Conflict Management and Peace Science, 22:327–339, 2005

ISSN: 0738-8942 print / 1549-9219 online
DOI: 10.1080/07388940500339167

Routledge
Taylor & Francis Group

Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong

CHRISTOPHER H. ACHEN

Department of Politics
Princeton University
Princeton, New Jersey, USA

Many social scientists believe that dumping long lists of explanatory variables into linear regression, probit, logit, and other statistical equations will successfully "control" for the effects of auxiliary factors. Encouraged by convenient software and ever more powerful computing, researchers also believe that this conventional approach gives the true explanatory variables the best chance to emerge. The present paper argues that these beliefs are false, and that without intensive data analysis, linear regression models are likely to be inaccurate. Instead, a quite different and less mechanical research methodology is needed, one that integrates contemporary powerful statistical methods with deep substantive knowledge and classic data–analytic techniques of creative engagement with the data.

Keywords regression analysis, linearity, data analysis, rule of three, monotonicity

Sometimes you can see a lot just by looking.
—attributed to former New York Yankees catcher Yogi Berra

Political researchers have long dreamed of a scientifically respectable theory of international politics. International peace and justice are painfully difficult to achieve, and some of the obstacles have an intellectual character. We do not understand what we most need to know.

In this quest, humanistic, interpretive, and historical methodologies have been profoundly valuable for more than two millennia. They have taught us most of what we know about international politics, and without question we will need their continuing insights for additional progress. Yet these traditional approaches encounter conceptual knots in

This research was partially supported by a sabbatical leave from the Department of Politics, Princeton University. I express my thanks to Jeff Herbst for arranging the leave time and to Sara McLaughlin Mitchell for inviting me to present this paper at the annual meeting of the Peace Science Society, Rice University, Houston, Texas, November 12–14, 2004. Thanks are also due to the many colleagues with whom I have discussed these issues over the years, including Larry Bartels, Jake Bowers, Henry Brady, Bear Braumoeller, Kevin Clarke, David Collier, Rob Franzese, David Freedman, John Jackson, Warren Miller, Bob Powell, Bruce Russett, Anne Sartori, Merrill Shanks, John Zaller, and many others, including several audience members in Houston. I apologize for not citing the many articles each of these notable scholars has written that have contributed to my understanding and argument. Bear Braumoeller, Doug Rivers, and Anne Sartori sent me helpful written comments on an earlier draft and saved me from careless language and misstatements, as did three anonymous reviewers. Of course, remaining errors are my own.

Address correspondence to Christopher H. Achen, Department of Politics, Princeton University, Princeton, NJ 08544, USA. E-mail: achen@princeton.edu

international politics that appear deeper than those in many other parts of political science. Game theory has exposed these counterintuitive aspects of reality. Mathematical modeling, more analytically powerful than human intuition and wisdom on their own, seems certain to become an integral part of the long-run intellectual progress that will reduce the scourge of war.

Yet game theory alone is insufficient. We are far enough along now in the study of international politics to see that there is no end to the making of formal models. Each little mathematical twist leads to another paper, complete with its own set of apparently supportive historical cases and ending with yet another slant on reality. The insights from each such effort range from the profound to the trivial, and researchers cannot always agree on which is which. Abstract theory on its own, however powerful, may be good applied mathematics, but it is not science. Once one has learned the mathematics, as the distinguished formal theorist Gerald Kramer (1986) remarked, theorizing is relatively easy. What is so much harder is the sorting out: Which theories tell us something consequential about the world?

This is where statistical analysis enters. Validation comes in many different forms, of course, and much good theory testing is qualitative in character. Yet when applicable, statistical theory is our most powerful inductive tool, and in the end, successful theories have to survive quantitative evaluation if they are to be taken seriously. Moreover, statistical analysis is not confined to theory evaluation. Quantitative analysis also *discovers* empirical generalizations that theory must account for. Scientific invention emerges from data and experiment as often as data and experiment are used to confirm prior theory. In international relations, the empirical finding (if that is what it is) that democracies do not fight each other has led to a great deal of intriguing theorizing. But all the theory is posterior to the raw empirical discovery.

