Annu. Rev. Polit. Sci. 2002. 5:423–50
DOI: 10.1146/annurev.polisci.5.112801.080943




# TOWARD A NEW POLITICAL METHODOLOGY: Microfoundations and ART

Christopher H. Achen
*Department of Political Science and Institute for Social Research, University of Michigan, 4252 ISR, Ann Arbor, Michigan 48106-1248; e-mail: achen@umich.edu*

■ **Abstract** The past two decades have brought revolutionary change to the field of political methodology. Steady gains in theoretical sophistication have combined with explosive increases in computing power to produce a profusion of new estimators for applied political researchers. Attendance at the annual Summer Meeting of the Methodology Section has multiplied many times, and section membership is among the largest in APSA. All these are signs of success. Yet there are warning signs, too. This paper attempts to critically summarize current developments in the young field of political methodology. It focuses on recent generalizations of dichotomous-dependent-variable estimators such as logit and probit, arguing that even our best new work needs a firmer connection to credible models of human behavior and deeper foundations in reliable empirical generalizations.

## INTRODUCTION

Decrying the scientific status of political science has a very long tradition, and not just from outside the discipline. Burgess (1891) condemned the low intellectual standards a century ago, and Bentley (1908, p. 162) shortly thereafter proclaimed, “We have a dead political science.” Catlin (1927, p. 142) found no sign of life a quarter century later: “There is as yet no such thing as a political science in any admissible sense.” The hue and cry has never ceased since.

Almost none of the critics has been entirely wrong. Political science really was too legalistic in the nineteenth century, too bereft of case studies and statistical evidence in the 1930s, too ignorant of survey research and statistical methods in the 1950s, and too resistant to rigorous theory in the1980s.

Even now, much remains to be done on all these fronts. If one puts side by side an introductory physics book, an introductory economics text, and an introductory treatment of the political process, it is difficult to be entirely happy with the current state of the profession. These other fields have serious imperfections and lacunae, but they also possess a broad-gauge, intellectually powerful, logically integrated, well-tested framework to convey to freshmen. We do not.

Methodology has customarily been supposed to be part of the solution. Beginning with Charles Merriam’s Chicago department in the 1920s and 1930s, and continuing in each of the succeeding generations, overcoming stasis and creating the scientific future of the discipline has meant disseminating the newest research techniques. When that has been done, we have always said, then political science will be scientific. We have worked hard, and the dissemination has always been achieved. Indeed, each step made us smarter. But disappointment has always followed. The current era is no exception.

Even at the most quantitative end of the profession, much contemporary empirical work has little long-term scientific value. "Theoretical models" are too often long lists of independent variables from social psychology, sociology, or just casual empiricism, tossed helter-skelter into canned linear regression packages. Among better empiricists, these "garbage-can regressions" have become a little less common, but they have too frequently been replaced by garbage-can maximum-likelihood estimates (MLEs).[1] Beginning graduate students sometimes say, "Well, I don't really understand how these variables relate to each other and the data are bad, but I did use the newest estimator, downloaded from the Internet, and I do report heteroskedasticity-consistent standard errors."

No wonder that a prominent applied statistician, looking recently at one of our more quantitative journals, said (no doubt with a bit of hyperbole), "There is only one item in here I would want to read." He then pointed to an article that was deeply informed about the substance of the problem it addressed but used only cross-tabulations (though it used them intensively and creatively).

Now, fairness requires that a survey of contemporary political methodology acknowledge the field's real achievements. When this author first wrote about the subject nearly 20 years ago (Achen 1983), there were relatively few scholars and accomplishments to report on. Now the field is much too large to cover in an essay, and the statistical sophistication of the discipline has been raised substantially. Although a little flim-flam has emerged to fleece the innocent, so too has much patient and serious development of genuinely new and more powerful statistical tools.

Nevertheless, the present state of the field is troubling. For all our hard work, we have yet to give most of our new statistical procedures legitimate theoretical microfoundations, and we have had difficulty with the real task of quantitative work—the discovery of reliable empirical generalizations. To help the reader see where we stand, the remainder of this essay develops this argument in the context of some recent interesting estimators proposed by prominent political methodologists. Setting aside those statistical proposals that have not stood up to peer criticism, the discussion focuses on some of the best recent work, which demonstrates most clearly what will be needed in the next decades.

The outline of the paper is as follows. First, I review estimators for dichotomous dependent variables, including one of the best-grounded and least-appreciated new estimators of recent years, Nagler's (1994) generalization of logit ("scobit"). This set of estimators is then shown to be easily generalizable beyond scobit to an unmanageably large class. The implication is that creating ever more "generalized" estimators without reference to substantive knowledge, a path we have often

[1]I owe the "garbage-can" epithet to Anne Sartori, who makes no claim of originality.

pursued in recent years, leads political methodology astray. Instead, we need reliable empirical knowledge in order to choose among the many possible estimators that might be used in each of our applications.

Next, the paper argues that little dependable empirical knowledge exists in political science because our conventional work habits squander our efforts. Two remedies are suggested. First, we need to exploit formal theory more often to structure our estimators. Second, when no formal theory is available, we need far more serious data analytic procedures to discipline our specifications.



## GENERALIZING FAMILIAR ESTIMATORS: FROM LOGIT TO SCOBIT

One group of estimators in very wide use in political science is the probit/logit group. Designed for discrete (often dichotomous) dependent variables, these estimators employ special techniques to keep forecasts meaningful. In the dichotomous case, for example, all logit and probit forecasts are probabilities. They never exceed one or fall below zero, as often happens when linear regression is applied to dichotomous dependent variables. These attractive fits, along with numerical tractability, account for the popularity of probit and logit in applied work. (The log-likelihood functions, though nonlinear in the parameters, are globally concave, so that numerical maximization is easy and reliable.)

In the dichotomous case ("success" or "failure"), both probit and logit generate the probability of a success as the value of a cumulative probability distribution function, that is, as a definite integral of a probability density function. To grasp the underlying intuition in a simple situation, suppose that there is just one independent variable and that it has a positive effect on success. Then the idea is that, if we plotted the probability of success against that variable, the shape of the graph would match some cumulative distribution function (cdf), perhaps a simple one with the stretched S-shape familiar from first courses in statistics. For this purpose, logit uses the standard logistic cdf, whereas probit uses the cdf of the standard normal. In both cases, the effects of the independent variables are nearly linear when probabilities of success are between 20% and 80%, but effects flatten at the extremes to keep probabilities bounded between zero and one.

Thus, to define the logit model, we first recall the density of the logistic distribution: $f_1(z) = e^{-z}/(1+e^{-z})^2$. Then if $P$ is the probability of success under the logit model, we set

$$P = \int_{-\infty}^{z} f_1(x)\,dx \qquad 1.$$

$$= F_1(z) = \frac{1}{1+e^{-z}}, \qquad 2.$$

where the second line is the cdf of the logistic distribution. If $Q$ is the probability of a failure, we also have

$$Q = 1 - P = \frac{1}{1 + e^z}, \tag{3.}$$

where the last equality follows from Equation 2.

In statistical applications, the setup is completed with a "link function": The argument $z$ of the cdf is expressed as a (typically linear) function of the explanatory variables. Subscripts are added to denote observation numbers. Thus, in most applications, $z_i = X_i\beta$, where $X_i$ is a (row) vector of explanatory variables for the $i$th observation, and $\beta$ is a fixed but unknown coefficient vector.

Under this specification, no matter what values on the real line $X_i\beta$ assumes, forecasts of $P_i = F_1(X_i\beta)$ always stay within the unit interval on which probabilities are defined, since the value of a cdf is itself a probability. This is the attraction of modeling a dichotomous dependent variable using a cdf.

In econometrics textbooks, logit and probit setups are usually generated from a random utility model. The quantity $z_i = X_i\beta + u_i$ is regarded as a utility for attaining or choosing success, where $u_i$ is a stochastic unobserved variable with a known distribution. Successes occur when utility falls above a threshold, conventionally set to zero:

$$p_i = Pr\,(X_i\beta + u_i > 0) \tag{4.}$$

$$= Pr\,(u_i > -X_i\beta), \tag{5.}$$

where $p_i$ denotes a probability derived from an arbitrary random utility model. Thus, when $u_i$ has some particular distribution with cdf $F_u$, successes have probability $p_i$ equal to the chance that a draw $u_i$ from its density $f_u$ falls to the right of $-X_i\beta$. This is simply the area under the density to the right of the point $-X_i\beta$, which is one minus the area to the left of the same point:

$$p_i = 1 - F_u(-X_i\beta). \tag{6.}$$

Now suppose that we model the left-hand-side probability $p_i$ in this equation as a cdf $F_p$ with density $f_p$, so that

$$F_p(X_i\beta) = 1 - F_u(-X_i\beta). \tag{7.}$$

Then the density $f_p$ must be the reflection (around zero) of the density of the disturbances $f_u$. To see this, observe that if $f_p$ and $f_u$ were reflections, then the area to the left of $X_i\beta$ under $F_p$ would equal the area to the right of $-X_i\beta$ on $F_u$. But this merely restates Equation 7.[2] Hence, in general, a random utility model

[2]Alternately, differentiating both sides of Equation 7 with respect to $X_i\beta$ gives $f_p\,(X_i\beta) = f_u\,(-X_i\beta)$, which restates the reflexivity in terms of heights of the densities rather than areas under them.

based on a particular density of the disturbances generates a functional form for the probability of success that is the cdf of another density, and the two densities are reflections of each other.

