Making the Most of Statistical Analyses: Improving Interpretation and Presentation[1]

Gary King
Department of Government
Harvard University

Michael Tomz
Department of Government
Harvard University

Jason Wittenberg
Department of Political Science
MIT[2]

August 10, 1998

[1]We thank Bruce Bueno de Mesquita, Jorge Domínguez, Geoff Garrett, Jay McCann, and Randy Siverson for their data; Jim Alt, Jorge Domínguez, Ken Scheve, Brian Silver, and Richard Tucker for their very helpful comments; and U.S. National Science Foundation (SBR-9729884), the Centers for Disease Control and Prevention (Division of Diabetes Translation), the National Institutes of Aging, the World Health Organization, and the Global Forum for Health Research, for research support.

[2]King: Littauer Center North Yard, Cambridge MA 02138; King@Harvard.Edu, http://GKing.Harvard.Edu, (617) 495-2027; Tomz: Littauer Center North Yard, Cambridge MA 02138; tomz@fas.harvard.edu. Wittenberg: E53-470, Cambridge, MA 02139; jwittenb@latte.harvard.edu

Abstract

We demonstrate that social scientists rarely take full advantage of the information available in their statistical results. As a consequence they miss opportunities to present quantities that are of greatest substantive interest for their research and to express their degree of certainty about these quantities. In this paper we offer an approach built on the technique of statistical simulation to extract the currently overlooked information from any statistical method no matter how complicated and to interpret and present it in a reader-friendly manner. Using this technique requires some sophistication which we try to provide herein but its application should make the results of quantitative articles more informative and transparent to all. To illustrate our recommendations we replicate the results of several published works showing in each case how the authors' own conclusions can be expressed more sharply and informatively and how our approach reveals important new information about the research questions at hand. We also offer very easy-to-use software that implements our suggestions.

1 Introduction

We show that social scientists rarely take full advantage of the information available in their statistical results. They do not calculate and present the quantities that are of most substantive interest and therefore miss much opportunity to bring the most important information to bear on their empirical questions. Unlike most methods papers, we do not ask researchers to make changes in their statistical model. Instead, we offer an approach, built on the technique of statistical simulation, to extract the currently overlooked information from any statistical method, no matter how complicated, and to interpret and present it in a reader-friendly manner.

To be more specific, we show how to convert the raw results of any statistical procedure into expressions that (1) *convey numerically precise estimates of the quantities of substantive interest*, (2) *include reasonable assessments of uncertainty about those estimates*, and (3) *require no specialized knowledge to understand*. One simple example of how to fulfill these criteria is with statements of results like this: "Other things being equal, an additional year of education would increase your annual income by $1,500 on average, plus or minus about $500." Any smart high school student would understand that sentence, no matter how sophisticated the statistical model and powerful the computers used to produce it. The sentence is substantively informative, because it conveys a key quantity of interest in the terms the reader wants to know: how much higher the starting salary would be if the student attended college for an extra year. At the same time, the sentence conveys the degree of uncertainty the analysis admits about the estimated quantity of interest. Inferences are never certain, so any honest presentation of statistical results must include some qualifier, such as the "plus or minus $500" in the present example.

In contrast, bad intepretations are substantively ambiguous and filled with methodological jargon: the "coefficient on education was statistically significant at the 0.05 level." Although descriptions like this are very common in social science, students, public officials and scholars should not need to understand phrases like "coefficient," "statistically significant," and "the 0.05 level" to learn from the research. Moreover, even statistically savvy readers should complain that the sentence does not convey the key quantity of interest: how much higher the starting salary would be if the student attended college for an extra year.

Our approach can help researchers do better in three ways. First, and most importantly, it can extract new quantities of interest from standard statistical models, thereby enriching the substance of social science research. Second, our approach allows scholars to assess the uncertainty surrounding any quantity of interest, so it should improve the candor and realism of statistical discourse about politics. Finally, our method can convert raw statistical results into results that everyone, regardless of statistical training, can comprehend. The examples in this paper should make all three benefits apparent.

The remainder of this paper describes our approach to interpreting and presenting statistical results. We begin by formalizing the problem of statistical interpretation (Section 2) and summarizing the existing methods of dealing with it (Section 3). Next, we introduce the technique of statistical simulation (Section 4) and present our approach and algorithms for using simulation to interpret and present statistical results (Section 5). Section 6 suggests a few tricks for enhancing the effectiveness of simulation. Finally, we illustrate our method in Section 7, where we replicate the results of three prominent published works, showing in each case how the authors' own conclusions can be expressed more sharply and informatively and how important new information comes to light without any new analyses. To assist researchers in implementing our technique, we have developed a very

easy-to-use, public domain software package called CLARIFY, which we describe in the Appendix.

2 Statistical Models to Interpret

We aim to interpret the raw results from any member of a very general class of statistical models. Using the notation from King (1989), we summarize the class with two equations:

$$\begin{aligned} Y_i &\sim f(\theta_i, \alpha) \\ \theta_i &= g(X_i, \beta) \end{aligned} \tag{1}$$

The first equation describes the *stochastic component* of the statistical model: the probability density (or mass function) that generates the dependent variable. For each observation i ($i = 1, \ldots, n$), the dependent variable Y_i is a random draw from the probability density $f(\theta_i, \alpha)$. Some characteristics of this function vary from one observation to the next, while others remain constant across all the i's. We represent the varying characteristics with the parameter vector θ_i and relegate non-varying features to the ancillary parameter matrix α.

The second equation gives the systematic component of the model; it indicates how θ_i changes across observations, depending on the values of the explanatory variables (normally including a constant) in the $1 \times k$ vector X_i and the effect parameters in the $k \times 1$ vector β. The functional form $g(\cdot, \cdot)$ specifies how the explanatory variables and effect parameters get translated into θ_i.

A simple member of this very general class is a linear-normal regression model, otherwise known as least squares regression. To see this, let $f(\cdot, \cdot)$ be the normal distribution $N(\cdot, \cdot)$, set the main parameter vector to the scalar mean $\theta_i = E(Y_i) = \mu_i$, and let the ancilliary parameter matrix be the scalar homoskedastic variance $\alpha = V(Y_i) = \sigma^2$. Finally, set the systematic component to the linear form $g(X_i, \beta) = X_i\beta$. The result is familiar:

$$\begin{aligned} Y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= X_i\beta \end{aligned} \tag{2}$$

Similarly, we can write a logit model by expressing the stochastic component as a Bernoulli distribution with parameter $\pi_i = \Pr(Y_i = 1)$ — no ancillary parameter is necessary — and setting the systematic component to the logistic form:

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\pi_i) \\ \pi_i &= \frac{1}{1 + e^{-X_i\beta}} \end{aligned} \tag{3}$$

Equations 1 also include as special cases nearly every other statistical model used in the social sciences, including multiple equation models in which Y_i is a vector, as well as specifications for which the probability distribution, functional form, or explanatory variables are estimated rather than assumed to be known.

3 Existing Methods of Model Interpretation

When interpreting the results of statistical analyses, many researchers still stop after a cursory look at "statistical significance" and the signs of the coefficients. This approach obviously fails to meet all of our three criteria given Section 1 for meaningful statistical

Explanatory Variable	Coefficient	Standard Error
Education	.181	.007
Age	.109	.006
$\text{Age}^2/100$	$-.078$	.007
Income	.151	.010
White	.116	.054
Constant	-4.715	.174

Table 1: *The Determinants of Voter Turnout, Presidential Election Years 1960-1996. The table lists logistic regression coefficients and corresponding standard errors for each explanatory variable and the constant term. The outcome variable is whether or not a citizen casts a ballot on election day.* $n = 15,873$.

communication. Some researchers do convey numerically precise estimates of the quantities of interest and in a way that requires no specialized knowledge to understand but the primary methods available for complex models (e.g. King 1989: Section 5.2) ignore both estimation and fundamental uncertainty. As a result they do not include reasonable assessments of the uncertainty of their estimates. We describe these procedures here and then improve on them in subsequent sections.

We discuss existing approaches — and their limitations — in the context of a simple logit example. Our example draws on the work of Rosenstone and Hansen (1993) who sought to explain why some individuals are more likely than others to vote in U.S. presidential elections. For expository purposes we focus on only a few demographic variables that Rosenstone and Hanson emphasized.

We pool every National Election Study since 1960 that was conducted during a presidential election year. The dependent variable Y_i is coded 1 if the respondent reported voting in the presidential election and 0 otherwise. The probability of voting should increase with the respondent's Education (E_i) in years and Income (I_i) in 10,000s of dollars. Turnout should also rise with years of Age (A_i) until the respondent nears retirement when the tendency should reverse. Finally turnout should be higher among Whites ($W_i = 1$) than non-whites ($W_i = 0$). Thus our set of explanatory variables is $X_i = \{\mathbf{1}, E_i, I_i, A_i, A_i^2, W_i\}$ where $\mathbf{1}$ is a constant and A_i^2 is a quadratic term.

After estimating a logit model like this one many researchers present only the coefficients and standard errors (see Table 1) generated by their statistical package pausing only to comment on statistical significance and the sign of the coefficients. Although these numbers are typically regarded as the "results" of the statistical procedure they are only one type of result and are in no way canonical. Statistical packages compute many quantities of marginal interest such as inverses of matrices and sums of squares in regression programs but researchers do not publish those results because they are not intrinsically interesting. Likewise political scientists should not publish $\hat{\beta}$ and $\hat{\alpha}$ if they are not inherently interesting as is usually the case with nonlinear models. At the very least such quantities should be moved to an appendix. Instead of focusing on parameter estimates and t-statistics researchers should present quantities of direct substantive interest. Examples include the probability that an individual will vote in the presidential election.