How is all this empirical creativity and validation to be achieved? Most empirical researchers in international politics, as in the rest of the discipline, believe that they know the answer. First, they say, decide which explanations of a given phenomenon are to be tested. One or more such hypotheses are set out. Then "control variables" are chosen—factors which also affect the phenomenon under study, but not in a way relevant to the hypotheses under discussion. Then measures of all these explanatory factors are entered into a regression equation (linearly), and each variable is assigned a coefficient with a standard error. Hypotheses whose factors acquire a substantively and statistically significant coefficient are taken to be influential, and those that do not are treated as rejected. Extraneous influences are assumed to be removed by the "controls."

Minor modifications may be made in carrying out this conventional research routine. Corrections may be made for heteroskedasticity or serial correlation. Asymptotically robust standard errors may be computed. Probit or logit may be used for discrete dependent variables, and duration models may be employed when lengths of time are to be explained. Lagged independent and dependent variables may appear in time series contexts, and models for counts may be used when counted quantities are to be explained.

Each of these techniques makes special statistical assumptions, and yet the models share a common framework. In the great majority of applied work with all these methods, a particular statistical distribution is specified for the dependent variable, conditional on the independent variables. The explanatory factors are postulated to exert their influence through one or more parameters, usually just the mean of the statistical distribution for the dependent variable. The function that connects the independent variables to the mean is known as the "link function." Thus if μ is the mean of the dependent variable and if the link $g(.)$ is a (typically nonlinear) function, we write $\mu = g(x_1, x_2, \ldots, x_k)$, where the x_j's are the explanatory factors.

In practice, researchers nearly always postulate a linear specification as the argument of the link function. That is, in ordinary regression, probit and logit, duration and count models, the independent variables usually enter the link in the standard linear form: $\mu = g(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k)$. Computer packages often make this easy: One just enters the variables into the specification, and linearity is automatically applied. In effect, we treat the independent variable list as a garbage can: Any variable with some claim to relevance can be tossed in. Then we carry out least squares or maximum likelihood estimation (MLE) or Bayesian estimation or generalized method of moments, perhaps with the latest robust standard errors. It all sounds very impressive. It is certainly easy: We just drop variables into our mindless linear functions, start up our computing routines, and let 'er rip.

James Ray (2003a, 2003b) has discussed several ways in which this garbage-can approach to research can go wrong, even in the simplest cases. First, researchers may be operating in a multi-equation system, perhaps with a triangular causal structure. For example, we may have three endogenous variables, with this causal order: $y_1 \rightarrow y_2 \rightarrow y_3$. If so, then y_1 has an *indirect* impact on y_3 (via y_2), but controlled for y_2, it has no *direct* impact. Ray emphasizes that if researchers want to know indirect causal effects (or *total* effects = direct + indirect), then they cannot run regressions such as

$$y_3 = \alpha + \beta_1 y_1 + \beta_2 y_2 + u. \tag{1}$$

For then they will get the wrong answer. Under the appropriate conditions, the estimated coefficient $\hat{\beta}_1$ will represent the *direct* effect of y_1, and that estimate will converge to zero in this case, even though the total effect may be substantial.[1] If a researcher foolishly concludes from this vanishing coefficient that the total effect of y_1 is zero, then, of course, a statistical error has been committed. It may be worth knowing for descriptive purposes that a variable like y_2 is correlated with the dependent variable, but as Ray says, that does not mean that it belongs in a regression as a control factor.[2]

The best solution is simply for researchers to be familiar with multi-equation systems and to recognize that their regressions yield only direct effects of right-hand-side variables. Put another way, any researcher intending to interpret regression (or probit, logit, etc.) coefficients as total effects has to be prepared to say, "It is obvious that none of my independent variables cause each other in any substantial way." If that statement is nonsensical, then the researcher is no longer in the single-equation world. Usually, ordinary regression then will be inappropriate, as any econometrics text explains.[3]

Ray (2003a, 2003b) also cautions sensibly against putting multiple measures of the same causal factor into a regression, against using exaggerated measures of causal impact, and against a host of other statistical sins visible in the journals. All these points are made clearly and deserve to be heard. However, Ray spends less time on another issue that seems to me of equal importance, and it is the one I wish to take up.