Now if the disturbance density is symmetric around zero, then the density is its own reflection, and therefore $F_u$ and $F_p$ in Equation 7 are the same. Replacing $F_p$ with $F_u$ in Equation 7 and substituting into Equation 6 gives

$$p_i = F_u(X_i\beta) \tag{8.}$$

for any arbitrary symmetric density such as the logistic or the normal. This is the familiar case seen in textbooks: The probability of observing a success has a cdf shape as a function of the explanatory variables, and that cdf is the same as the cdf of the disturbances in the underlying random utility model.

In particular, when the logistic cdf $F_1$ with logit success probability $P_i$ is used, then

$$P_i = F_1(X_i\beta) = \frac{1}{1+e^{-X_i\beta}} \tag{9.}$$

in parallel with Equation 2; here, again, $z_i = X_i\beta$. Similarly, for the probability of failure, we have

$$Q_i = 1 - F_1(z_i) = \frac{1}{1+e^{X_i\beta}} \tag{10.}$$

in parallel with Equation 3. Hence, the random utility approach to motivating the logit model is equivalent to the purely statistical specification in Equations 1 and 2. [The same is true for probit, the sole difference being that the normal (Gaussian) cdf replaces the logistic.] Note, however, the crucial importance of distributional symmetry of the disturbance in moving from Equation 6 to Equation 8, a point to which we shall return.

Taking derivatives in Equation 9 quickly establishes the familiar properties of the logit, for example that in a linear specification, explanatory variables have maximum marginal effect when $P_i = 0.5$, and that marginal effects diminish monotonically and symmetrically around that value, tending to zero as $P_i \to 0$ or $P_i \to 1$.

## SCOBIT

In a particular application, a researcher might prefer a cdf different from the logistic or the normal. Perhaps theory or experience indicates that large positive values of the disturbance term in the random utility model are more likely than large negative values, or that the maximum effects of the independent variables occur at a different probability value than 0.5.[3] One way to create such a specification is to note that any real number in the unit interval raised to a positive power remains in the unit interval. In particular, for the logit success and failure probabilities $P_i$ and $Q_i$, we have that $0 \leq P_i^\alpha, Q_i^\alpha \leq 1$ for any $\alpha > 0$.

After taking note of these considerations, Nagler (1994) uses them to define another estimator, called "scobit" (skewed logit).[4] The idea is to let the new probability of failure be the logit failure probability raised to the power $\alpha$. Thus, if $P_i^*$ and $Q_i^*$ are the scobit probabilities of success and failure, respectively, we set

$$Q_i^* = Q_i^\alpha = \frac{1}{(1 + e^{X_i\beta})^\alpha} \qquad 11.$$

using Equation 10, and then we adjust the success probability accordingly:

$$P_i^* = 1 - Q_i^\alpha = 1 - \frac{1}{(1 + e^{X_i\beta})^\alpha}, \qquad 12.$$

where we customarily require $\alpha > 0$. Obviously, when $\alpha = 1$, scobit reduces to logit. Thus, scobit is a legitimate generalization of logit; logit nests within it. A routine for estimating the model is now included in the statistical software package STATA.

A useful way to interpret scobit is to use Equation 12 to define a cdf:[5] $F^*(X_i\beta) = P_i^*$. There is no accepted name for this distribution defined by $F^*$, though it is closely related to the Burr distribution, whose cdf is

$$F_{\text{Burr}}(x) = 1 - (1 + x^c)^{-k} \quad (x \geq 0) \qquad 13.$$

[3]There is a technical point here: Because the underlying scale for these threshold models is arbitrary, one can always transform both sides of Equation 7 to get any cdf one likes for the functional form without affecting the fit at all. For example, if one wants the probit cdf $\Phi$ to replace some other distribution with cdf $F_p$ on the left-hand side of Equation 7, one would apply the compound function $\Phi\,[F_p^{-1}(.)]$ to both sides of Equation 7. Thus, in some sense, every threshold model for dichotomous dependent variables is equivalent to a probit setup. But the transforming function nearly always produces elaborately complicated functional forms for the explanatory variables on the right-hand side, with no clear substantive interpretation, and so the point is of no practical importance.

[4]His work is a rediscovery; the estimator was popularized in the statistical literature by Aranda-Ordaz (1981) and is often referred to by his name. The originator is Prentice (1976, p. 766). These earlier authors specify the same likelihood function slightly differently, which obscures their identity with scobit. Prentice, who derives Equation 11 from a more general estimator, multiplies both numerator and denominator on the right-hand side by $(e^{-X_i\beta})^\alpha$. In contrast, Aranda-Ordaz writes $Q_i = 1/(1+ \alpha^{-1} e^{X_i\beta})^\alpha$, which differs from Nagler's $Q_i^*$ by the addition of the constant $\alpha^{-1}$. But if we let $\alpha^* = \log \alpha^{-1}$, then we can replace $\alpha^{-1}$ with $e^{\alpha^*}$ and simply absorb the constant $\alpha^*$ into the intercept term in $X_i\beta$. This leaves us with the scobit likelihood.

[5]It is easily shown that $P_i^*$ meets the conditions to be a cdf. In particular, it is monotonic in $X_i\beta$.

and 0 otherwise (Burr 1942, p. 217, Equation 20). In fact, it may be shown that the $F^*$ distribution is exponential-Burr. That is, if $z$ has the $F^*$ distribution, then $x = e^z$ is distributed Burr.[6] Because the unlovely name "exponential-Burr" is not used in the literature, I will refer to the distribution most often as the "scobit distribution."[7]

As intuition suggests and Nagler demonstrates, the shape of the underlying density is in general no longer symmetric under scobit, and therefore marginal effects of independent variables in linear specifications are no longer symmetric around $P_i^* = 0.5$. Setting the second derivative of the cdf in Equation 12 to zero gives the (unique) maximum of the density and hence the point of greatest marginal impact:

$$\frac{\partial^2 P_i^*}{\partial z^2} = \frac{1}{(1+e^z)^{\alpha+1}} - \frac{(\alpha+1)e^z}{(1+e^z)^{\alpha+2}} = 0. \qquad 14.$$

Solving gives

$$z = -\log \alpha, \qquad 15.$$

and substitution into Equation 12 gives, for the point of largest marginal impact under scobit, $P^*$,

$$P^* = 1 - \left[\frac{\alpha}{\alpha+1}\right]^{\alpha}. \qquad 16.$$

Hence, for example, $P^* \to 0$ as $\alpha \to 0$, and $P^* > 0.5$ if $[\alpha/(\alpha+1)]^{\alpha} < 0.5$, which occurs when $\alpha > 1$.

Thus, maximal marginal effects under scobit need not occur where the probability of success is 50%, as in logit or probit. Under scobit, maximum impact may occur where the success probability exceeds 50% ($\alpha > 1$) or where it falls below 50% ($\alpha < 1$), an important and potentially useful generalization. As Nagler (1994, p. 253) notes, Equation 16 implies that the point of maximum impact is confined to the interval $(0, 1 - e^{-1})$, or approximately (0, 0.63). When larger points of maximum impact are needed, he essentially proposes switching to the power logit estimator, defined below.

Nagler (1994) applies probit, logit, and scobit to U.S. voter turnout data from the 1984 Current Population Survey from the Census Bureau, showing that scobit gives

[6]Morgan (1992, p. 186) sets this derivation of the exponential-Burr distribution as a problem for the student. If one takes derivatives of the $F^*$ and Burr cdfs to get the densities, then standard change-of-variable arithmetic suffices for the demonstration.

[7]Morgan (1992, p. 147) calls the scobit $F^*$ distribution "log-Burr," but this is a verbal slip. As his mathematics demonstrates, scobit is not log-Burr; rather, Burr is log-scobit. (Equivalently, scobit is exponential-Burr.) To see the plausibility of this claim, note that the Burr distribution is non-negative like the log-normal, whereas the scobit distribution, like the normal, covers the entire real line. Thus, the Burr relates to the scobit distribution in the same way that the log-normal relates to the normal, that is, Burr is log-scobit.

a slightly better statistical fit. He also finds that $\alpha \approx 0.4$, implying, from Equation 16, that voters with turnout probabilities of approximately 40% are most affected by the explanatory variables. Of course, probit and logit would have imposed a value of 50% as the point of maximum impact. Thus scobit yields a genuinely different substantive interpretation.