Some researchers go a step further: to see how much the probability of voting changes

with a given explanatory variable, they take the partial derivative of $\hat{\theta}_i$ with respect to the variable of interest. This *derivative method* is equivalent to drawing a linear approximation to the nonlinear curve at one point, and then reporting the slope of that one line.

The derivative method can shed light on our logit results. In our model, $\hat{\theta}_i = \hat{\pi}_i \equiv \Pr(Y_i = 1)$, the probability of voting, is a quantity that is more interesting and comprehensible than the effect parameters listed in Table 1. The partial derivative of $\hat{\pi}_i$ with respect to E_i is $\partial \pi_i / \partial E_i = \pi_i(1 - \pi_i)\beta_E$, where β_E is the effect parameter corresponding to the education variable. This derivative reaches a maximum at $\pi_i = 1/2$, in which case $\partial \pi_i / \partial E_i = \beta_E / 4$. By substituting in the point estimate, $\hat{\beta}_E$, we find that the instantaneous effect of an additional year of education is to increase the probability of voting by 0.045. The derivative method is easy to implement — just divide the coefficients by four — but it has important limitations: it applies only at the maximum point, not even for one unit around that point, and certainly not for any other point. The method also requires that the researcher compute a different derivative for each type of nonlinear model. Since this is only meant to provide a quick and dirty method of interpretation, we do not consider it further.

Another approach involves choosing a group of substantively interesting "observation-types" defined by their values on the explanatory variables, and listing the *fitted values* for each. Unlike the derivative approach, this method does not attempt to interpret non-linear models as if they were linear. For instance, we might be interested in the probability of voting for a 42 year old ($A = 42$) college educated ($E = 16$) black citizen ($W = 0$) making an annual salary of \$57,000 ($I = 5.7$). To determine the probability that someone with these characteristics votes, we plug these values and the corresponding $\hat{\beta}$'s into the logistic functional form in the second line of Equation 3. In our case, the probability of voting was 0.9.

A third method, now called *first differences*, conveniently provides one number for the effect of each explanatory variable. This number indicates how much the expected value changes when the investigator shifts one explanatory variable from an initial level to a different level, while holding other variables constant at their means. In our case, increasing education from high school ($E = 12$) to college ($E = 16$) boosted the probability of voting by 0.09.

These three approaches — the derivative method, fitted values, and first differences — represent improvements over simply listing parameter estimates, but they are ultimately inadequate because they ignore uncertainty. The statistical model given in Equation 1 contains two forms of uncertainty. *Estimation uncertainty* results from not knowing β and α perfectly, an unavoidable consequence of having fewer than an infinite number of observations. Researchers usually acknowledge this uncertainty by reporting standard errors or t-statistics from their computer output. Since β and α are uncertain, however, any calculations based on estimates of those parameters must also be uncertain, a fact that almost no researchers take into account. In contrast, *fundamental uncertainty*, represented by the stochastic component (the distribution f) in the first line of Equation 1, results from the innumerable chance events such as weather or illness that may influence Y but are not included in X. Even if we could know the exact values of the parameters (thus eliminating estimation uncertainty), fundamental uncertainty would prevent us from predicting Y with certainty. Our methods for computing quantities of interest must account for both types of uncertainty.

4 An Introduction to Statistical Simulation

Researchers could use advanced mathematics to derive measures of uncertainty such as standard errors and confidence intervals for any of the methods described in Section 3. The calculations would however be technically demanding. Thus it is not surprising that most scholars never make these calculations even though the importance of reporting uncertainty is widely appreciated. We now introduce *simulation* as an easy method of calculating uncertainty estimates and more broadly of understanding the entire statistical model (see Mooney 1993; Tanner 1996).

Statistical simulation uses the logic of survey sampling to approximate complicated mathematical calculations. The trick is that drawing random numbers (or "simulations") from probability distributions is often quite easy and the numbers can be used to *approximate* quantities of interest just as random samples from a population can be used to *estimate* unobserved but interesting features of that population.

For example we could compute the mean of a probability distribution $P(y)$ by taking the integral $E(Y) = \int_{-\infty}^{\infty} yP(y)dy$ but this is not always be the most pleasant of experiences! Alternatively we could approximate the mean through simulation by drawing many random numbers from $P(y)$ and computing their average. To approximate the theoretical variance of Y we could take the sample variance of a large number of random draws and to approximate the probability that $Y > 0.8$ we could count the fraction of draws that are larger than 0.8. Likewise we could approximate the variance of a *function* of Y by drawing values of Y computing the function for each draw and then calculating the variance. Even the complete distribution of say $\sqrt{Y}$ could be computed by plotting a histogram of the square root of a large number of simulations of Y.

The approximation can be computed to *any* desired degree of accuracy by increasing the number of simulations so little is lost by simulation — except a bit of computer time — and much is gained in ease of use. Indeed it is well-known among teachers of statistics that no matter how well the analytical methods are taught students get the right answer far more frequently when taught simulation. Most methodologists now use simulation to check their own analytical calculations. To assess the precision of the approximation run the same procedure with the same number of simulations repeatedly; if the answer is the same to within 4 decimal points then that is how accurate the approximation is. If more accuracy is needed increase the number of simulations and try again.

To be more formal about it the same rules apply here as apply to the analysis of simple random samples. Suppose we are approximating a mean of some function of the data (a point estimate). If σ^2 is the simulation variance and M is the number of simulations the approximation error (in standard deviation units) e is $e = \sigma/\sqrt{M}$. We can also turn around this equation to find the number of simulations necessary given the specified variance of the quantity being approximated and the acceptable approximate error: $M = \sigma^2/e^2$.

As an example in forecasting congressional elections the error in predicting the vote is normally around $\sigma = 0.06$ (6% of the vote). If we are willing to accept an approximation error of 0.002 the necessary number of simulations is $0.06^2/0.002^2 = 900$. Approximating more complicated statistics such as variances is harder to analyze formally but they usually require more simulations.

5 Algorithms for Translating Statistical Results

We now explain how to use statistical simulation to compute quantities of interest and account for estimation and fundamental uncertainty. Our method approximates the *entire* probability distribution of the quantity of interest—allowing researchers to compute point estimates—standard errors—confidence intervals—and most anything else desired.

We begin with a procedure for computing estimation uncertainty—which arises because we cannot fully know the true values of the effect coefficients and ancillary parameters. For example—standard errors provide a shorthand measure of uncertainty about the estimated coefficients—but they do not capture all aspects of the estimation uncertainty associated with the model's parameters. In particular—they do not tell us anything about covariances among different parameter estimates.

Most statistics packages store not only the standard errors but also the covariances—in a variance matrix—which is generated during the estimation step. The standard errors are the square root of the diagonal elements of this matrix; the off-diagonal elements express the covariances between one parameter estimate and another in repeated draws from the same probability distribution. If a model contained any ancillary parameters—they would appear along with the other coefficients and would have their own rows and columns in the variance matrix.

The key to our method is in understanding how to use both the parameter estimates and the variance matrix to produce many plausible sets of parameter estimates—some of which may be larger or smaller than the estimates produced during the estimation phase—but all consistent with and drawn from the model. As shorthand—designate $\hat{\gamma}$ as the column vector produced by stacking the estimated effect coefficients ($\hat{\beta}$) on top of the ancillary parameters ($\hat{\alpha}$). More formally—$\hat{\gamma} = \text{vec}(\hat{\beta}, \hat{\alpha})$—where "vec" stacks the unique elements of $\hat{\beta}$ and $\hat{\alpha}$ in a column vector. Then let $\hat{\text{V}}(\hat{\gamma})$ designate the variance matrix associated with these estimates.

The central limit theorem tells us that with a large enough sample and when the usual regularity conditions are met—we can draw (i.e.—simulate) our parameters from a multivariate normal distribution with mean equal to $\hat{\gamma}$ and variance equal to the variance matrix of estimates $\hat{\text{V}}(\hat{\gamma})$.[1] Using our notation—

$$\gamma \sim \text{N}\left(\hat{\gamma}, \hat{\text{V}}(\hat{\gamma})\right) \tag{4}$$

Appendix A explains how to make these random draws using a popular software package.

At the extreme—if we knew the elements of γ perfectly—the sets of draws would all be identical; the more they vary (due to larger elements in the variance matrix)—the less we would know about γ; the specific manner in which they vary summarizes all knowledge about the parameters that we can obtain from the statistical procedure. We still need to translate γ into quantities of more substantive interest—but now that we have summarized all knowledge about γ we are in a position to make the translation.

In most cases—the assumptions necessary to use our methods are exactly those made by applied social scientists in using available statistical methods. That is—we all assume that the statistical model is identified and the central limit theorem holds sufficiently for the available sample size so that the sampling distribution (not the stochastic component) can be described by a normal distribution. From a Bayesian perspective—we exclude the

[1] This distributional statement is a shorthand summary of the Bayesian, likelihood, and Neyman-Pearson theories of statistical inference. The interpretive differences among these theories (such as whether θ or $\hat{\theta}$ is the random variable) are important but need not concern us here as our approach can usually be used with any of these and most other theories of inference (see Barnett, 1982).

unusual cases where our flat prior generates an improper posterior. Although we focus on asymptotic results for simplicity, as do the vast majority of the applied researchers using nonlinear statistical models, the approach we recommend here also works as easily with finite sample distributions, which of course are preferable when feasible. Our presentational methods are conditional on the model and so they work whenever these usual assumptions work.