[1] Of course, triangular systems of equations can be estimated by ordinary least squares only if their disturbances are uncorrelated—the "hierarchical" case.

[2] A reviewer suggested that "descriptive" be substituted for "direct"effect. But since direct effects can be large when the descriptive correlation is zero, and vice-versa, I have not adopted this advice.

[3] The distinction here is between *multiple* regression, which means "more than one independent variable (other than an intercept)," and a *multivariate* statistical method, which means"more than one dependent variable. Thus simultaneous equation estimation, factor analysis, and scaling techniques are all multivariate techniques, but "multivariate regression" is a misnomer when applied to a single regression. The term "multivariate regression" applies when regression methods are used to estimate an entire system of equations, as in seemingly unrelated regressions (SUR) and related techniques.

Monotonic Relationships and Linearity

In this paper, my central question is this: When researchers actually do need to control for certain variables, do linear specifications accomplish that task?

Now, of course, no linear specification will work if the relationship is quadratic or nonmonotonic in some other way. So let us assume that problem away and imagine that all the variables in our link functions are monotonically related (positively or negatively) to the dependent variable, controlled for everything else.[4] That is, we are discussing the case in which, no matter what the values of other variables are, an increase in the value of any of the independent variables always leads to an (expected) increase in the dependent variable, or else it always leads to an (expected) decrease in the dependent variable. That is what is meant by (strict) positive and negative monotonicity in the mean of the dependent variable, respectively. This is the kind of relationship most researchers have in mind when they turn to statistical work: Do tighter alliances lead to more war? Does more trade lead to less war? Does more democracy lead to fewer militarized interstate disputes? All these are questions about *conditional monotonic* relationships: Conditional on the other explanatory variables, the expected effect of the independent variable is monotonic. Indeed, we rarely have intuitions about linearity in substantive problems. Monotonicity is what we understand best.

In practice, conditional monotonic relationships are nearly always modeled with linear link functions, as we have noted. Linear links assume that relationships are conditionally monotonic, but they also assume something more. They assume that the relationships are conditionally *linear*, a stronger statement than conditional monotonicity. Linearity requires that a one unit increase in a given independent variable always lead to the *same* expected change in the dependent variable, no matter what the values of the other independent variables. Monotonicity requires only that a one unit change lead to *some* change in the dependent variable, always in the same direction, no matter what the values of the other independent variables.

In practice, we just assume that linearity is a good approximation to monotonicity. I am not sure that we think about this very much, and I am certain that econometrics textbooks discuss it far too little, if they mention it at all. Implicitly, we treat the difference between monotonicity and linearity as unimportant. That is, we assume that the following First Pseudo-Theorem is true.

First Pseudo-Theorem: Dropping a list of conditionally monotonic control variables into a linear link function controls for their effects, so that the other variables of interest will take on appropriate coefficients.

But is this pseudo-theorem true? No doubt it's not *exactly* true. But is it pretty close to being true, so that we're unlikely to be misled in practice?

A closely related notion turns up in hypothesis testing. When researchers need to test which of several hypotheses is correct and they have an independent variable measuring each of them, they typically ask themselves whether the effects of the hypotheses are conditionally monotonic. If so, then they assume that a linear specification or linear link function will sort out which hypothesis is correct. Implicitly, most researchers assume that the following Second Pseudo-Theorem is approximately correct.

Second Pseudo-Theorem: Dropping a list of conditionally monotonic variables into a linear link function assigns each of them their appropriate explanatory impact, so that the power of each hypothesis can be assessed from its coefficient and standard error.