An alternate approach to scobit derives it from a random utility model. Smith (1989, p. 186) and Nagler (1994, pp. 253–54) take this approach, assuming the distribution of the disturbance term to be Burr II.[8] If the Burr II cdf is denoted $F_u^*(z)$, then by definition (Burr 1942, p. 217),

$$F_u^*(z) = \frac{1}{(1+e^{-z})^\alpha}. \qquad 17.$$

Substituting this $F_u^*$ for $F_u$ in Equation 6 and again using $z_i = X_i\beta$ produces

$$P_i^* = 1 - \frac{1}{(1+e^{X_i\beta})^\alpha}, \qquad 18.$$

and Equations 11 and 12 follow immediately, as desired.

Thus, as with logit, we may arrive at scobit via purely statistical considerations or by the econometric route of specifying a random utility model for individual choice. [In fact, since scobit may be derived in this way from the Burr II distribution, Smith (1989, p. 186) proposed calling the estimator "Burrit."] However, the scobit derivation differs in a crucial way from more familiar estimators. When logit is derived from a random utility model, the symmetric logistic disturbances lead to a logistic cdf functional form for the probabilities. Similarly, for probit, the symmetric normally distributed disturbances imply a normal cdf functional form for the probabilities. For scobit, however, the asymmetric density assumed for the disturbances does not lead to a cdf for the probability of success that has the same distribution. Instead, the assumption of Burr II disturbances leads to a scobit (exponential-Burr) cdf for the functional form.

The Burr II and exponential-Burr distributions are distinct, though closely related, as the disturbance cdf and the cdf for the probability of success must be in any random utility model. They have the relationship shown in Equation 7. As the discussion there implies, the Burr II and exponential-Burr densities must be reflections of each other. Informally speaking, any Burr II density may be converted to the corresponding exponential-Burr density by flipping it so that the left side becomes the right, as the discussion above at Equation 7 implies. In summary, then, under a random utility model, Burr II disturbances generate a cdf for the probability of success $P_i^*$ that corresponds to the scobit (exponential-Burr)

[8]Burr (1942) proposed a dozen (unnamed) distributions, of which this is the second. Subsequent authors have usually referred to them by Roman numeral (as in Johnson et al. 1994, pp. 53–54). The "Burr" distribution we have already encountered is Burr XII. Nagler (1994, p. 234, fn. 3) refers to Burr II as "Burr-10," since it appears in Burr's Equation 10.

distribution, and these two distributions have densities that are reflections of each other.

## AN ALTERNATE GENERALIZATION OF LOGIT: POWER LOGIT

In Nagler's scobit, it is the logit probability of *failure* that is subject to exponentiation. The probability of success is then chosen so that the two probabilities add to unity. Of course, one might have proceeded the other way around, raising the logit probability of *success* to the power $\alpha$ and forcing the probability of failure to adjust so that they sum to one. This is the "skewed logistic" of Robertson & Cryer (1974).[9] Because scobit and power logit are both "skewed logistics," however, and because "skewed logistic" is easily confused with "scobit," I have adopted Morgan's (1992, p. 186) alternate name for this estimator, "power logit."

Again using $P_i$ and $Q_i$ to represent the logit probabilities of success and failure, and defining $P_i^{**}$ and $Q_i^{**}$ to be the probabilities of success and failure under power logit, we set

$$P_i^{**} = P_i^{\alpha} = \frac{1}{(1 + e^{-X_i\beta})^{\alpha}} \qquad 19.$$

and

$$Q_i^{**} = 1 - P_i^{\alpha}, \qquad 20.$$

where the first line follows from Equation 9. We again require $\alpha > 0$. Of course, like scobit, this estimator reduces to logit when $\alpha = 1$.

If we interpret $P_i^{**}$ as a cdf, so that $P_i^{**} = F^{**}(X_i\beta)$, then the $F^{**}$ distribution is Burr II. (To see this, compare the definition of Burr II in Equation 17 to the definition of power logit in Equation 19.) That is, the cdf used in the functional form for power logit is the Burr II cdf. Like the scobit density, the Burr II density is asymmetric, so that again, this model allows the independent variables to have a point of maximum influence at probabilities different from 0.5. The largest marginal impact occurs at the point $P^{**}$, which is

$$P^{**} = \left[\frac{\alpha}{\alpha + 1}\right]^{\alpha}. \qquad 21.$$

Thus, $P^{*} \to 1$ as $\alpha \to 0$, and $P^{*} < 0.5$ if $[\alpha/(\alpha + 1)]^{\alpha} < 0.5.$, which occurs when $\alpha > 1$. In contrast to scobit, large values of $\alpha$ reduce the point of maximum impact, whereas small $\alpha$ values increase it. Power logit's point of maximum influence for the independent variables is confined to the interval $(e^{-1}, 1)$, approximately (0.37, 1). This is simply the reflection of the corresponding interval for scobit.

[9]Robertson & Cryer set out only the case $\alpha = 2$. Prentice (1976, p. 765) proposed the more general form shown here. It has been studied by Wu (1985) and McLeish & Tosh (1990).

Power logit seems never to have been derived from a random utility model, but it is easy to do so. To make the derivation successful, the density of the disturbances must be the reflection of the density of the power logit (Burr II) cdf $P_i^{**}$. However, we have already seen that the scobit (exponential-Burr) density is the reflection of the Burr II density. It follows immediately that we need to assume scobit-distributed disturbances here. That is, in a random utility framework, scobit disturbances generate the power logit (Burr II) functional form. A direct proof is straightforward.[10]

In summary, then, the random utility approach to generating scobit and power logit yields the following dual relationship, apparently not previously noticed:

Scobit: Burr II disturbances $\Rightarrow$ exponential-Burr cdf functional form

and

Power logit: exponential-Burr disturbances $\Rightarrow$ Burr II cdf functional form

Put more colloquially, in a random utility framework, scobit disturbances lead to the power logit model, and power logit disturbances imply the scobit model.

Perhaps the clearest way to see the duality relationship between these two estimators is to compare the scobit equation for failure (Equation 11), $Q_i^* = Q_i^\alpha$, and the power logit equation for success (Equation 19), $P_i^{**} = P_i^\alpha$, where again $P_i$ and $Q_i$ are the logit equations for success and failure, respectively.[11] Now from Equations 9 and 10, $P_i$ evaluated at $X_i\beta$ is identical to $Q_i$ evaluated at $-X_i\beta$. Hence, Equations 11 and 19 imply immediately that the probability of obtaining a "failure" under scobit with coefficient vector $\beta$ is the same as the probability of a "success" under power logit with coefficient vector $-\beta$. Thus, if we give one of these estimators a dataset in which successes and failures have been reversed, the maximum likelihood estimates will not remain the same except for the sign of the coefficients, as they would in logit or probit. Instead, the best fit will switch to a completely different model.

This seemingly minor point has a major consequence for empirical work. With logit and probit, researchers studying turnout, for example, are accustomed to ignoring whether voting should be coded as one and abstention as zero, or vice versa. Reversing the zeroes and ones on the dependent variable has no real statistical consequences. Scobit and power logit do not have that property, however. Reversing the zeroes and ones on the dependent variable for either one of them causes the estimator to switch to the other model. Thus, coding who is a zero and who is a one in a dataset is not a small step with these two estimators: Different choices

[10]Use Equation 12 as $F_u$ in Equation 7. This yields a cdf defining $F_p$ on the right-hand side. It has the same form as Equation 19, as desired.

[11]Incidentally, Equations 11 and 19 are not the usual notation for these estimators: I hope that writing them in this fashion makes the relationship and distinction between them clearer than it is in much of the literature.

produce genuinely different fits. In particular, the zero-one reversed fit for scobit yields the coefficients from power logit (with reversed sign), and vice versa.

The good aspect of this model-switching feature of scobit and power logit is that, although we may not have known it, we already have software for power logit. The scobit software in STATA can be used to estimate the power logit model—just reverse the zeroes and ones on the dependent variable, and then at the end, change back the sign of the resulting coefficients. The standard errors, log-likelihoods, and other features of the fit apart from the coefficients will be correct as printed out by STATA.

In summary, both scobit and power logit generalize the logit model. Each offers potential for fitting datasets not well modeled by the symmetric logit and probit estimators. Moreover, for each of them, at least a partial rational choice microfoundation has been successfully laid, since each has been derived rigorously from a particular distribution of the disturbances in a random utility model. Quantitatively skilled graduate students will want both estimators in their toolkits, particularly now that STATA makes appropriate software available.

Political methodologists have long suspected that our familiar estimators were often too restrictive. Dichotomous-dependent-variable models were thought to be a good example. Now we have generated freer models with more parameters and fewer limitations. And we have believed that more generality is always good.

## THE PERILS OF GENERALIZING FAMILIAR ESTIMATORS

Social scientists currently have a wealth of dichotomous-dependent-variable models from which to choose, including many not mentioned here (e.g., Prentice 1976, Stukel 1988, Morgan 1992). Moreover, now that Nagler has shown political scientists the way, other dichotomous-dependent-variable estimators can be generated for our purposes freely and pleasantly.

