research on Senator Dan Quayle.

Fenno "won the lottery," as he sometimes puts it, by having a completed 250-page manuscript on Dan Quayle in the Senate sitting in his desk drawer when George Bush sprang Quayle on the world at the 1988 Republican convention. But what would he do with the winning ticket? Cash it in at the Washington Post? Add a quick chapter on Paula Parkinson and the National Guard and rush to print in time to hit the supermarket bookracks before Election Day? The first of Fenno's Jefferson Lectures addresses this dilemma, weaving it into a meditation on the boundary between scholarship and politics and on the chemistry connecting participant observers and the people they observe.

The second of Fenno's Jefferson Lectures explores a different sort of boundary, the one between political scientists and journalists. Having followed Quayle's legislative career over a six year period, Fenno was impressed with the young senator's ability to cross partisan and ideological lines to forge a successful coalition with Ted Kennedy on the consequential Job Training Partnership Act. By contrast, the pack of reporters that besieged Quayle after his burst into national prominence was fixated on old incidents dredged up from his private life, and on the fraction of comments by Washington insiders that confirmed their preconception of Quayle as a lazy rich kid, a "lightweight." Fenno concludes that "journalists will come to judgments about politicians too quickly, too superficially, and too inflexibly to fill the political scientists' need for an understanding of politicians. ...That is the bad news and the good news for political scientists who wish to do it themselves."

Much in these essays can be read as straightout advocacy of field observation as a distinctive and valuable political science research method, not only against the appearance of redundancy with journalists (the tone of the second Quayle lecture is presaged by an uncharacteristically combative turn of phrase in Fenno's presidential address to the APSA: "if my observations are interchangeable with those of a journalist, then I shall desist"), but also against the "imbalance" produced by the current dominance of secondary analysis—often but by no means always quantitative and statistical—in empirical political science.

Fenno offers some sense of the depth and individuality of his own first-hand data in "Observation, Context, and Sequence," where in nine pages he turns his six senators' apparently unproblematic votes on selling AWACS planes to Saudi Arabia into six distinctive miniature dramas of ambition, calculation, and timing. He offers more of the same in another essay that will be new to most readers, "What Are They Like?", a profile of Jim Johnson, one of the eighteen congressmen from *Homestyle*, a Shakespeare-reading lawyer, former Marine jet pilot, water-toothbrush entrepreneur, and amateur theologian representing the farms, mines, foothills, and resorts of northern Colorado. So you think congressmen only care about getting reelected? Johnson is an independent, anti-war Republican who walked away in 1980 after eight years in office because he was bored ("Most of your time is spent listening to bullshit and jollying") and frustrated by the course of public policy ("Right now, the people want a military buildup—the Carter doctrine, the MX missile. You can feel it coming—rolling, rolling—and there is no way I can make a difference on that issue"). Fenno concludes, characteristically, with a whole series of general categorizations suggested by this single case: congressmen motivated by policy, citizen legislators uninterested in Washington careerism, congressmen whose independence on roll call votes is "underwritten by the quality" of their constituency relations, and so on.

Here, as elsewhere, Fenno views close-up observation as one link in a chain of analysis connecting political theory and the real world of politics. His job, in his own deceptively simple words, is "to go take a first-hand look at our politicians and report back." By "back" he means very specifically "back to other political scientists," a fact eloquently

confirmed by the fate of his Quayle manuscript (available from CQ Press, but not in supermarkets).

Fenno ends with a plea to the discipline to join in his work and to train his successors. "How much value do we place on observation as a research mode? How much legitimacy do we wish to bestow on observation-based research? Have we tried our very best to teach observation and failed?"

One reaction to Fenno's plea is to notice that the essays collected here provide several excellent arguments for sitting in Rochester spinning computer tapes. I have never had to fix a congressman's flat tire on a mountain road in the hope that he might later answer some questions (though I do often spend fruitless days building rapport of another sort with my own data). Barring major failures of hardware or software, I can count on access to my subjects (or at least to Warren Miller's); Fenno frets about access constantly. Perhaps most importantly, I usually know what I'm studying before I start, whereas Fenno lives on an intellectual high wire, investing months or years in projects the fruits of which he can scarcely foresee.