Again, no one imagines that this theorem is precisely correct. The issue is whether it's close enough for government department work. In practice, nearly all of us assume nearly all of the time that the pseudo-theorem is very nearly correct.

[4]Explanatory variables are sometimes used to explain other parameters, such as variances. Such cases raise the same concerns as those discussed below for regression analysis.

Why are these pseudo-theorems so important? The answer is straightforward: If they are approximately true, then most of the empirical work done by social scientists is reliable science. But if they are not true, then most of the statistical work appearing in the journals is under suspicion. And if the problem is sufficiently drastic—for example, if linear link functions can make conditionally monotonic variables with positive effects have statistically significant negative coefficients—then garbage-can regressions, garbage-can probits and logits, and garbage-can MLE and Bayesian estimators are not good science. It would follow that drastic changes in ordinary statistical practice in political science would be called for.

Can Linear Links Make Good Variables Go Bad?

With patience, anyone can concoct a large N, highly nonlinear problem in which regression analysis fails to get the right signs on the coefficients. The trick is just to put a few outliers in places where a variable has high leverage. Examples of that kind do not bear on the question asked here.

Instead, we seek a small problem where graphs will show us that the data are reasonably well behaved and that no egregious outliers occur. The closer the true fit, the better. Then we introduce a small amount of nonlinearity and assess the effects. Might the problem be bad enough in a data set with 15 observations, for example, that we get a fairly large, statistically significant coefficient of the wrong sign?

The data set I will use is given in Table 1. The independent variables x_1 and x_2 are the true explanatory factors. The variable x_1 has a slightly nonlinear effect on the dependent variable y, and it enters the equation in the form $z = f(x_1)$. To avoid outliers and to eliminate any possibility of stochastic accidents, when z and x_1 are entered, the regression fit is perfect: $R^2 = 1$. Thus the dependent variable y is constructed in a very simple linear way to ensure that its relationship to the independent variables involves nothing unusual:

$$y = z + 0.1x_2. \tag{2}$$

TABLE 1 Data

Obs.	z	x_1	x_2	y
1	0	0	0	0
2	0	0	1	.1
3	0	0	1	.1
4	1	3	1	1.1
5	1	3	1	1.1
6	1	3	1	1.1
7	2	6	2	2.2
8	2	6	2	2.2
9	2	6	2	2.2
10	8	9	2	8.2
11	8	9	2	8.2
12	8	9	2.1	8.21
13	12	12	2.2	12.22
14	12	12	2.2	12.22
15	12	12	2.2	12.22

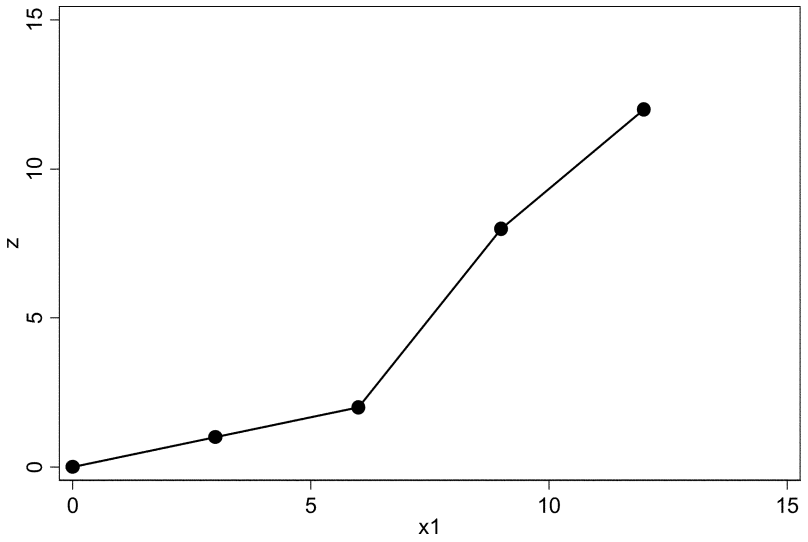

FIGURE 1 The function $z = f(x1)$.