For example, all the estimators discussed above might be nested inside a single estimator. One way to do this would be to add one new parameter $\gamma$ and then write the probability of success as a mixture of the scobit and power logit probabilities ("mixit"):

$$P_i^{\text{mix}} = \gamma P_i^* + (1-\gamma) P_i^{**}, \qquad 22.$$

where $0 \leq \gamma \leq 1$. Obviously, scobit and power logit are the special cases in which $\gamma = 1$ and $\gamma = 0$, respectively. This new estimator also allows for functional relationships in the data that logit, scobit, and power logit cannot include; it has considerable flexibility. In the contemporary style, this estimator might be proclaimed Generalized Scobit and Power Logit, and preached as GSPL.

Alternatively, rather than constructing a weighted sum of the probabilities of success from the scobit and power logit, we might multiply them instead ("clumpit"):

$$P_i^{\text{clump}} = \frac{(P_i^*)^\gamma (P_i^{**})^{1-\gamma}}{(P_i^*)^\gamma (P_i^{**})^{1-\gamma} + (Q_i^*)^\gamma (Q_i^{**})^{1-\gamma}}, \qquad 23.$$

where again $0 \leq \gamma \leq 1$, and scobit and power logit are the special cases in which $\gamma = 1$ and $\gamma = 0$, respectively. Here, as for all the previous estimators, it is not hard to demonstrate that the standard features of a cumulative distribution function hold for the function defining $P_i^{\text{clump}}$. (In particular, $P_i^{\text{clump}}$ is monotonic in its argument.) Like mixit, clumpit has substantial flexibility of fit, and values of all its parameters can be computed by maximum-likelihood estimation, or, if priors are imposed on the parameters, by Bayesian computations.

Still more statistical models for dichotomous dependent variables might be created. All the estimators discussed above start from the logit cdf $F_1$. They use that cdf to define probabilities of success and failure, and then transform the probabilities in some fashion. Instead, one might start from the normal cdf, define the corresponding probit probabilities, and then transform the probit probabilities in the same ways. Or one might start with the cdf from t-distributions, or the double exponential, or Cauchy, or many others.[12] Combining these possibilities with scobit and power logit, plus the new mixit and clumpit, we have painlessly created in one paragraph more than a dozen brand-new dichotomous-dependent-variable estimators. Extending each of them to polychotomous responses is straightforward, too: One proceeds just as with polychotomous probit. There is no end of opportunities.

By now, though, a concern should have arisen in the reader's mind. For this generality is all too quick. Yes, dozens of estimators are easily created for any situation. Unfortunately, they often fit approximately equally well but give quite different answers. If any of them might plausibly be used on statistical grounds, which one is best for a given problem? Trying them all, besides being unreasonably burdensome, is not even possible; there will always be another ten untried. Purely statistical considerations cannot tell us what to do.

Worse yet, generality is not free. These setups with additional parameters often require surprisingly large datasets to be successful. Consider the best-known generalization of logit, namely scobit. Scobit adds only a single parameter to logit. Yet computational experience with it indicates that samples of 500 are often too small for reliable results when that parameter is added. In Nagler's (1994) own simulations with samples of 500, scobit sampling variances for coefficients were routinely five to ten times larger than those of the corresponding logit, and sometimes 100 or even 1000 times larger. Even in samples of 2000, some coefficients had sampling variances 25 to 100 times larger than logit's. Only in Nagler's study of eligible voters, with nearly 100,000 observations, did the scobit sampling variances

[12]It is convenient to use distributions whose support is the entire real line so that out-of-bounds forecasts do not occur, but this allows for log chi-square, log exponential, and many others, as well as those listed above.

settle down to averaging only about twice the size of logit's, a reasonable statistical price to pay for the increased flexibility of fit.

These features of scobit have been investigated by Hanmer, who replicated Nagler's simulations.[13] He finds the source of the problem in occasional wild misestimates of $\alpha$, the additional scobit parameter, which then cause serious errors in the other coefficients. He also finds that estimates of $\alpha$ are very sensitive to functional form, so that including squared terms in a specification (whether they belong or not) can cause dramatic changes in the estimate of $\alpha$. Often, the $\alpha$ term seems to capitalize on chance, changing dramatically to try to accommodate one or two data points. In one run, Hanmer found that dropping one observation out of 500 changed the estimated $\alpha$ from 680,000 to 38. Removing one more observation reduced $\alpha$ to 5. The other coefficients sometimes doubled or were cut in half as $\alpha$ changed.

These upheavals took place in data simulated with the same distributions and parameters Nagler used in his own simulations, guaranteed to meet scobit's assumptions, and estimated using the model known to be correct. (The real world would no doubt have been more devious.) Even so, a sample with a truly dramatic error in the estimated $\alpha$ turned up in the first 100 simulated samples Hanmer tried. Serious errors of estimation occurred in about 5% of all 500-observation datasets. Moreover, none of this trouble is unique to scobit. All these findings apply to power logit as well, by the usual trick of reversing the zeroes and ones. And one shudders to imagine empirical and computational experience with mixit and clumpit, which add *two* parameters to logit. In short, if the reader has not already guessed, mixit and clumpit are fakes—mathematically correct but not to be taken seriously. Many a "generalized" estimator glitters emptily.

It is important to understand that nothing in the previous paragraphs indicates that scobit and power logit have no uses, or that the software used to generate their estimates is misleading. To the contrary, the estimators are genuine advances and the software generally works well on what is a difficult numerical estimation.[14] The point is rather that generalizing logit can be very expensive in statistical precision, a point confirmed by theoretical work on scobit (Taylor 1988). Precision is much less an issue when samples have 100,000 cases, as in Nagler's substantive study with Census Bureau data. Then one can let the data speak relatively unaided. But in survey samples of 1000 to 2000, typical of political science work with dichotomous

[13]See MJ Hanmer, "An Investigation of Scobit," unpublished manuscript, Department of Political Science, University of Michigan.

[14]Altman & McDonald (2002) find that the scobit maximum-likelihood estimates (MLEs) are numerically hard to compute even in routine cases and that some standard packages, such as GAUSS, occasionally fail to find the true MLEs, even getting the sign wrong on some estimated coefficients. It is possible that the Aranda-Ordaz version of this estimator, which reparameterizes the distribution to lessen the correlation between $\alpha$ and the other coefficients, might help. In any case, this issue (whether the answer printed by the computer program is the correct estimate) is distinct from that discussed by Hanmer (whether the correct estimate is near the truth).

variables, one needs a strong formal-theoretic or detailed data-analytic reason to be using scobit or power logit.

Some readers of this argument have imagined that it applied only to scobit, an estimator not much used in practice. Certainly, they have said, scobit has problems with its standard errors. But that need not stop us from happily creating and using our other substantively atheoretical generalized estimators and MLEs. Hence, the concerns of this paper are easily dismissed.

In fact, however, this defense of conventional wisdom resembles that of the Hapsburgs, who were secure in their belief that the Empire's weaknesses were confined to Serbia. Like Serbia, scobit may expose the issues a little more clearly, but nearly all the new estimators proposed in political methodology in recent years raise the same concerns as does each application of scobit. Since each new estimator imposes a certain structure on the data and often uses up additional degrees of freedom to create statistical generality, why should we believe these assumptions in this problem? Typically, no formal model supports the assumptions, and no close data analysis is presented in their favor. In fact, no matter how devastating those absences, we often write as if we didn't care. For both the creators and the users of our new estimators, simply listing the assumptions seems satisfactory, and we treat the ensuing estimates as findings. Statistical estimators have the logical form *If A, then B.* "Therefore B," we cry.

We have now come to the central issue facing contemporary political methodology. Dozens of estimators might be used in any of our empirical applications. Too often, applied researchers choose the standard ones because they believe methodologists approve of them, whereas methodologists prefer some new, complicated, untested alternative because they know that the standard estimators are often ungrounded in substantive theory, and they hope that the new one might stumble onto something better. Few researchers in either group make a convincing case that their estimator is humming rather than clanking on their dataset. Even the creators of estimators usually do not prove that the supporting assumptions would make rational sense or common sense for the political actors being studied. Nor do they carry out the patient data analysis required to show that their estimator, an arbitrary selection from among dozens that might have been proposed, is more than just computable and plausible, but that its assumptions really match up in detail to the data for which it is intended. If the thing might work on some planet, we think our job is done.

Too many of the new estimators in political methodology are justified solely because they are one conceivable way to take account of some special feature of the data. Perhaps the dependent variable is discrete, or a duration, or a count, or an ecological average, or perhaps partially missing data. Then under some all-too-convenient assumptions, we show that the implied estimates are MLE or Bayes, and we demonstrate that our computers can solve for the parameters. Applied researchers are grateful: "An estimator that takes account of the special features of my data in a way that ordinary regression never did—hooray!" Too often, they rush out to adopt it, not noticing that it may give bizarre answers that standard, simpler, better-tested estimators, perhaps unfamiliar to them, would have avoided.