In the next three sub-sections, we describe algorithms for converting the simulated effect and ancillary parameters into quantities of interest. The algorithms will account not only for the estimation variability that we have just discussed, but also for the fundamental variability arising from the stochastic component of the statistical model. We provide algorithms for generating predicted values, expected values, and first differences, but researchers could, and in most cases should, use simulation to compute other quantities of interest, as we show in section 7.

5.1 Predicted Values

Our task is to draw one value of Y, conditional on one chosen value of each of the explanatory variables, which we represent as the vector X_c. Denote the simulated θ as $\tilde{\theta}_c$ and the corresponding simulated Y as $\tilde{Y}_c$. The result is called a simulated predicted value, although X_c may correspond to the future (in which case $\tilde{Y}_c$ is a simulated forecast), a real situation described by observed data (so that $\tilde{Y}_c$ is a simulated fitted value), or a hypothetical situation not necessarily in the future (so that $\tilde{Y}_c$ is a simulated counterfactual fitted value). (None of these, of course, is the y-hat from regression which is in that context an expected value, about which more in Section 5.2.) To draw one such simulation, follow these steps.

1. Estimate the model by running the usual software program (which normally maximizes a likelihood function), and record the point estimates $\hat{\gamma}$ and variance matrix $\hat{\mathrm{V}}(\hat{\gamma})$.

2. Draw one value of the vector γ from the multivariate normal distribution in Equation 4. Denote the draw $\tilde{\gamma} = \mathrm{vec}(\tilde{\beta}, \tilde{\alpha})$. This step simulates estimation uncertainty.

3. Decide what predicted value you wish to compute and on that basis choose one value for each explanatory variable, and denote the vector of such values X_c.

4. Using the simulated effect coefficients from the top portion of $\tilde{\gamma}$, compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$, where $g(\cdot, \cdot)$ is the systematic component of the statistical model.

5. Simulate the outcome variable $\tilde{Y}_c$ by taking a random draw from $f(\tilde{\theta}_c, \tilde{\alpha})$, the stochastic component of the statistical model. This step simulates fundamental uncertainty.

We repeat this algorithm, say, $M = 1000$ times to produce 1000 predicted values, thereby approximating the entire probability distribution of Y_c. From these simulations, the researcher can compute not only the average predicted value, but also measures of uncertainty around that prediction.

To illustrate how the algorithm works in practice, we estimate a log-linear regression model of the size of government in the U.S. states. Our dependent variable, Y_i, is the natural log of the number of people (measured in 1000s) that the state government employed on a full-time basis in 1990 (see Tufte, 1974). We hypothesize that the number of employees might depend on the proportion of Democrats in the state legislature, since Democrats are reputed to favor bigger government than Republicans, even after adjusting for state

population. Thus, our two main explanatory variables are the log of state population (P_i) in 1000s and the logged proportion of lower-house legislators who identified themselves as Democrats (D_i).

We now apply our algorithm to predict the number of government employees in a state with 6 million people and an 80% Democratic house. First, we use the statistical software described in Appendix A to estimate the log-linear model and simulate 1 set of values for the effect coefficients ($\tilde{\beta}$) and the ancillary parameter ($\tilde{\sigma}$). Next, we set the main explanatory variables at $P_c = \ln(6000)$ and $D_c = \ln(0.8)$, so we could construct X_c and compute $\tilde{\theta}_c = X_c\tilde{\beta}$. We then draw one value of $\tilde{Y}_c$ from the normal distribution $N(\tilde{\theta}_c, \tilde{\sigma}^2)$. Finally, we calculate $\exp(\tilde{Y}_c)$ to transform our simulated value into the actual number of government employees, a quantity that is more understandable than a natural logarithm.

By repeating this process $M = 1000$ times, we generated 1000 predicted values, which we sorted from lowest to highest. The numbers in the 25th and the 976th positions represented the upper and lower bounds of a 95 percent confidence interval. Thus, we predict with 95 percent confidence that the state government would employ between 74,000 and 151,000 people. Our best guess was 107,000 full-time employees, the average of the predicted values.[2] (Democratic state legislators do appear to produce states with more government employees; See Section 5.3.)

5.2 Expected Values

Depending on the issue being studied, the *expected* or mean value of the dependent variable may be more interesting than a predicted value. The difference is subtle but important. A predicted value contains both fundamental and estimation uncertainty, whereas an expected value averages over the fundamental variability arising from sheer randomness in the world, leaving only the estimation uncertainty caused by not having an infinite number of observations. Thus, predicted values have a larger variance than expected values, even though the average should be nearly the same in both cases.[3]

When choosing between these two quantities of interest, researchers should reflect on the importance of fundamental uncertainty for the conclusions that they are drawing. For certain applications, such as forecasting the actual result of an election or predicting next month's foreign exchange rate, scholars and politicians — as well as investors — want to know not only the expected outcome, but also how far the outcome could deviate from expectation due to unmodeled random factors. Here, a predicted value seems most appropriate. For other applications, the researcher may want to highlight the mean effect of a particular explanatory variable, so an expected value would be the best choice.

We now offer an algorithm for creating *one* simulation of an expected value, given chosen values for the explanatory variables.

1. Estimate the model by running the usual software program, and record the point estimates of $\hat{\gamma}$ and variance matrix $\hat{\mathrm{V}}(\hat{\gamma})$.

2. Draw one value of the vector γ from the multivariate normal distribution in Equation 4. Denote the simulations as $\tilde{\gamma} = \mathrm{vec}(\tilde{\beta}, \tilde{\alpha})$. This step simulates estimation uncertainty.

[2] We verified that $M = 1000$ was high enough by rerunning the entire procedure; nothing of substantive interest changed.

[3] In linear models, the average predicted value is identical to the expected value. For non-linear cases, the two can differ but are often close if the nonlinearity is not severe.

3. Decide what expected value you wish to compute and on that basis choose one value for each explanatory variable, and denote the vector of such values X_c.

4. Using the simulated effect coefficients from the top portion of $\tilde{\gamma}$, compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$, where $g(\cdot, \cdot)$ is the systematic component of the statistical model.

5. Draw m simulations of the outcome variable $\tilde{Y}_c^{(k)}$ $(k = 1, \ldots, m)$ from the stochastic component $f(\tilde{\theta}_c, \tilde{\alpha})$ ($m = 1000$ will usually be sufficient). This step simulates fundamental uncertainty.

6. Average over the fundamental uncertainty by calculating the the numerical average of the m simulations to yield one simulation of the expected value $\tilde{\mathrm{E}}(Y_c) = \sum_{k=1}^{m} \tilde{Y}_c^{(k)}/m$.

When $m = 1$, this algorithm reduces to the one for predicted values. If m is a larger number, Step 4 accurately portrays the fundamental variability, which step 5 averages-away to produce an expected value. The larger the value of m, the more successful the algorithm will be in purging $\tilde{\mathrm{E}}(Y_c)$ of any fundamental uncertainty.

To generate 1000 simulations of the expected value, repeat steps 2-5 of the algorithm $M = 1000$ times for some fixed value of m. The resulting expected values will differ from each other due to estimation uncertainty, since each expected value will correspond to a different $\tilde{\gamma}$. These M simulations approximate the entire probability distribution of $E(Y_c)$, enabling the researcher to make inferences about the expected value.

Our algorithm works in all cases, but it does involve some approximation error, which can be reduced by setting both m and M sufficiently high. For some statistical models, there is a shortcut that curtails both computation time and approximation error. Whenever $E(Y_c) = \theta_c$, the researcher can skip steps 4-5 of the expected value algorithm, since steps 1-3 suffice to simulate one expected value. This shortcut is appropriate for logistic regression, which we now use to illustrate our technique.

In our running logit example, the probability of voting in a presidential election is $E(Y) = \theta = \pi$. We estimated this probability, and the uncertainty surrounding it, for several interesting "observation types." First, we used the software described in Appendix A to estimate our model of voter turnout and draw one set of coefficients $(\tilde{\beta})$ from a multivariate normal distribution. Next, we constructed X_c by setting education at 12 years (equivalent to a high school degree), age at 30 years, and other variables at their mean values. This enabled us to calculate $\tilde{\pi}_c = 1/(1 + e^{-X_c\tilde{\beta}})$, the probability of voting.

By repeating steps 2-3 of the expected value algorithm $M = 1000$ times, we generated 1000 predicted probabilities of voting for an individual with characteristics X_c. Running the algorithm 1000 times was our equivalent of re-running the election 1000 times, given the same election campaign and citizens but allowing other forms of variability. The differences in $\tilde{\pi}_c$ across the various runs were due to estimation uncertainty. By sorting these probabilities and recording the values in the 10th and 991st positions, we obtained a 99 percent confidence interval, which appears as the horizontal bar in the lower left corner of Figure 1. The confidence interval is quite narrow, because the large number of observations $(n = 15,873)$ eliminated most of the estimation uncertainty.

We used a similar procedure to obtain confidence intervals for 60 year olds with a high school degree, as well as college-educated people at ages 30 and 60. These intervals appear in Figure 1, which shows that being college-educated, as opposed to having only a high school degree, increases the probability of voting substantially, and more for 30 year olds than for 60 year olds (because the 60 year olds already have high probabilities.) The

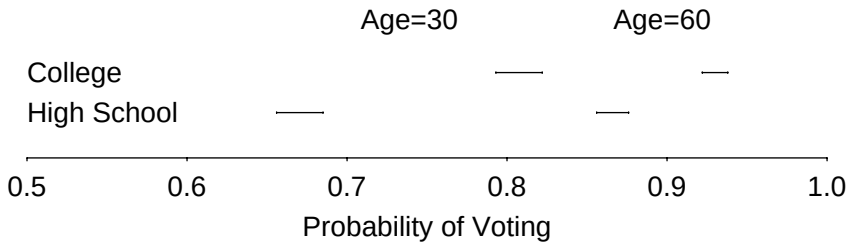

Figure 1: *Probabilities of Voting: The horizontal lines represent 99% confidence intervals for the probablity of voting, conditional on age and education, holding other variables constant at their mean values.*

difference in the probability of voting across the two age categories is also quite large but the gap is bigger for those with only a high school degree.