It is good to be reminded of these advantages of my own style of research, but better still to be reminded of the corresponding disadvantage—to recall that my work is the sort that reduces complex motivations, strategies, and behaviors like those of the six Senators in Fenno's AWACS example to the roll call vote vector [0 0 0 1 1 1], that turns a fascinating human being like Jim Johnson into a residual. No one who reads these essays will believe that that is a small price to pay, or that the many who choose to pay it are any more serious or sophisticated political scientists than the hardy few who do not.

Contents of *Political Analysis*, Volume 2

The second volume of *Political Analysis*, edited by James Stimson, will be published shortly by the University of Michigan Press. *Political Analysis* is the journal of the political methodology section of APSA and provides the primary outlet for research in political methodology. Special discount pricing for *PA* is available to members of the methodology section. Here is an advance peek at the contents of the coming issue.

Call for Methodology Papers for the 1991 APSA Annual Meeting

Charles H. Franklin
Washington University

It is time to get those proposals in for the 1991 Annual Meeting of the American Political Science Association. As section chair for political methodology, I want to especially encourage readers of *TPM* to submit proposals which advance the state of both substantive and methodological work in political science. The deadline for proposals is December 1, 1990. Proposals should be sent to: Charles H. Franklin, Department of Political Science, Washington University, One Brookings Drive, St. Louis, MO 63130 (314) 889-5874 (Bitnet: C38871CF@WUVMD.BITNET)

I am looking for two types of papers for this section. One is sophisticated substantive papers which "show how the best work in the field should be done". These papers should exemplify the best work which is simultaneously innovative methodologically and substantively. I hope we will have such panels in several subfields, such as international relations, comparative politics, electoral behavior and public opinion.

The second sort of paper is explicitly methodological and represents the development of new techniques which may be applicable to a variety of substantive settings. Topics might include selection bias, models for discrete data, maximum likelihood methods, survey sampling, Bayesian analysis, time series, scaling, aggregation and ecological inference.

I am anxious to see innovative work from the broad spectrum of methodology, including research design and other non-statistical aspects of methods. I would especially welcome papers on experimental designs.

Call for Paper Proposals, 1991 Political Methodology Conference

John R. Freeman
University of Minnesota

The Eighth Annual Political Methodology Conference will be held Wednesday July 17 through Sunday July 21, 1991, at Duke University.

Faculty who would like to participate in the conference should submit a 3-5 page proposal describing the research which they wish to present. Of particular interest are papers on long-standing methodological problems of theoretical import. Put another way, emphasis will be on work which both advances our technical understanding of and ability to cope with important methodological issues and which advances our theoretical understanding of politics.

A small number of advanced graduate students will be invited to attend the conference. These individuals will *not* be required to present a paper, although they are welcome to propose to do so. Graduate students should be nominated by their faculty advisors. The advisor's letter should include a brief description of the student's methodological training and research interests. A vita for the student should accompany this letter.

The National Science Foundation and Duke University will cover the costs of attending the conference. A combined total of 30 faculty and graduate students will receive invitations. Other interested scholars are welcome to attend the sessions. However, only the 30 individuals who receive formal invitations will be reimbursed for travel expenses and provided with lodging and meals.

Proposals should be sent to John R. Freeman, Department of Political Science, 1414 Social Sciences Building, Minneapolis, Minnesota 55455. The deadline for receipt of proposals is March 1, 1991.

Program of the 1990 Political Methodology Annual Meeting

The Seventh Annual Political Methodology conference was held at Washington University, in St. Louis, July 19-21, 1990. Thirty-nine participants made up the official program. Attendance by political scientists from St. Louis brought attendance at each session to about 50. This was the largest turnout in the seven year history of the summer meetings.

The summer Political Methodology meetings are supported by a grant from the National Science Foundation. Additional support for housing and meals is provided by the host institutions. Previous meetings have been hosted by The University of Michigan, The University of California, Berkeley, Harvard University, Duke University, The University of California, Los Angeles, The University of Minnesota and Washington University.

In addition to the usual paper givers and discussants, this year's meetings continued the practice begun in 1989 of inviting several graduate students to attend the meetings without the obligation of presenting a paper. Graduate student participation has proven beneficial to both students (who gain a better perspective on the state of political methodology) and faculty (who get to meet their soon to be colleagues).