Once upon a time, our tools were very limited, and econometrics texts taught us, "Decide what sort of data you have, and look up the corresponding estimator." Few questioned the assumptions closely; there were no real alternatives. But those days are long gone. No researcher should suppose now that there is just one statistically reliable technique for a given class of data. There are many, and dozens more are easily created. No one should imagine that some particular newly invented estimator emerging in a prominent political science journal is the only or best way to analyze a dataset. Applied political researchers need to wise up, and political methodologists need to stop ill-using them by promoting particular estimators on abstract grounds of greater generality. The truth is that, for virtually any political dataset in common use, dozens of statistical estimators might be tried, and we simply have not done the work needed to recommend any one of them with scientific honesty.

In short, creating more and more abstract estimators, unrelated to well-grounded empirical generalizations, cannot be the right way to define our job as political methodologists. Statisticians do that for a living, and we will never be as good at their job as they are. Trying to keep up will leave us forever second-rate—at best—and, more importantly, irrelevant to genuine empirical advance in the discipline.

We have a different agenda. One can see it in good statistics texts, wherein the statistician is constantly advised that many techniques are available, and that choosing the right one requires consulting the quantitatively sophisticated researchers in a given field. Inventing new applied estimators is relatively easy, statisticians are told; the trick is to find those that truly fit the data on a particular subject. Ask the specialists, who know the statistical characteristics of the data in detail; then, the texts say, select an estimator on that basis. Right now, though, if statisticians consulted political methodologists concerning the statistical character of our observations, we would have too many second-rate estimators and not enough first-rate answers. What can be done?

## MICROFOUNDATIONS

A "microfoundation" for a statistical specification is a formal model of the behavior of the political actors under study. The model might emerge from decision theory, game theory, or some other formalism. Then the statistical setup is derived mathematically from the model, with no further ad hoc adjustments. An independent, normally distributed error term ("white noise") may be added for the inevitable random, nonsystematic deviations from the model.

The simplest example of a dichotomous-dependent-variable estimator that is microfoundation-ready is the probit model. Suppose that some formal model explains the probability of success (say, a country signing a particular treaty) as a function $g(.)$ of certain exogenous variables $X_i$. Success occurs (the country signs the treaty) according to the threshold model in Equation 4, with many small, random, nonsystematic factors incorporated into an additive disturbance term $u_i$. (Deriving the existence of additive normal disturbances from the formal model is even better, of course, but not always possible or sensible.) It follows from the formal model, let us suppose, that success occurs when $g(X_i) + u_i > 0$. We may then derive rigorously, as we did in going from Equation 4 to Equation 8,

$$p_i = \Phi[g(X_i)], \qquad 24.$$

where $\Phi$ is the cdf of the standard normal distribution.

Equation 24 is the probit model, with some (perhaps complicated) function of the exogenous variables as its explanatory foundation. To use this model is to assert that the model $g$ is the correct representation for the systematic part of the behavior. The Central Limit Theorem justifies the claim that if each of the other, residual factors is nonsystematic and not too intercorrelated, with no subset of those factors dominant over the others, then the disturbance term should be independent and normally distributed. The idea is that a good model fitted to data results in white noise disturbances. That assumption may be disputed, of course, and tested against the data, but it is intellectually coherent and has standard justifications. Thus, probit is not itself a formal model, but it combines easily with one. Much the same can be said for logit, which is essentially indistinguishable from probit in applications. In either case, a formal model plus a justification for white noise disturbances yields a probit or logit setup as a statistical model with microfoundations.

Now let us take up the case of scobit. We have seen that scobit requires disturbances distributed as Burr II. Because there is no reason to believe that the Burr II distribution would occur by chance, it cannot be assumed merely for convenience in a particular application. The same is true for other distributional assumptions used in our estimators, whether Cauchy, t-distributions, truncated normals, or beta distributions. If knowledgeable people are to take the resulting estimators seriously, the distributional assumptions must be defended theoretically and justified with a formal model of the behavior of the political actors. Atheoretical assertions that an estimator follows from *some* arbitrary assumptions, no matter how rigorously, will not persuade. As we have already seen, there are too many possible estimators for any problem. For an estimator to be believed, it requires microfoundations. But when an estimator such as scobit does not use white noise disturbances, how might microfoundations be supplied?

A somewhat artificial case is apparent in the treaty-signing example. Suppose that some formal model implies that a rational governmental decision maker will sign a treaty if any one of the major interest groups in the country supports the treaty. Suppose further that there are $\alpha$ such groups in country $i$, and that each of them will support the treaty with probability $P_i = F_1(X_i\beta)$, where $P_i$ is the logit probability of support as a function of exogenous factors $X_i$ related to characteristics of the country and the treaty. These probabilities must be identical for each group.

Then the probability $P_i^*$ that the leader will sign the treaty is the probability that at least one of the groups will support it, which is one minus the probability that no group will support it. If $Q_i = 1 - P_i$ is the logit probability of opposition for each group, then the probability of the treating being signed by country $i$ is

$$P_i^* = 1 - Q_i^\alpha, \tag{25.}$$

which is the scobit model of Equation 12.

Other, similar situations might also generate a scobit specification: "If anyone in the family wants to drive down to vote, we will all go," or "If you can give me one good reason to send a check to Snooky for Senate, I'll send one." When $\alpha$ different conditions are each *sufficient* for success and all have the same logit-based probability of occurring, then the scobit model is mathematically implied. Political actors will behave as if they obeyed a threshold model of choice with Burr II disturbances, but the Burr II assumption will not be arbitrary. Instead, it will be a logical consequence of an underlying formal model with white noise disturbances.[15]

Readers may wish to verify that power logit has much the same potential justification. When $\alpha$ different conditions are each *necessary* for success and all have the same logit-based probability of occurring, then the power logit model is implied. In that case, actors will behave as if they followed a threshold model of choice with exponential-Burr disturbances, but again, the claim that they do so is not arbitrary.[16]

Thus, substantive formal models of a certain kind would give microfoundations to scobit and power logit. They would tell researchers that employing these special estimators is indicated, or even required. As with other estimators, arguing for the use of obscure distributions purely on grounds of computational convenience or aesthetic attractiveness should be avoided. The Burr II and exponential-Burr distributions would be derived from a clean foundational model with routine, conventional logistic errors that required no special pleading for credibility.

The formal model justifying a particular application of scobit or power logit has to be plausible, of course, if the microfoundation strategy is to be successful. The examples of informal models just discussed are all questionable, and they seem to show that model-based justifications for scobit and power logit occur only occasionally. When the posited model justifying an estimator is not persuasive, then a debate will break out. But at least the debate can focus on the signing of treaties, about which political scientists are likely to be knowledgeable, rather than on the occurrence of Burr II disturbances, about which our expertise is negligible. In fact, the latter topic can be ignored. The outcome of the debate on treaty accession will logically determine the choice of estimator. That is what microfoundations are for.

[15]For careful thinking about the statistical implications of models with necessary and sufficient conditions, see BF Braumoeller, "Causal Complexity and the Study of Politics," unpublished manuscript, Harvard University.

[16]Microfoundations can be constructed for mixit and clumpit as well, but they are even more specialized than those for scobit and power logit and thus are not to be taken seriously.

Thus, occasionally, models such as scobit, power logit, and other MLEs will be implied by a theoretical model. When they are, they have microfoundations and should be the estimator of choice. More often, though, their usefulness will be found in checking for specification errors. Like many other specification checks and tests, they can help us find model errors. When logit follows from a formal model and power logit does not, but power logit fits better, then we know something is wrong in the formal theory supporting logit or in the implementation of the logit specification. For finding our mistakes, scobit, power logit, and their estimator cousins in other applications are most helpful.

Nagler's (1994) study, for example, shows that our standard specifications for voter turnout are not working in logit and probit. That is an enormously valuable contribution. But in the modern view, the implication is not necessarily that we should abandon logit and switch to one of its generalizations. It is rather that we need to think hard both about the formal theory of turnout and about the specifications we use in logit and probit to study it. (An important step toward a theoretically grounded empirical study of turnout is Sanders 2001.) If we cannot think of any reason why scobit has formal-theory support, however, then jumping to it bears a heavy burden of proof and should be considered with skepticism. Instead, the theoretically defensible goal is either to re-do the theory or, perhaps more commonly, to find and fix the specification errors in the link function. When that has been done in a context where logit has a strong, persuasive formal-theoretic justification, we expect that in the end, logit will usually turn out to have the best fit. Good theory will then be vindicated, and scobit will have played a key auxiliary role in that outcome.