Simulation enables us to present the expected values in other ways too. For two different levels of education we calculated 95 percent confidence intervals around the predicted probability of voting across the entire range of age. To do this we repeated our expected value algorithm $M = 1000$ times for each configuration of education and age. The results appear in Figure 2 which illustrates the conclusions of Rosenstone and Hansen quite sharply: the probability of voting rises steadily to a plateau between the ages of 45 and 65 and then tapers downward through the retirement years. The figure also reveals that uncertainty associated with the expected value is greatest at the two extremes of age: the vertical bars which represent 95 percent confidence intervals are longest when the respondent is very young or old.

5.3 First Differences

A simulated first difference is the difference between two expected rather than predicted values. To simulate a first difference researchers need only run the expected value algorithm twice using different settings for the explanatory variables.

For instance to simulate a first difference for the first explanatory variable set the values for all explanatory variables except the first at their means and fix first one at its starting point. Denote this vector of values of the explanatory variables as X_s and run the expected value algorithm once to generate $\tilde{\mathrm{E}}(Y_s)$ the average value of Y conditional on X_s. Next change the value of the first explanatory variable to its ending point leaving the others at their means as before. Denote these values as X_e and re-run the algorithm to get $\tilde{\mathrm{E}}(Y_e)$ the mean of Y conditional on X_e. The first difference is simply $\tilde{\mathrm{E}}(Y_e) - \tilde{\mathrm{E}}(Y_s)$.

Repeat the first difference algorithm $M = 1000$ times to approximate the distribution of first differences. Average these simulated values to obtain a point estimate compute the standard deviation to obtain a standard error or sort the values to approximate a confidence interval.

We offer two brief examples of first differences. In our log-linear regression example increasing Democratic control from half to two-thirds of the lower house tended to raise government employment by 7000 people on average. The 95 percent confidence interval around this first difference ranged from 3000 to 12000 full-time employees. (This result seems well worth following up as to the best of our knowledge the relationship has not been studied in the state politics literature.) In our logit example a rise in education from

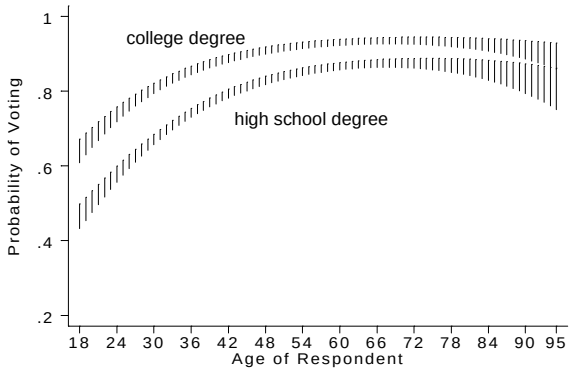

Figure 2: Probability of Voting by Age: Vertical bars indicate 95 percent confidence intervals.

12 to 16 years typically boosed the probability of voting by 0.086Γwith a 99% confidence interval running from 0.078 to 0.094.

6 Tricks of the Trade

The algorithms in Section 5 apply for all statistical modelsΓbut they can be made to work better by following a few tricks and avoiding some common misunderstandings.

FirstΓas discussedΓwe should assess the precision of any quantity computed by simulation by repeating the entire algorithmΓand seeing if anything of substantive importance changes. If something does changeΓincrease the number of simulations and try again. If you are using the expected value algorithmΓbe sure to adjust both m and M. Numbers should be reported to the correct level of precisionΓso for example if repeated runs with the same number of simulations produce an estimate that changes only in the fourth decimal pointΓthen (assuming this is sufficient for substantive purposes) the number reported should be rounded to two or three decimal points.

One common mistake is to exclude some parameters in the step where estimation uncertainty is simulated. Parameters have different logical statusesΓsuch as the effect parameters β vs the ancillary parameters αΓbut the statistical algorithm does not distinguish between the two: both are uncertain and so both should be simulated — even if only one turns out to be of use in later calculations. A variety of other tricks exist to speed up various proceduresΓsuch as making calculations to avoid simulating some of the relevant parametersΓbut for the vast majority of applications these are unnecessary and risk mistakes in implementation. We sacrifice nothing of importance (except a few computer cycles)Γrisk fewer mistakesΓand have an easier time if the rules above are followed without deviationΓexcept when necessary.

Another trick is to make sure that you always calculate simulations of Y firstΓand then calculate your quantity of interest from those. Another way to verify that no mistake is madeΓif this procedure of always simulating Y first is followedΓis to be sure that all estimated parameters are simulated and all simulated parameters are used somewhere in the calculations. We violated this rule when computing expected values by using $\tilde{\pi}$ directly (because drawing dichotomous Y's and averaging would yield exactly $\tilde{\pi}$). If you are unsureΓdon't stop until you have simulations of your outcome variable.

If some function of YΓsuch as $\ln(Y)$Γis used for the analysisΓthe inverse function

$\exp(Y)$ can be applied after the simulations are drawn to reveal Y since $\exp(\ln(Y)) = Y$. We adopted this procedure in the log-linear regression example in Section 5.1. This is important since the usual ad hoc procedure of calculating the predicted value of $\widehat{\mathrm{E}(\ln(Y))}$ without simulation and then exponentiating is inappropriate since $\mathrm{E}[\ln(Y)] \neq \ln[\mathrm{E}(Y)]$. With simulation both $\mathrm{E}(Y)$ and Y can be computed easily whatever scale of estimation the researcher happens to use.

Statistical programs generally report point estimates for the parameters along with standard errors. However we need point estimates and the full variance matrix. (Simulating each parameter independently is incorrect since it misses the covariances.) Virtually all good statistical programs can report the full variance matrix but almost all require you to ask for it explicitly such as by setting some global variable or option.

In some statistical models elements of γ are orthogonal and so separate variance matrices for each set are provided. To follow the procedure above it may be easiest to create a bloc diagonal matrix with the separately estimated variance matrices on the diagonal (i.e. filled in with zeros elsewhere) and then proceed with the algorithms outlined in Section 5. Obviously if the subsets of γ are really orthogonal equivalent draws from the two sets can be made from independent multivariate normal distributions but to avoid mistakes its easier not to deviate from our algorithm.

An important trick is to use reparameterizations of elements of γ so that the asympotic multivariate normal approximation for estimation uncertainty is more likely to hold in finite samples. In general all parameters should be reparameterized unless they are already unbounded and logically symmetric as a Normal must be. For example a variance parameter like σ^2 must be greater than zero and so it will generally pay to reparameterize by using an expression like $\sigma^2 = e^{\eta}$. This allows us to estimate η which is on the scale from $-\infty$ to ∞ as one of the elements of γ which is assumed to be multivariate normal. When making reparameterizations of course we add an extra step to the algorithms above: after drawing γ from the multivariate normal we reparameterize back to the original scale (by computing $\tilde{\sigma}^2 = e^{\tilde{\eta}}$).[4]

Several other commonly used reparameterizations come in handy too. For example a correlation parameter ρ ranging from -1 to 1 can be reparameterized to η (reusing the same symbol) on an unbounded scale with the inverse of Fisher's Z transformation:

$$\rho = \frac{e^{2\eta} - 1}{e^{2\eta} + 1}$$

as used in King (1997: 136 201). Or a parameter representing a probability π can be made unbounded using the logistic transformation $\pi = [1 + e^{-\eta}]^{-1}$. These and other tricks should enhance the effectiveness of simulation.

In certain unusual instances be aware that the number of simulations required to accurately approximate an expected value may be larger than normal. In our experience this has always resulted when the researcher has badly misspecified a nonlinear statistical model. As a rough test of whether the algorithm is providing correct estimates you can compare the mean simulated expected value with the calculation using the point estimates $\hat{\theta} = g(X_c, \hat{\beta})$. Even better increase both m and M and run the algorithm again.

[4] Reparameterization also makes likelihood maximization algorithms easier to use by avoiding problems caused by the optimization procedure choosing inadmissable parameter values (which often result in the program terminating abnormally because of attempts to divide by zero or take the log of negative numbers). Since maximum likelihood estimates are invariant to reparameterization, the reparameterization has no effect except on the finite sample distribution around the point estimate. For example, estimating $\hat{\sigma}^2$ directly gives the same maximum likelihood estimate as estimating $\hat{\eta}$ and transforming to $\hat{\sigma}^2$ by using $\hat{\sigma}^2 = e^{\hat{\eta}}$.

The simulation procedures given in this paper span virtually all quantities that might be of interest and all statistical models that scholars might wish to interpret. As such they are in many ways canonical methods of simulation. However *many* other simulation algorithms are available in the context of specific models to speed the approximation or to make it more accurate for a fixed number of simulations. When designing computer programs for general use these special algorithms should be used as they will often improve on our suggestions. Fortunately we can still look to the algorithms presented here to understand what is going on since as long as M and m are large enough all the algorithms should give identical answers (see Mooney 1993; Tanner 1996).

7 Empirical Examples

We include here replications of three empirical works but instead of choosing the most egregious (or even the median which isn't far from the most egregious) we choose a large number of the best works from our most prestigious journals and presses written by some of our most distinguished authors. Within this group we focused on recent works eliminated the (many!) publications we were unable to replicate and then picked three to illustrate a diverse range of models and interpretative issues. The procedures for model interpretation in all three were exemplary. If we all followed their examples reporting practices in the discipline would be greatly improved. For each article we describe the substantive problem posed and statistical model chosen; we also accept rather than evaluate their statistical procedures. We then detail how the authors interpreted their results and demonstrate how our procedures advance this state of the art.