Papers and Discussants

W. Phillips Shively, University of Minnesota, "Cross Level Inference"

Disc: Gary King, Harvard University

Henry Brady, University of Chicago, "Statistical Methods for Analyzing Strategic Voting"

Disc: Stanley Feldman, SUNY–Stony Brook

Melvin J. Hinich, University of Texas, "Applying Diagnostic Tests for Improving the Fit of Apparently Correctly Specified Forecasting Models"

Disc: John Williams, Indiana University

B. Dan Wood, Texas A&M University, "Resolving Problems of Temporal and Spatial Aggregation with Pooled Granger Methods"

Disc: John Williams, Indiana University

Larry Bartels, University of Rochester, "Instrumental and 'Quasi-Instrumental' Variables (II)"

Disc: Douglas Rivers, Stanford University

Robert S. Erikson, University of Houston, "Estimating the Incumbency Advantage in Congressional Elections"

Disc: Charles H. Franklin, Washington University

John E. Jackson, University of Michigan, "A Model of Endogenous Voter and Party Preferences"

Disc: Charles H. Franklin, Washington University

Nathaniel Beck, University of California, San Diego, "Discerning Cycles in Political Data"

Disc: John Freeman, University of Minnesota

Tse-min Lin, SUNY Stony Brook, "Equilibrium Cycles in Presidential Elections"

Disc: John Freeman, University of Minnesota

Charles W. Ostrom, Jr. and Renée M. Smith, Michigan State University, "Cointegration and Error Correction: Examining the Presidential Approval-Economy Connection"

Disc: Jim Stimson, University of Iowa

Lois W. Sayrs, University of Iowa, "Slutzky's Problem and War Cycles"

Disc: Christopher H. Achen, University of Michigan

Renée Marlin-Bennett, The American University, James C. Roberts, Towson State University, and Alan Rosenblatt, The American University, "A Methodology for Testing Game Theoretic Hypotheses Using Events Data"

Disc: Christopher H. Achen, University of Michigan

Walter R. Mebane, Jr., Cornell University, "Spatially Aggregated Analysis using the Censuses and Annual Surveys of Governments"

Disc: Herbert M. Kritzer, University of Wisconsin

Graduate Student Participants

R. Michael Alvarez, Duke University
Fay E. Booker, University of Chicago
Janet Box-Steffensmeier, University of Texas
Nancy Burns, Harvard University
Elizabeth Gerber, University of Michigan
Victoria Lynn Gerus, Harvard University
Mike Gilligan, Harvard University
James S. Granato, Duke University
Simon Jackman, University of Rochester
Glenn E. Mitchell II, University of Iowa
Renée M. Smith, Michigan State University
Marco R. Steenbergen, SUNY Stony Brook
Margaret C. Trevor, University of Chicago

Textbooks Needed for Methodologists in Czechoslovakia

Neal Beck
University of California, San Diego

I visited with some methodologists in Prague this summer. After 20 years they are starting to rebuild the social sciences in Czechoslovakia. The repression that followed the 1968 invasion makes this a difficult task, since a whole generation of social scientists was, more or less, not trained. My friend in Prague was not allowed to teach after 1968 (wrong political beliefs). This problem is compounded by a lack of hard currency. A box of disks, for example, costs about one week's salary for a professor. Books and software are prohibitively expensive.

Both Gauss and Limdep have agreed to send versions of their programs to the Sociological Institute in Prague. Sage Publishers has kindly made available many of its volumes to that institute.

I would like to be able to send good statistics and methods texts to that institute. If anyone has spare copies of any good texts that they would be willing to donate to the institute, I would appreciate it if they would send them to me and I will send them on to Prague.

Many thanks.

Neal Beck
Department of Political Science
University of California, San Diego
La Jolla, CA 92093

The Political Methodologist
Department of Political Science
Washington University
One Brookings Drive
St. Louis, MO 63130

Submissions to *TPM* should be sent to Charles Franklin, Department of Political Science, Washington University, St. Louis, MO 63130. (314) 889-5874. (Bitnet: C38871CF@WUVMD.Bitnet). We prefer submissions in either TEX or LATEX formats, or ASCII files. Best of all, is submission via Bitnet. Subscriptions to *The Political Methodologist* are free to members of APSA's Political Methodology Section and $15.00/year to others.