At this point, no doubt, empirical investigators and methodologists accustomed to contemporary political science norms will object. "Look," they will say, "this new Glockenspiel estimator may not have those frou-frou microfoundations you insist on, but it makes theoretical sense by my lights: It takes account of the yodeled nature of my dependent variable, which ordinary regression ignores. Plus it can be derived rigorously from the Cuckoo distribution. Besides, it fits better. The graphs are pretty, at least if not looked at too closely, and the likelihood ratio test rejects the ordinary regression fit at the 0.05 level. Theory-schmeary. Our job is to let the data decide. I'm going to use Glockenspiel. Anything else is choosing a poorer fit." Nearly all of us methodologists have shared these views at some stage of our professional lives.

Nowadays, this is the battle line where the old political methodology and the old political science confront the new. Devotees of the old computing-power-plus-MLE viewpoint are "fitness buffs." If Glockenspiel fits a little better than regression, we have traditionally told ourselves, then it is a better answer than regression or probit. But as we have all learned by our own painful experience, good statistical fitness is not enough. That training regimen too often drives out thinking.

The old style, in which so many of us were trained and which increasing computing power makes even more seductive, is content with purely statistical derivations from substantively unjustified assumptions. The modern style insists on formal theory. The old style dumps its specification problems into a strangely distributed disturbance term and tries to model or correct the resulting mess; the new style insists on starting from a formal model plus white noise errors. The old style thinks that if we try two or three familiar estimators out of 50 possible ones, each with some arbitrary list of linear explanatory variables and fabricated distributional assumptions, and one of them fits better, then it is the right answer. The modern style insists that, just because one atheoretical fit is better than another, that does not make any of them intellectually coherent or satisfying. Instead, a new estimator should be adopted only when formal theory supports it, and not otherwise.

Empirical research closely informed by formal theory has made significant headway in certain fields of political science, notably in studies of U.S. legislators, bureaucrats, interest groups, and the relationships among them—a literature which would require a review article of its own (but see Morton 1999, especially ch. 8). Other examples would include Bartels (1998), who estimates the changing voting power of various American subgroups by exploiting both formal theory about the Electoral College and the extensive datasets provided by the National Election Studies. Bartels has remarked to me that his calculations, like many other formal-theoretic studies of voter turnout, implicitly rely on the scobit $\alpha$ equaling unity, which seems untrue—an example of methodological development influencing theoretical conclusions.

In international relations, Schultz (2001) constructs a model of international bargaining with a free domestic opposition. His explanation for the "democratic peace" (the observation that democracies do not fight each other) implies fresh ways to test for it. Building on related theoretical work by McKelvey & Palfrey (1995), Signorino (1999) shows the striking difference that a formal model makes in the statistical study of international crisis behavior. He pioneers the stage-by-stage statistical modeling of real-world political games. Sartori (2002) exploits a game-theoretic argument to impose a new identifying condition on a selection-bias model of crisis bargaining, and she goes on to provide the most statistically sophisticated analysis of an MLE model yet done by a political scientist. Both the Signorino and the Sartori papers show the power of contemporary formal theory: No methodologist in the old tradition would have thought to propose either of these unfamiliar statistical setups. Both emerge directly from theory, not from econometrics books with their convenient, familiar, substantively unjustified distributional assumptions and functional forms.

## TOWARD RELIABLE EMPIRICAL GENERALIZATIONS

Thus far, the discussion has emphasized methodology as the testing of theory. Morton (1999) has admirably reviewed the literature from this perspective. Certainly theory testing is a central task for methodologists. However, methodologists have another role as well, at least equally important. A theory needs things to explain, and finding them is part of our job, too. Much useful theory has emerged bottom-up rather than top-down. One example is the growing literature on Bayes models of public opinion (Zechman 1979, Achen 1992, Bartels 1993, Gerber & Green 1998).

The discovery of thoroughly reliable quantitative generalizations with theoretical bite is often more crucial to the discipline than theory testing. Fecund empirical generalizations certainly exist in political science. The democratic peace may be one such generalization; "party identification predicts the vote very well" seems to be another. Both these propositions have engendered substantial decision- and game-theoretic literatures. Admittedly, both would be more helpful if we knew precisely what "democracy" meant in the first instance and "party identification" in the second, but progress is occurring on both definitions. (On democracy, see Munck & Verkuilen 2002; the revisionist theory of party identification begins with Jackson 1975.)

Neither of these two generalizations about political life came from prior theory. (Yes, Kant had proposed the democratic peace, but almost nobody believed him, and his arguments had been forgotten until empirical researchers surprised everyone with strong evidence.) Both generalizations are important discoveries, and both demonstrate that empirical work often comes before smart theorizing rather than following it, a phenomenon familiar from the natural sciences. Kepler's laws preceded Newton and structured his theorizing; the surprising discovery that black box radiation arrived in discrete units led to quantum mechanics. In short, empirical research has an essential role that involves its own kind of imagination and creativity apart from theory. Empiricists are not simply slack-jawed, dwarfish varlets following the theorist around and washing up the glassware.

We methodologists often find ourselves in Hempel's "context of discovery," with no theories, formal or otherwise, to guide us—a little social psychology, perhaps, but nothing up to the task of making our inferences reliable. Microfoundations remain the gold standard, but often we have to begin with less and search for the empirical regularities that might lead to theory. In that context of high-dimensional problems with too little theoretical structure, how can careless curve-fitting and unreliable findings be avoided?

The usual answer is that, in research problems without microfoundations, we need hard work, insight, and art to see patterns and establish credible empirical regularities. We think of ourselves as following that advice. But our conventional procedures have let us down, and we have had little success. None of the important empirical generalizations in the discipline has emerged from high-powered methodological research. Instead, almost without exception, they were found with graphs and cross-tabulations. Methodological advances, from multiple regression onward, have largely been irrelevant.

To enjoy better times, quantitatively sophisticated empiricists will have to change their way of thinking. Kramer (1986) once wrote that creating a theory is relatively easy; it is learning whether the theory is true that is hard. And he added that political scientists tend to believe the reverse.

Empirical work, the way too many political scientists do it, is indeed relatively easy. Gather the data, run the regression/MLE with the usual linear list of control variables, report the significance tests, and announce that one's pet variable "passed." This dreary hypothesis-testing framework is sometimes seized upon by beginners. Being purely mechanical, it saves a great deal of thinking and anxiety, and cannot help being popular. But obviously, it has to go. Our best empirical generalizations do not derive from that kind of work.

How to stop it? The key point is that no one can know whether regressions and MLEs actually fit the data when there are more than two or three independent variables. These high-dimensional explanatory spaces will wrap themselves around any dataset, typically by distorting what is going on. They find the crudest of correlations, of course: Education increases support for liberal abortion laws, for example. In the behavioral tradition, that counts as a reliable finding. But no one knows why education is associated with that moral position (higher intellect discovering the truth? Mindless adoption of elite tribal norms? Coincidence due to correlation with something else entirely?), and that leaves open the possibility that abortion attitudes do not work the way our simple linear statistical models assume that they do.

Are educated Protestant evangelicals more enthusiastic about relaxed abortion laws than less-educated members of their denominations, for example? In the political science literature, at least, almost no one knows; we have not published the relevant cross-tabulations, and so we know very little about interactions of that kind. Instead, we proceed as we have been trained, looking at the coefficients in large statistical models. Hence, we know only that when linear probit models have mushed their way helplessly through national samples with jumbles of Baptists, Quakers, agnostics, Mormons, Christian Scientists, Jews, Catholics, and Presbyterians—some black, some white, some Asian, and some Hispanic—then education acquires a positive coefficient in predicting liberalism concerning abortion. Whether these different groups of people have unique histories, respond to their own special circumstances, and obey distinctive causal patterns, we do not know because we do not check. In consequence, no real knowledge about the influence of education on abortion attitudes follows from the positive coefficient. Getting rid of this cheap sense of "empirical findings" is probably the central task that quantitative political science faces.

Consider, for example, Nagler's (1994) statistical finding from scobit, discussed above, that the maximum impact of the variables explaining voter turnout occurs when the probability of turnout is approximately 40%. If true, this would be a highly consequential finding, both for political scientists trying to understand why people vote and for political practitioners seeking to target their mobilization efforts. How might it be verified credibly and in detail? Begin with education: Those who have attended college vote more frequently than those who finished only high school. From the cross-tabulations, is the turnout gap between these two groups really largest when the model predicts 40% turnout? Or consider age: Forty-year-olds vote more than thirty-year-olds. Is the gap largest at 40% turnout? How about election-day registration? Is turnout in states with this provision larger than in those without it, and is the gap largest for citizens with a 40% chance of voting?

This is the sort of detailed investigation that truly convinces an alert reader and builds reliable empirical generalizations.

## TOWARD RELIABLE ESTIMATORS

Each estimator requires the investigator to be sensitive to its own special features. Consider, for example, the attractive application of heteroskedastic probit to opinion data (Alvarez & Brehm 1995, 1997, 2002). Heteroskedastic probit has the same structure as Equation 8: The probability of a success is given by the cdf of a particular distribution, in this case the normal distribution. Customarily, the normal distribution is derived from a random utility model with normally distributed disturbances, as we have seen.