7.1 A Time Series-Cross-Sectional Model

Advanced industrial democracies are on average more exposed to international commercial and capital flows today than at any time in the past 100 years. Conventional wisdom holds that this globalization of markets has compelled governments to slash public spending but a new book by Geoffrey Garrett (1998) offers evidence to the contrary. Where strong leftist parties and encompassing trade unions have coincided Garrett argues globalization has led to *greater* government spending as a percentage of GDP whereas the opposite has occurred in countries where the left and labor have been weak.

Garrett constructed a panel of economic and political variables measured annually for fourteen industrial democracies during the period 1966–1990. He then estimated a linear-normal (least squares) regression model (see Equations 2) where the dependent variable Y_i is government spending as a percentage of GDP for each country-year in the data set. The three key explanatory variables were Capital mobility C_i (higher values indicate fewer government restrictions on cross-border financial flows) Trade T_i (larger values mean more foreign trade as a percentage of GDP) and left Labor power L_i (higher scores denote a stronger combination of leftist parties and labor unions). Garrett also focused on two interactions among the variables: $C_i L_i$ and $T_i L_i$. Finally he included a battery of business cycle and demographic controls as well as the lagged level of government spending and dummy variables for countries and time periods.

To interpret his results Garrett computed selected predicted values by making

> a series of counterfactual estimates of government spending under different constellations of domestic political conditions and integration into global markets. This was done by setting all the other variables in the regression equations

left-labor power	trade: low	trade: high	capital mobility: low	capital mobility: high
low	43.1	41.9	42.8	42.3
high	43.5	44.2	43.1	44.5

Table 2: *Garrett's Counterfactual Effects on Government Spending (% of GDP): Each entry is a predicted value of government spending for given values left-labor power and trade or capital mobility, holding all other variables constant at their means.*

> equal to their mean levels and multiplying these means by their corresponding coefficients and then by examining the counterfactual impact of various combinations of left-labor power and globalization ... (p. 82).

In particular, Garrett distinguished beween *low* and *high* levels of L_i, T_i, and C_i. For these variables, the 14th percentile in the dataset represented a low value, whereas the 86th percentile represented a high one.[5] The counterfactual estimates that Garrett calculated appear in Table 2. Garrett used the counterfactuals to draw three conclusions. First, "government spending was always greater when left-labor power was high than when it was low, irrespective of the level of market integration." (Entries in the second row in each table have larger values than the first.) Second, "where left-labor power was low, government spending decreased if one moved from low to high levels of market integration, but the converse was true at high levels of left-labor power." (See the table on the left.) Finally, "the gap between the low and high left-labor power cells was larger in the high trade and capital mobility cases than in the cells with low market integration," (differences in the entries in the second column of each table are larger than in the first) implying that "partisan politics had more impact on government spending where countries were highly integrated into the international economy than in more closed contexts" (p. 83).

Garrett's counterfactuals go far beyond the customary list of coefficients and t-tests, but our tools can help us extract even more information from his model and data. For instance, simulation can reveal whether the differences in values across the cells might have arisen by chance alone. To make this assessment, we re-estimated the parameters in Garrett's regression equation[6] and drew 1000 sets of simulated coefficients from their posterior distribution, using the algorithm in Section 5.1. Then we fixed L_c and T_c at their 14th percentiles, held other variables at their means, and calculated 1000 (counterfactual) expected values, one for each set of simulated coefficients. Following the same procedure, we produced counterfactuals for the other combinations of L_c, T_c, and C_c represented by the cells of Table 2.

Finally, we plotted "density estimates" (which are smooth versions of histograms) of the counterfactuals; these appear in Figure 3. We can think of each density estimate as a

[5] "So as not to exaggerate the substantive effects" of the relationships he was studying, Garrett "relied on combinations of the 20th and 80th percentile scores" (p. 82). Unfortunately, due to a minor arithmetic error, the values he reports (p. 84) correspond only to the 14th and 86th percentiles. To facilitate comparison with Garrett, we use the 14th and 86th percentiles in our simulations.

[6] Our estimated coefficients differed from those in Garrett (1998, pp. 80-81) by only 0.3 percent, on average. Standard errors diverged by 6.8 percent, on average, apparently due to discrepancies in the method of calculating panel-corrected standard errors (Franzese, 1996). None of the differences made any substantive difference in the conclusions.

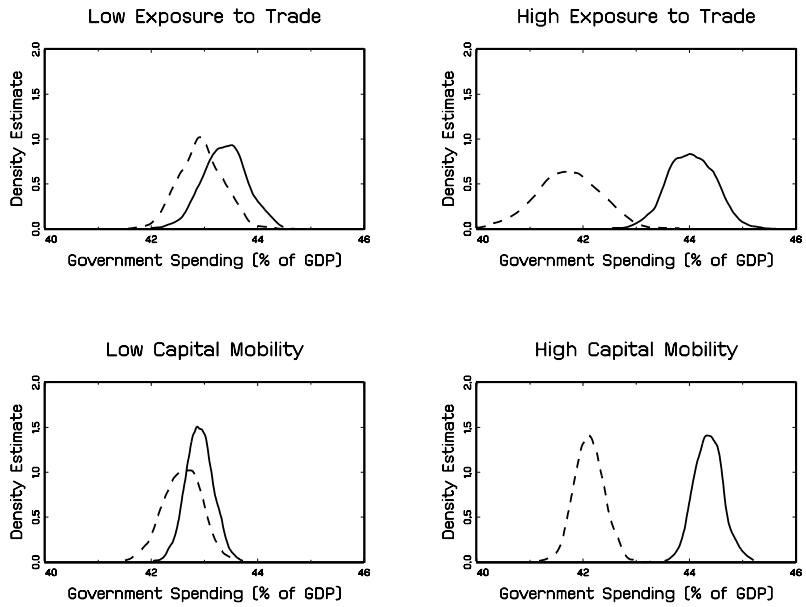

Figure 3: *Simulated Levels of Government Spending: These panels contain density estimates (smooth versions of histograms) of expected government spending for countries where left-labor power is high (the solid curve) and low (the dotted curve). The panels, which add uncertainty estimates to the concepts in Table 2, demonstrate that left-labor power has a distinguishable effect only when exposure to trade or capital mobility is high.*

pile of simulations distributed over the value government spending. The larger the pile for each point on the curveΓat any given level of government spendingΓthe more simulations took place near that point.

Figure 3 shows that when globalization of trade or capital mobility is lowΓleftist governments spend only slightly more than rightist ones; the density estimates overlap so thoroughly that it is difficult to distinguish the two spending patterns with much confidence. That isΓrandom drawsΓand thus presumably real outomesΓfrom one distribution will not be easily distinguished from the other. In the era of globalizationΓby contrastΓ domestic politics exerts a powerful effect on fiscal policy: leftist governments outspend rightist ones by more than two percent of GDP on averageΓa difference we can affirm with great certaintyΓsince the density estimates for the two regime-types are far apart. In summaryΓour simulations temper Garrett's claim that left-labor governments always outspend the rightΓregardless of the level of market integration: the tendency is correctΓ but there is much variability. At the same timeΓthe simulations support Garrett's central claim that globalization has intensified the relationship between partisan politics and government spending.

Simulation also can provide insight into the *dynamics* of government spending. Each cell in Table 2 contains a static quantityΓthe expected level of spending for a given configuration of globalization and domestic politicsΓbut how might a permanent one-time change a key explanatory variable affect the trajectory of spending over many years? The algorithms desccribed here can provide the answer. (See Alt and Lowry (1994: Appendix B) for a similar simulation in the context of a 3SLS model.)

To see how a permanentΓone-time increase in exposure to trade from low to high

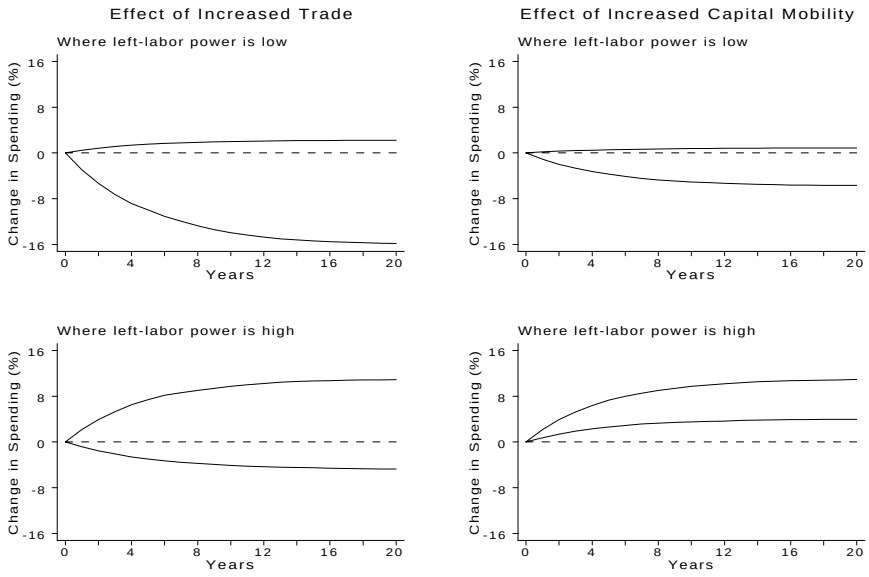

Figure 4: *Simulated Dynamics of Government Spending: These graphs depict the 95 percent confidence intervals (solid lines) in government spending in response to a permanent one-time increase (at time 0) in globalization. The dotted line denotes 0% increase in spending.*

would affect the path of goverment spending over a 20-year periodΓwe used the algorithm for simulating first differences (see Section 5.3). FirstΓwe established our "starting and ending points" by fixing left-labor power and trade at low levels (L_{low} and T_{low}) and setting other variables at their means. We then simulated 1000 expected values for the level of government spending in "year 1". These simulations became our lagged dependent variables for 'year 2Γ and the 1000 simulations for year 2 became the lags for year 3. By repeating this process for two decadesΓone simulation at a timeΓwe generated 1000 plausible trajectories for government spending under the assumption that left-labor power and exposure to trade remained low.