Unlike ordinary probit, however, in which the disturbances are assumed to be distributed standard normal with fixed variance, heteroskedastic probit allows their variance to depend on exogenous variables. Thus, in the notation of Equation 8,

$$p_i = \Phi_{\sigma_i}(X_i\beta), \tag{26.}$$

where $\Phi_{\sigma_i}$ is the cdf of a normal distribution with mean zero and variance $\sigma_i$.

The specification is then completed by setting the standard deviation of the disturbances, $\sigma_i$, equal to a function of exogenous variables $Z_i$, for example,

$$\log \sigma_i = Z_i\gamma. \tag{27.}$$

Thus, heteroskedastic probit generalizes ordinary probit in the same way that heteroskedastic regression generalizes ordinary regression, and it is an important model for the same familiar reasons.

The interpretation of the variance part of the model is tricky, however. Anything that generates higher variance will improve the fit. With opinion data, ambivalence is one possibility, if it causes large error variance in responses. Alvarez & Brehm stress this source of higher variance. However, extreme but opposed opinions among respondents, with no ambivalence at all, are another possible source. Careful investigation will be needed to distinguish between these alternatives.

The sorting is made particularly difficult because probit has no natural scale. A heteroskedastic probit model with explanatory variables $X_i$ is completely equivalent to a homoskedastic probit with explanatory variables $X_i/exp(Z_i\gamma)$. (This is the usual "correction for heteroskedasticity" transformation familiar from regression analysis.) In other words, the $Z_i$ variables might enter the equation either because they affect the disturbance variance or because they affect the responses directly.

Nor will it be easy to use the functional form of Equation 27 to separate the two possibilities. The standard deviation $\sigma_i$ varies only modestly around unity in most applications. Hence, by the usual Taylor series expansion, to a very good approximation,

$$exp(Z_i\gamma) \approx 1 + Z_i\gamma \tag{28.}$$

where $Z_i\gamma$ is small.

It follows that the multiplicative inverse of $exp(Z_i\gamma)$ is approximately $1 - Z_i\gamma$ plus very small higher-order terms. Assuming that $\beta$ contains an intercept term $\beta_0$, and writing $X_i\beta$ as $\beta_0 + X_{1i}\beta_1$, we find

$$X_i\beta/exp(Z_i\gamma) \approx X_i\beta - \beta_0 Z_i\gamma + \text{small interaction terms in } X_{1i} \text{ and } Z_i. \quad 29.$$

The left-hand side of this equation was constructed because it was the link function in an ordinary probit equation statistically indistinguishable from the heteroskedastic case. But it has turned out to be very nearly, apart from the difficult-to-detect interaction terms, a simple linear specification in $X_i$ and $Z_i$, the collection of variables that influence the dependent variable directly and those that influence the disturbance variance. (The latter have their sign reversed.)

In short, it will be challenging to distinguish a variable's positive effects on the disturbance variance from its negative effects on the dependent variable (and vice versa). Does education reduce ambivalence, or does it just move opinions in a positive direction? We will be hard-pressed to tell the difference. Trying to estimate both at the same time will make the estimator nearly collinear. Small specifications, carefully formulated with formal theory in mind and relentless data analysis, will be needed to make heteroskedastic probit models yield findings we can rely on with confidence.

Similar remarks might be made about applications of multivariate probit models to vote choice among multiple candidates, with which Alvarez & Nagler (1995, 1998) have done important pioneering work. Such models require careful specification of covariances among error terms if the models are to be identified, and careful testing of the resulting forecasts to check whether the strong assumptions of multivariate normality truly describe the nature of voters' decision making. Much data-analytic experience will be needed before multivariate probit is ready for routine production work.

Making a serious case that an estimator is working well is like validating an empirical generalization—very hard work. Traditionally, we have tried to do both with informal assumptions about the right list of control variables, linearity assumptions, distributional assumptions, and a host of other assumptions, followed by a significance test on a coefficient. But since all the assumptions are somewhat doubtful and largely untested, so are the estimators and the conclusions. The depressing consequence is that at present we have very little useful empirical work with which to guide formal theory. Behavioral work too often ignores formal theory. That might not be so bad if it did its own job well. But it produces few reliable empirical generalizations because its tests are rarely sharp or persuasive. Thus, empirical findings accumulate but do not cumulate.

## A RULE OF THREE

Only a more modern approach can halt the proliferation of noncumulative studies. As an instance of the altered perspective I have in mind, I propose the following simple rule, to be applied when no formal theory structures the investigation and we must rely on the art of data analysis:

**A Rule of Three (ART):**
**A statistical specification with more than three explanatory variables is meaningless.**

ART may sound draconian, but in fact, it is no more than sound science. With more than three independent variables, no one can do the careful data analysis to ensure that the model specification is accurate and that the assumptions fit as well as the researcher claims.

Why a rule of three, and not four or two? Rigidity is inappropriate, of course, but the number three is not wholly arbitrary. The guideline is derived from many researchers' experience. Close study of two explanatory factors is usually easy. However, the curse of dimensionality sets in quickly. Collinearity among explanatory factors plagues social science and multiplies the pains of data analysis rapidly as the number of factors rises. Serious data analysis with three explanatory factors is not much like using two, and using four is so hard and so time-intensive that it is almost never done astutely and thoroughly. Sorting out the effects of three variables is a daunting but not impossible task. Hence the rule of thumb: Truly justifying, with careful data analysis, a specification with three explanatory variables is usually appropriately demanding—neither too easy nor too hard—for any single paper.

If one needs several more controls, then there is too much going on in the sample for reliable inference. No one statistical specification can cope with the religious diversity of the American people with respect to abortion attitudes, for example. We have all done estimations like these, underestimating American differences and damaging our inferences by throwing everyone into one specification and using dummy variables for race and denomination. It's easy, but it's useless, and we need to stop.

In any study of political thinking or action, whether abortion attitudes, voter turnout, or international crisis behavior, the various subgroups of actors must be taken seriously and looked at separately and in detail. Cross-tabulation and plotting enforce this mental discipline, and they are the way to start any analysis. But the same logic also implies that when we use our more powerful contemporary statistical tools, we need to subset the sample. Some religious and philosophical communities, for example, have to be set aside in the study of abortion attitudes because we lack adequate data about them. Put bluntly, in most of our empirical analyses, some groups of observations should typically be discarded to create a meaningful sample with a unified causal structure.

Data collection is expensive, and discarding observations will initially seem wasteful. Why confine a probit analysis to African-American abortion attitudes, for instance? The subsample will be much smaller than the full dataset, and it will be harder to speak with confidence about the findings. Instead, why not just throw half a dozen dummy variables and another several linear control variables into the probit analysis to mop up diversity? That would save all the observations. After all, these control variables "matter." Let's put them all in and use all the observations. So goes the conventional wisdom.

Unfortunately, the conventional approach creates devastating inferential consequences. As a brief look at relevant data quickly shows, no one should be studying black Americans' abortion attitudes with a dummy variable for race. A study that gets the unique causal patterns of black Protestants approximately right and throws everyone else out of the sample is better than an analysis that tosses every group into the statistical soup and gets them all wrong. A phony big-sample certitude is no help to anyone.

Similar remarks apply to virtually everything we study. Sometimes patient investigation will show that coefficients vary only a little from one observation to the next, and then our customary procedures will work adequately when applied to the full dataset. But often the causal patterns are dramatically different across the cases. In those instances, subsetting the sample and doing the statistical analysis separately for each distinct causal pattern is critical. Happily, these causally homogeneous samples will need far fewer control variables and make the application of ART easier, because irrelevant subgroups will have been set aside for separate analysis and the corresponding control variables will be unnecessary. Attractive examples of this style of empirical work include Gowa (1999) and Miller (1999).

To do contemporary data analysis, then, we need to consider carefully what explanatory situation we are in. Do the data contain a homogeneous causal path, or several? Because thorough checking is essentially impossible with more than three explanatory variables, ART is crucial to reliable empirical work. Contrary to the received wisdom, it is not the "too small" regressions on modest subsamples with accompanying plots that should be under suspicion. Instead, the big analyses that use all the observations and have a dozen control variables are the ones that should be met with incredulity.

The result of ART, and other rules like it emerging from the new methodology, would be more careful and appropriate choice of samples and much more detailed attention to what the data really say. Political scientists would develop the intimate knowledge of their observations that would constrain our choice of estimators and discipline our formal theories. The easy proliferation of conceivable estimators discussed above would be limited, since assumptions would have to match up to what we knew about our data. Substantively, too, phony generalizations would be caught more often; truly reliable generalizations would have a fighting chance. Political science would have hope, at least, of developing a firm base of empirical knowledge and substantively relevant econometric estimators on which to build.

Some of these substantive generalizations will have been suggested by theory: We will be searching under the streetlamp. But others will have to come from the darkness, unillumined by theory. Searching in darkness requires more self-discipline than we have mustered thus far. ART is meant to help.