To complete the first-difference algorithmΓwe changed the value of one explanatory variableΓexposure to tradeΓfrom T_{low} to T_{high} and re-ran the algorithmΓbeginning at year 1. This resulted in 1000 paths for spending consistent with our data in a country where left-labor power was low but trade exposure was high. To estimate the impact of a permanent one-time increase in exposure to tradeΓwe subtracted this second set of trajectories (based on T_{high}) from the first set (based on T_{low}). This yielded 1000 trajectories for the cumulative change in spending over 20 years in response to an increase in trade. For any given yearΓthe mean value of these trajectories represented our best estimate of the cumulative change that had taken place so far. Figure 4 gives the 95 percent confidence intervals. Repeating this algorithm generated the four panels in the figure.

The panels show that left-labor and rightist governments embark on distinct spending trajectories in response to globalizationΓbut most of the difference is due to capital mobility rather than higher trade. Where left-labor power is highΓan increase in capital mobility from low to high will raise government spending as a percentage of GDP by 4.9 percentage points over five years and 7.4 percentage points over 20. The 95% confidence interval around this trend is well above the (dotted) zero-lineΓso we can be very certain that the

change in spending will be positive. The panels inspire less confidence in the proposition that left-labor governments will boost spending in response to foreign trade. The 95% confidence intervals in the lower left-hand plot span both sides of the zero line, suggesting that in a large fraction of cases spending could *fall* in the context of increasing exposure to international commerce. The plots therefore remind us that globalization is not a monolithic phenomenon, and that each component of globalization deserves individual attention.

7.2 Multinomial Logit Models

How do citizens in a country ruled for years by one party vote when there appears to be an opportunity to turn that party out of office? Jorge Domínguez and James McCann (1996) address this question with regard to Mexico, arguing that during the 1988 presidential election Mexicans were focused on the fate of the ruling party, and only secondarily concerned with policy issues. They find that opinions on the parties, the presidency, and the presidential candidates retain the most explanatory significance.

One method with which they investigate these questions is a multinomial logit analysis of survey data from the 1988 presidential election. The model, which has multiple equations, fits into the general framework of Equations 1 by letting Y_i and π_i each be 3×1 vectors:

$$\begin{aligned} Y_i &\sim \text{Multinomial}(\pi_i) \\ \pi_i &= \frac{e^{X_i \beta_j}}{\sum_{k=1}^{3} e^{X_k \beta_j}} \qquad \text{where } j = 1, 2, 3 \text{ candidates.} \end{aligned} \tag{5}$$

where X_i is the vector of explanatory variables for observation i, and β_1, β_2, and β_3 each are $k \times 1$ vectors of coefficients. the three candidates are Carlos Salinas (from the ruling PRI), Manuel Clouthier (representing the PAN, a right-wing party), and Cuauhtémoc Cárdenas (head of a leftist coalition). The survey data used to estimate this model consists of a battery of demographic, economic, and political questions on a national sample of 1,359 people (after removing missing values), taken before the 1988 presidential election. 1988 was electorally significant in Mexico, because for the first time all three presidential candidates appeared to be highly competitive.[7]

They argue that "voters reasoned through two separate electoral decisions: a negative judgement on the 'party of the state', and then a choice of an opposition vehicle" (p. 112). They further argue that a minority of Mexicans exhibit strategic voting behavior, voting for a candidate other than their first choice if it means increasing the likelihood of the ruling party's defeat. To explore their hypotheses about this two-step electoral decision they compute hypothetical probabilities of voting for the two opposition parties for two types of voters:

> One we call a left-wing voter. Such a person would come from the working class and would belong to a union, would never attend church, and would oppose

[7] The model includes 31 explanatory variables: Previous PRI voter, Previous PAN voter, Not previously mobilized, Favor capital punishment, Increase foreign investment, Limit imports, Pay foreign debt, Retain state industries, Political interest, Preference for "strong leaders", Current national economy, Future national economy, Current personal finances, Future personal finances, Education level, Age, Female, Professional class, Working class, Union member, Population of city/town, North, South, Federal District, Church attendance, PRI getting stronger, Economy–other party, Social unrest–other party, Presidential approval, Personality trait index: Clouthier, Personality trait index: Cárdenas. See Domínguez and McCann (1996: 213-216) for question wording.

	Prefer Clouthier (PAN)	Prefer Cárdenas
Voters with right-wing dispositions		
PRI becoming weaker	.51	.17
	(.34Γ68)	(.09Γ29)
PRI becoming stronger	.07	.03
	(.03Γ15)	(.01Γ06)
Voters with left-wing dispositions		
PRI becoming weaker	.07	.65
	(.02Γ15)	(.45Γ81)
PRI becoming stronger	.01	.13
	(.003Γ03)	(.05Γ26)

Table 3: *Ideological Predispositions and Beliefs about the PRI's Long-Term Viability in 1988: Probabilities of Voting for Opposition Candidates. Cell entries are probabilities, a replication of Domínguez and McCann (1996: Table 4–16, p. 111). Numbers in parentheses are 95% confidence intervals we computed by simulation.*

> foreign investmentΓ freer tradeΓ continued payment of the foreign debtΓ and privatization of state enterprises. The other we call a right-wing voter. Such a person would come from the professional class and not belong to a unionΓbut would attend church once a week and favor the four policies opposed by the left-wing voter (pp. 110–111).

Each of these two population types is further divided into those who believed the PRI was becoming weakerΓ and those who believed it was becoming stronger. They get the expected values in Table 3Γand interpret these results as follows (p.112):

> Among voters ready to live without the PRIΓ thereforeΓ the PAN benefitted substantially among right-wing voters and lost considerably among left-wing voters. Cárdenas benefitted enormously among left-wing voters but lost more than half his vote among right-wing voters. NonethelessΓ Cárdenas's support among right-wing voters was more than double the PAN's support among left-wing voters; this suggests strategic voting by right-wing voters ready to support Cárdenas because he seemed more likely than Clouthier to beat Salinas. In 1988Γ among voters ready to replace the party in powerΓ issues and demographic cleavages mattered. Provided voters thought that the PRI was becoming weaker and that the economy and social peace were not in danger if an opposition party were to winΓthen such voters were much more likely to support the opposition.

The probabilities (expected values) in Table 3 that Domínguez and McCann computed are far better than the usual presentation of multinomial logistic resultsΓbut we now have the tools to go farther. For one exampleΓwe can compute confidence intervals for these predicted values. To do thisΓwe first replicated their model and applied the algorithm of Section 5.2 for each of their four population types. That isΓfor each of these population types we fixed the values of the ideological predisposition and PRI strength variables as indicatedΓheld other variables constant at their meansΓdrew 1000 simulated coefficients (of β_1Γ β_2Γand β_3)Γand computed 1000 simulations of the 3×1 vector π_i (representing

the probabilities of voting for each of the three candidates) by applying the second line of Equation 5 for $j = 1$ and $j = 2$. π_3 is computed using $\pi_3 = 1 - \pi_1 - \pi_2$. Then we computed 95% confidence intervals for each population type by sorting the 1000 values and taking the 25th and 976th values for each of the three candidates. The confidence intervals for the two opposition candidates are also displayed in Table 3 in parentheses beneath the expected value. These uncertainty estimates provide some interesting clarifications. For example among right-wing voters who think the PRI is becoming stronger there is statistically no difference between support for Clouthier and support for Cárdenas as the confidence intervals overlap.

With our new tools we can extract more information from the model than just expected values and confidence intervals. For example we can use *ternary diagrams* to display the three-party vote probabilities simultaneously in a triangle in two dimensions (see Miller 1977; Katz and King 1997). Roughly speaking the closer a point appears to one of the vertices the larger the vote probability is for that candidate whose name appears at that vertex. A point that appears at the middle of the triangle indicates probabilities split one-third for each of the candidates; points that appear on a vertex are 100% for that candidate and 0% for the other two.

For example the two panels of Figure 5 show for left-wing and right-wing voters respectively how the probabilities of voting for each of the three candidates change depending on whether they think the PRI (Salinas's party) is strengthening or weakening. Each arrow represents the expected value for one individual under these two conditions and leaving other explanatory variables set at their observed values. The base of each arrow indicates the predicted voting probability of an individual set to think that the PRI is strengthening whereas the arrowhead shows the voting probabilities for the same individual set to think the PRI is weakening. Thus the set of arrows help explore what would happen if the entire electorate moved in the direction most voters seem to be heading. (For graphical clarity but without changing any patterns we show a 10% random sample of the original data.)

In both ternary diagrams most of the arrows begin in areas of the diagram indicating support for Salinas (i.e. close to the Salinas vertex). Opinion about the ruling PRI suppresses any left-right divisions within the population. Once voters think that the PRI is weakening the underlying ideologically-based voting preferences emerge which can be seen by the large movement and diverse opinions at the end of the arrow heads. The apparent strategic voting behavior of some right-wing voters is evident in panel (b). The "natural" candidate of choice for right-wingers would be Clouthier and yet a fair number of these individuals who thought the PRI was weakening actually said they preferred Cárdenas the candidate thought to have the best chance of ousting Salinas. These are the arrows that start near the Salinas vertex and point toward the Cárdenas vertex. Strategic behavior of left-wing voters can not however be detected in panel (a) since for these voters the natural candidate was Cárdenas. Figure 5 does support Domínguez and McCann's conclusion that "Cárdenas benefitted enormously among left-wing voters." With a few exceptions left-wing voters who thought the PRI was weakening supported Cárdenas as can be seen in panel (a) from the arrows that lie along the axis running from Salinas to Cárdenas.

Domínguez and McCann focus primarily on voting behavior as is traditional in survey research. But their question also is related to the election outcome. We extend their ideas by simulating nationwide *electoral outcomes* — the fraction of the vote each candidate receives — conditional only on whether all voters believe the PRI is weakening or strengthening. We thus generated 100 simulated elections for each of these two situations.

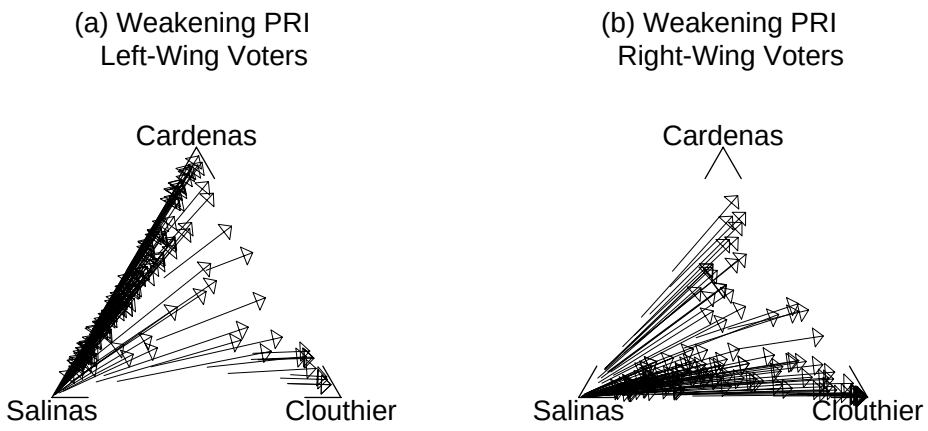

Figure 5: *Candidate Choice by Beliefs about the PRI. The base of each arrow gives the predicted probability of voting for each of the three parties (portrayed as the proximity to the three labeled vertices) for an individual who believes that the Salinas' party is weakening, with other variables held at their observed values. The end of the arrow is the probability that the same individual would be predicted to have if that person believed Salinas' party was strengthening. For clarity, a random 10% of the observations appears in this graph.*

To do this for a given population type we applied the algorithm to generate expected values for the 3×1 vector Y_i (one of which equals 1 the other two 0) in Section 5.2 holding each voter's other characteristics as they are except that we set the PRI strength variable as indicated. For each person in the sample we drew one simulated vote and then counted the fraction of simulated votes each candidate received to produce one simulated election outcome. We repeated this 100 times generating 100 simulated election outcomes (i.e. estimates of what would happen when repeating the election with the same campaign and voters).[8]

The results are displayed in the ternary diagram in Figure 6. Unlike the last coordinates in this figure represent predicted fractions of the vote received by each candidate under a different simulated election outcome (instead of probabilities for individual voters). Simulations represented by "."s (all near the middle) are electoral outcomes where everyone in the sample is coded to think the PRI is weakening; the "o"s (all on the bottom left) are simulated outcomes where everyone thinks the PRI is strengthening but in both cases characteristics measured by other variables are held constant at their observed values. We also added "win lines" to the figure that divide the ternary diagram into areas that indicate which candidate receives a majority and thus wins the simulated election (e.g. points that appear in the top third of the triangle are simulated election outcomes where Cárdenas receives a plurality).

Figure 6 demonstrates that when the country believes the PRI is strengthening Salinas wins hands down In fact he wins every one of the simulated elections (indicating that the probability that he would win an actual election under these conditions is greater than 0.99). However if every voter believed that the PRI was weakening the election

[8]Mexico's proportional representation system means that aggregation from a probability sample is relatively straightforward. For expository purposes, we ignore issues of sample design that may affect aggregate estimates like these.

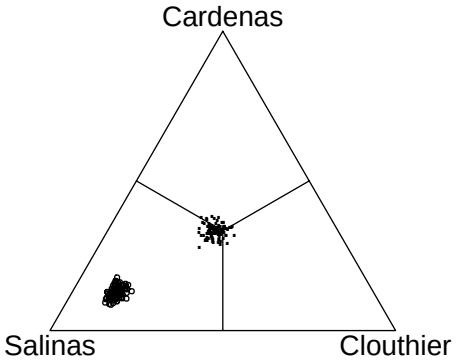

Figure 6: *Simulated Electoral Outcomes: Coordinates in this ternary diagram are predicted fractions of the vote received by each of the three candidates. Each point is an election outcome drawn randomly from a world in which all voters believe Salinas' PRI party is strengthing (for the "o"'s in the bottom left) or weakening (for the "."'s in the middle), with other variables held constant at their means.*

outcome would be essentially a tossup with each of the three parties having about an equal chance of victory. Thus Salinas's PRI would still have about a 1/3 probability of winning. This must be a sobering thought for those hoping to bring an end to one party politics in Mexico. Hope of defeating the PRI even under these optimistic conditions clearly requires some kind of compromise between the two opposition parties or changing conditions other than beliefs about party strengths. The figure also supports the argument that despite much voter fraud Carlos Salinas probably did win the presidency in 1988. He may well have won it by a lesser margin than reported but Figure 6 is strong evidence that he probably did in fact defeat a divided opposition.[9]

7.3 Censored Weibull Regression Models

How does going to war affect the survival of political leaders? Bruce Bueno de Mesquita and Randolph Siverson (1995) investigate this question with a data set on international wars between 1816 and 1975. They find that leaders who take their states into war risk their own political power at home. Longer pre-war leadership mitigates this risk but more so for authoritarian leaders than for democratic ones. Defeat and high war costs increase the risk of removal for all leaders. Authoritarians are found to fight longer wars than democrats and also select wars to participate in that have a higher risk of defeat.

Bueno de Mesquita and Siverson evaluate their argument using censored Weibull regression a form of duration model on a data set consisting of 191 cases of state war participation between 1823 and 1974. For fully observed cases (i.e. cases where the leader had left office at the time of the study) the model for Y_i years of duration in office after the onset of war is

$$\begin{aligned} Y_i &\sim \text{Weibull}(\mu_i, \sigma) \\ \mu_i &\equiv E(Y_i|X_i) = \left(e^{X_i\beta}\right)^{-\sigma} \Gamma(1+\sigma) \end{aligned} \tag{6}$$

where i is an index over leaders and σ is an ancilliary (shape) parameter. (Γ is the gamma function an interpolated factorial that works for continuous values of its argument.) There

[9] See Scheve and Tomz (in press) on simulation of counterfactual predictions and Sterman and Wittenberg (in press) on simulation of fitted values, both in the context of binary logit models.

are four explanatory variables in X_iΓT_i (the log of Tenure in office prior to the onset of warΓin years)ΓD_i (the log of pre-war tenure just for Democratic leaders)ΓB_i (the log of Battle deaths per 10Γ000 population)Γand W_i (a dummy variable indicating whether the war was Won or not).

Bueno de Mesquita and Siverson discuss the marginal impact of their explanatory variables by computing the "hazard rate" associated with each variable. Hazard rates are the traditional method of interpretation in the literatureΓbut understanding them requires a lot of statistical knowledge. They find that an increase in the length of pre-war tenure for authoritarian leaders increases leader survivabilityΓbut "similar tenure for democratic leaders produces no siginficant benefit in survival . . . . Thus regime type makes an appreciable difference in the prospects of surviving a war politicallyΓwith democratic leaders placed at considerably higher risk than their authoritarian counterparts."[10]

Simulation can help here in several ways. For exampleΓwe can answer a more direct question than one about the hazard rate: How many more months will a leader survive in office given (say) a ten year increase in his or her pre-war tenure? To accomplish this we simulate survival time in office for democrats with median pre-war tenureΓfor authoritarians with median pre-war tenureΓand for each again with these median pre-war tenures increased by ten years. For each case we set the relevant explanatory variablesΓand held the other variables constant at their means. We then use the algorithm in Section 5.1 to generate 500 simulated durationsΓwhich we then plot as a density estimate (a smooth version of a histogram) of years of survivalΓin each case.

These results appear in panels (a) and (b) of Figure 7. In each plot the dotted curve represents the density of our simulated values of survival in office for leaders with median pre-war tenure. The solid lines represent the densities if pre-war tenure is increased by ten years.[11] The arrows in the graphs pointing to the axis indicate median values of survival.

Figure 7 supports Bueno de Mesquita and Siverson's proposition that pre-war tenure matters more for authoritarians than it does for democrats. This can be seen by noting that the medians are further apart on the authoritarian panel (a) than on the democratic panel (b). A ten year increase in pre-war tenure for authoritarians clearly yields a larger gain in survival than a similar increase for democratic leaders. But we can also add some numerical precision: the gain is only 2.8 years on average for democrats (which have a mean duration of 5.6 minus 2.8)Γwhereas authoritarians last 11.8 years longer (15.1 minus 3.3). This figure is also a good illustration of a case where the entire distribution of the quantity of interest yields insights that the point estimates alone do not. For exampleΓ because of the asymmetries in the distributionsΓthe modal survival time (the top of each distribution) does not closely correspond to the more interesting median survival time. The exact nature of the dramatic skewness in the distribution is also importantΓin that it shows clearly how most of the survival times are relatively short (under 5 years) and highly clusterΓwith longer times distributed over a much wider range (5–20 years and more).