## SUMMARY AND CONCLUSION

This is the way we political methodologists have thought we should proceed: Pick a problem applied researchers care about. Set up some convenient distributional assumptions, mathematically generalizing what has been done before but not worrying overmuch about the corresponding reality. Then hammer the resulting (perhaps messy) likelihood functions or Bayesian posteriors with relentless computing. A careless substantive example may be included for illustration; there is no need to take it seriously. The enterprise is fun, and it looks fancy and sophisticated.

This approach defines the old political methodology. Helpful as it may have been at one stage of our subfield's development, it is now outdated, for it is profoundly atheoretical. Contrary to what those outside the field often believe, inventing new estimators is not very difficult. With a little work and creativity, dozens can be constructed for any class of estimation problem that interests us so long as substantive theory imposes no constraints. What is horribly difficult is to justify the use of a particular estimator in a given social science dataset—not just hand-wave, but truly justify with theory and evidence, so that a fair-minded but skeptical reader would be convinced. That problem has been almost entirely ignored by the old approach, with the result that political methodology has played little or no role in the key empirical discoveries of the past 30 years in political science.

In a more modern view, radical changes in our work habits are needed. Two avenues for justification of our inferences are open to us, neither of which we have exploited well thus far. The first is to develop microfoundations. This approach ties our estimators to formal theory, letting theory decide which assumptions we should make. In particular, it puts a premium on estimators that can be derived rigorously from a formal model of political actors' behavior, perhaps with the addition of white noise disturbances. Then substantive theoretical foundations are not decorative; they are required. Arbitrary, substantively unjustified distributional assumptions are banned.

The second approach applies when theory is unavailable, perhaps the usual case. Then the requirement is that all the assumptions in the analysis be subjected to ruthless data analysis to assess their validity. No more casual assertions of linearity, no more garbage cans of variables from different literatures, no more endless lists of control variables, no more dubious distributions, no more substantively atheoretical, one-size-fits-all estimators to be applied whenever a certain kind of dependent variable or a certain kind of statistical problem appears. Instead, patient data analysis is required—a clear, detailed demonstration in print that in all the parts of the sample, the same model works in the same way, and that the assumptions hold throughout.

Because doing serious data analysis of this kind is demanding work, I have suggested A Rule of Three (ART). No specification with more than three explanatory variables is at all likely to have been checked adequately. Samples should be chosen and, if necessary, pruned so that three control variables are sufficient. Nothing else should be believed.

Political methodology is a very young field. In its early days, onlookers were delighted by every sign of growth and mastery, no matter how modest. Now adolescence has arrived. Necessary and natural as they were at one time, the old work habits and the old goals suddenly look immature. If further development is to occur, then it is time to insist on different standards of achievement. Formal theory and serious data analysis would remake political methodology, and would give us a far better chance than we now have to contribute to the discipline's search for theoretical understanding of politics.

## ACKNOWLEDGMENTS

An earlier version was presented at the Annual Meeting of the American Political Science Association, San Francisco, California, August 29–September 2, 2001. My thanks to many colleagues who attended that panel and made helpful suggestions, including Mike Alvarez, Neal Beck, Henry Brady, Simon Jackman, Gary King, and Jonathan Nagler. Thanks also to Micah Altman, Larry Bartels, David Collier, Jim Granato, John Jackson, Anne Sartori, Phil Schrodt, and John Zaller for recent conversations about the topic of this paper. A fellowship from the Center for the Study of Democratic Politics at Princeton University supported the research, as did the Department of Political Science at the University of Michigan. The paper is dedicated to the memory of my respected Michigan colleague and irreplaceable friend, Harold K. ("Jake") Jacobson, who died unexpectedly while it was being written.

**The *Annual Review of Political Science* is online at http://polisci.annualreviews.org**



## LITERATURE CITED

Achen CH. 1983. Toward theories of data: the state of political methodology. In *Political Science: The State of the Discipline,* ed. A Finifter, pp. 69–93. Washington, DC: Am. Polit. Sci. Assoc.

Achen CH. 1992. Social psychology, demographic variables, and linear regression: breaking the iron triangle in voting research. *Polit. Behav.* 14:195–211

Altman M, McDonald M. 2002. Replication with attention to numerical accuracy. *Polit. Anal.* In press

Alvarez RM, Brehm J. 1995. American ambivalence towards abortion policy. *Am. J. Polit. Sci.* 39:1055–82

Alvarez RM, Brehm J. 1997. Are Americans ambivalent towards racial policies? *Am. J. Polit. Sci.* 41:345–74

Alvarez RM, Brehm J. 2002. *Hard Choices, Easy Answers: Values, Information, and American Public Opinion*. Princeton, NJ: Princeton University Press

Alvarez RM, Nagler J. 1995. Economics, issues and the Perot candidacy. *Am. J. Polit. Sci.* 39:714–44

Alvarez RM, Nagler J. 1998. When politics and models collide: estimating models of multiparty elections. *Am. J. Polit. Sci.* 42:55–96

Aranda-Ordaz FJ. 1981. On two families of transformations to additivity for binary response data. *Biometrika* 68:357–64. Erratum, *Biometrika* 70:303

Bartels LM. 1993. Messages received: the political impact of media exposure. *Am. Polit. Sci. Rev.* 87:267–85

Bartels LM. 1998. Where the ducks are. In

*Politicians and Party Politics*, ed. JG Geer, pp. 43–79. Baltimore, MD: John Hopkins Univ. Press

Bentley AF. 1908. *The Process of Government: A Study of Social Pressures*. Chicago: Univ. Chicago Press

Burgess JW. 1891. *Political Science and Comparative Constitutional Law*. Boston: Ginn

Burr IW. 1942. Cumulative frequency functions. *Ann. Math. Stat.* 13:215–32

Catlin GEC. 1927. *The Science and Method of Politics*. New York: Knopf

Gerber A, Green DP. 1998. Rational learning and partisan attitudes. *Am. J. Polit. Sci.* 42: 794–818

Gowa J. 1999. *Ballots and Bullets*. Princeton, NJ: Princeton Univ. Press

Jackson JE. 1975. Issues, party choices and presidential votes. *Am. J. Polit. Sci.* 19:161–5

Johnson NL, Kotz S, Balakrishnan N. 1994. *Continuous Univariate Distributions*. New York: Wiley

Kramer GH. 1986. Political science as science. In *Political Science: The Science of Politics*, ed. HF Weisberg, pp. 11–23. Washington, DC: Am. Polit. Sci. Assoc.

McKelvey RD, Palfrey TR. 1995. Quantal response equilibria for normal form games. *Games Econ. Behav.* 10:6–38

McLeish DL, Tosh DH. 1990. Sequential designs in bioassay. *Biometrics* 46:103–16

Miller WE. 1999. Temporal order and causal inference. *Polit. Anal.* 8:119–42

Morgan BJT. 1992. *Analysis of Quantal Response Data*. London: Chapman & Hall

Morton RB. *Methods and Models*. Cambridge, UK: Cambridge Univ. Press

Munck GL, Verkuilen J. 2002. Conceptualizing and measuring democracy. *Comp. Polit. Stud.* 35: In press

Nagler J. 1994. Scobit: an alternative estimator to logit and probit. *Am. J. Polit. Sci.* 38:230–55

Prentice RL. 1976. A generalization of the probit and logit methods for dose response curves. *Biometrics* 32:761–68

Robertson T, Cryer JD. 1974. An iterative procedure for estimating the mode. *J. Am. Stat. Assoc.* 69(48):1012–16

Sanders MS. 2001. Uncertainty and turnout. *Polit. Anal.* 9:45–57

Sartori AE. 2002. An estimator for some binary-outcome selection models without exclusion restrictions. *Polit. Anal.* In press

Schultz K. 2001. *Democracy and Coercive Diplomacy*. Cambridge, UK: Cambridge Univ. Press

Signorino CS. 1999. Strategic interaction and the statistical analysis of international conflict. *Am. Polit. Sci. Rev.* 93:279–97

Smith R. 1989. On the use of distributional mis-specification checks in limited dependent variable models. *Econ. J.* 99:178–92 (Suppl: Conf. papers)

Stukel TA. 1988. Generalized logistic models. *J. Am. Stat. Assoc.* 83:426–31

Taylor JMG. 1988. The cost of generalizing logistic regression. *J. Am. Stat. Assoc.* 83: 1078–83

Wu CFJ. 1985. Efficient sequential designs with binary data. *J. Am. Stat. Assoc.* 80:974–84

Zechman MJ. 1979. Dynamic models of the voter's decision calculus. *Public Choice* 34: 297–315

Annual Review of Political Science
Volume 5, 2002

# CONTENTS

INDEXES



ERRATA

An online log of corrections to *The Annual Review of Political Science* chapters (if any have yet been occasioned, 1997 to the present) may be found at http://polisci.annualreviews.org/