By judging significance tests for the differences between coefficients in the Weibull regression modelΓBueno de Mesquita and Siverson conclude that pre-war tenure does not affect the survivability of democrats in office. HoweverΓdifferences in these coefficients do not directly translate into differences in survival timesΓsince the mapping between the two is highly nonlinear. Simulation allows us to investigate in a more intuitive way whetherΓ

[10] Although it does not appear to make a big difference substantively, Bueno de Mesquita and Siverson's specific numerical calculations turn out to be incorrect except when $\sigma = 1$.

[11] Large numbers of random draws from a Weibull tend to yield a few very large values. We include these points in the calculation of the densities, but for clarity truncate the graph at 20 years.

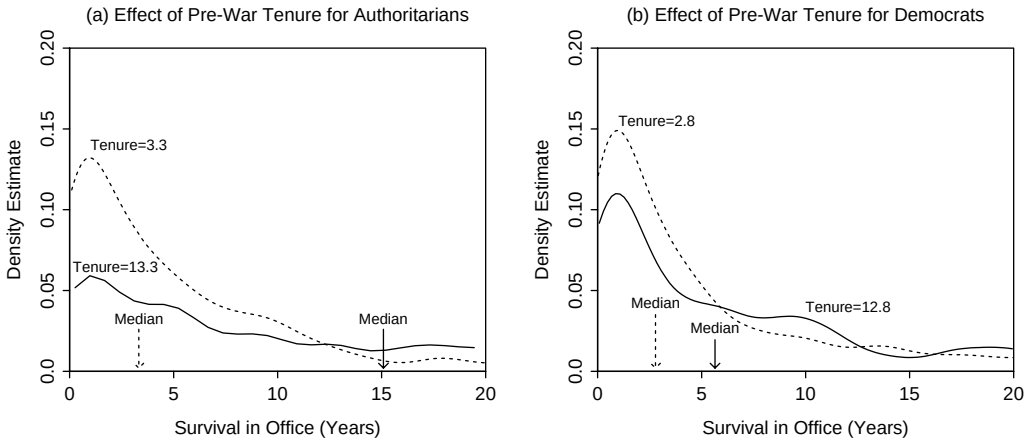

Figure 7: *Regime Type and and Political Survivability in Wars: Density estimates of the number of years of survival in office for (a) authoritarian and (b) democratic leaders with median pre-war tenure (dotted line) and long pre-war tenure (solid line).*

on averageΓpre-war tenure matters for democratic leaders. To do this we generated data for 500 sets of graphs like Figure 7Γand then calculated the difference in medians for both the democratic and authoritarian leaders. We then plot the density estimate of these differences in Figure 8.

Contrary to Bueno de Mesquita and Siverson's conclusionΓour representation of their results shows that pre-war tenure does matter somewhat for democratic leaders: on averageΓa ten year increase in tenure enhances the leader's survivability by approximately two years (the mode of the "Democrats" curve). HoweverΓthis graph also supports their main pointΓnamelyΓthat authoritarian leaders benefit much more from pre-war tenure than do democratic ones.

8 Concluding Remarks

Our proposals for extracting new information from existing statistical models should enable scholars to interpret and present their results in ways that convey numerically precise estimates of the quantities of substantive interestΓinclude reasonable assessments of uncertainty about those estimatesΓand require no specialized knowledge to understand. The methods we propose are obviously more onerous than the methods currently used in political science. They require more computationΓand researchers who put them into practice will have to think much harder about exactly how to translate their findings into quantities of interest to a wider audience. But implementing them may improving current empirical research. Our approach may even help bridge the acrimonious and regrettable chasm that often separates quantitative and non-quantitative scholarsΓand make the fruits of statistical research accessible to all who have a substantive interest in the issue under study.

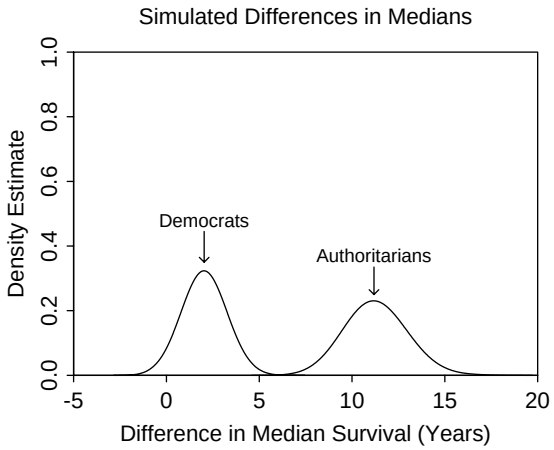

Figure 8: *Regime Type and and Political Survivability in Wars: Density estimates of the difference in median survival time in office for authoritarian and democratic leaders with average pre-war tenure.*

Appendix: Software

Although understanding the principles behind our method requires some extra effortΓ we have written very easy-to-use statistical software to implement the method in practice. This softwareΓa set of macros for use with the Stata statistics package which we call *CLARIFY: Software for Interpreting and Presenting Statistical Results*Γwill calculate quantities of interest for the most commonly used statistical modelsΓincluding linear regressionΓbinary logitΓbinary probitΓordered logitΓordered probitΓmultinomial logitΓ Poisson regressionΓnegative binomial regressionΓand a growing number of others. The software and detailed documentation are available at `http://GKing.Harvard.Edu`. We provide a brief description here.

The package includes three macros that are intended to be run in this order:

`ESTSIMP` estimates a chosen model and generates random draws from the multivariate normal distribution (i.e. computes $\tilde{\gamma}$)Γas we describe in Section 5.

`SETX` sets X_c to desired values such as meansΓmediansΓpercentilesΓminimaΓmaximaΓ specified valuesΓmixtures of theseΓand others.

`SIMQI` computes desired quantities of interest such as *predicted values* (Section 5.1) and *expected values* (Section 5.2).

These programs come with many optionsΓbut to show how easy they can be to useΓ we provide one brief example. Suppose we have an ordered probit modelΓin which the dependent variable `y` takes on the values 1Γ2Γ3Γ4Γor 5Γand the explanatory variables are `x1` and `x2`. Suppose we want to find the probability that `y` has the value 4 when `x1`= 12.8 and `x2` is fixed at its meanΓand want a 90% confidence interval around that probability. To generate this quantity of interestΓwe would type the following 3 commands from the Stata command prompt:

```
estsimp oprobit y x1 x2
setx x1 12.8 x2 mean
```

```
simqi, prval(4) level(90)
```

The first line estimates the ordered probit model of `y` on `x1` and `x2`and generates and stores simulated values of all estimated parameters. The second line sets `x1` to 12.8 and `x2` to its mean. The third line computes the desired quantity of interest (a probability value of 4) and the (90 percent) level of confidence associated with an interval to be computed around it.

These programs are very flexible and will compute many more quantities of interest than included in this brief example. The on-line help gives detailed descriptions. We invite others to write us with contributions to this set of macros to cover additional statistical models or other quantities of interest. We also plan to continue to add to them.

References

Alt, James E. and Robert C. Lowry. 1994. "Divided Government, Fiscal Institutions, and Budget Deficits: Evidence from the States." *American Political Science Review*, 88, 4 (December): 811–828.

Barnett, Vic. 1982. *Comparative Statistical Inference*. 2nd ed. New York: Wiley.

Bueno de Mesquita, Bruce and Randolph M. Siverson. 1995. "War and the Survival of Political Leaders: A Comparative Study of Regime Types and Political Accountability." *American Political Science Review*, Vol. 89, No. 4, (December): 841-855.

Domínguez, Jorge I. and James A. McCann. 1996. *Democratizing Mexico: Public Opinion and Electoral Choices*. Baltimore: The Johns Hopkins University Press.

Franzese, Jr., Robert J. 1996. "PCSE.G: A Gauss Procedure to Implement Panel-Corrected Standard-Errors in Non-Rectangular Data Sets." (Cambridge, MA: Harvard University, 3/19/96).

Garrett, Geoffrey. 1998. *Partisan Politics in the Global Economy*. New York: Cambridge University Press.

Gelman, Andrew et al. 1997. *Bayesian Data Analysis*, New York: Chapman and Hall.

King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*, Princeton: Princeton University Press.

King, Gary. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*, New York: Cambridge University Press.

Katz, Jonathan and Gary King. 1997. "A Statistical Model for Multiparty Electoral Data" presented at the annual meetings of the Midwest Political Science Association, Chicago, Illinois, April.

Long, Scott J. 1997. *Regression Models for Categorical and Limited Dependent Variables*, Thousand Oaks, CA: Sage Publications.

Miller, William L. 1977. *Electoral Dynamics in Britain since 1918*. London: Macmillan.

Mooney, Christopher. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference*, Newbury Park, CA: Sage publications.

Rosenstone, Steven J. and John Mark Hansen. 1993. *Mobilization, Participation, and Democracy in America*. New York: MacMillan.

Scheve, Kenneth and Michael Tomz. In press. "Electoral Surprise and the Midterm Loss in U.S. Congressional Elections." *British Journal of Political Science*, forthcoming.

Sterman, John D. and Jason Wittenberg. In press. "Path-Dependence, Competition, and Succession in the Dynamics of Scientific Revolution." *Organization Science.*

Tanner, Martin A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, third edition, New York: Springer-Verlag.

Tufte, Edward R. 1974. *Data Analysis for Politics and Policy.* Englewood Cliffs, NJ: Prentice-Hall.