Since the dependent variable in this regression is exactly predicted by the independent variables, if these variable are used in a regression equation of the form

$$y = \alpha + \beta_1 z + \beta_2 x_2 + u, \tag{3}$$

then the correct estimates $\hat{\alpha} = 0$, $\hat{\beta}_1 = 1$, and $\hat{\beta}_2 = 0.1$ are returned, with $R^2 = 1.0$, as the reader can verify using the data in Table 1.

Now suppose that, as usual in practical applications, the form of the function $f(x_1)$ is not known. That is, we do not have the variable z. Instead, we have x_1, whose effect on y is slightly nonlinear and unknown. The equation to be estimated is then

$$y = \alpha + \beta_1 f(x_1) + \beta_2 x_2 + u, \tag{4}$$

where the function $f : x_1 \rightarrow z$ is to be determined from the data.

The actual function f used to construct Table 1 is plotted in Figure 1. As is obvious, x_1 and $f(x_1)$ are nearly the same variable. The form of f is only slightly nonlinear, implying that the relationship of x_1 to the dependent variable is only slightly nonlinear. Moreover, inspection of Table 1 shows that in the true form of the relationship, both x_1 and x_2 are conditionally monotonically related to y. Thus we are meeting all the conditions in the pseudo-theorems. It ought to be safe to just use both x_1 and x_2 linearly and forget about the unknown, small amount of nonlinearity. The pseudo-theorems should save us. But do they?[5]

If we take the conventional approach of treating monotonicity as linearity and just use x_1 to replace z, the bivariate regressions (with standard errors in parentheses) look fine: They show the expected positive relationships, with R^2 of .89 and .53, respectively, and all

[5]There is another way to interpret this situation, too. As Table 1 shows, the variable z takes on just five values: 0, 1, 2, 8, 12. These might have come to the researcher in the form Conciliate, Warn,

slopes highly statistically significant with t-ratios above 4:

$$y = -1.524 + 1.047x_1, \quad (.7069) \quad (.0962) \tag{5}$$

$$y = -3.680 + 5.341x_2. \quad (2.233) \quad (1.305) \tag{6}$$

Thus nothing seems amiss except a bit of imperfection in the fit—actually, quite a bit less imperfection than in many international politics data sets! Everything appears positively monotonic as it should, and the bivariate relationships look right. We have only deviated a little bit from linearity, and no harm has befallen us. So far, so good.

Now if the pseudo-theorems are correct, when the dependent variable is regressed on both x_1 and x_2, we ought to get two positive, statistically significant coefficients. Indeed, when this regression is carried out, statistical significance does hold comfortably for both slopes, and the adjusted R^2 rises to .92. Alas, though, a disaster occurs. The coefficient on x_2 is now substantially and statistically significant ($t = -2.50$), but it has become 28 times larger in magnitude than its true value. Worse, it has the wrong sign:

$$y = 0.5888 + 1.427x_1 - 2.780x_2. \quad (1.034) \quad (.1722) \quad (1.111) \tag{7}$$

What is particularly odd here is that the messed-up coefficient applies to x_2, the independent variable without nonlinearities.

Nothing in this finding depends on doing one regression or having a small sample. If one prefers the less theoretical, computer-intensive style of simulating large numbers of regression runs, it is easy to come to the same conclusion. Just treat Table 1 as the joint distribution of x_1, x_2, and y, with each of the 15 observations equally likely. Then draw from this distribution under independent random sampling, and compute regression equations with various sample sizes. It is easy to prove that the coefficients will converge to those given in the preceding equation and that t-ratios will become arbitrarily large as the sample size goes to infinity. That is, with enough data, the coefficient on x_2 is essentially *always* large and of the wrong sign.

In short, both pseudo-theorems are false. Garbage-can regressions, whether used to control for extraneous factors or to test competing hypotheses, just do not work. Not all empirical work with small nonlinearities comes out as poorly as this example, of course.

Threaten, Display Force, Begin War. Alternately, they might have arrived as Strongly Disagree, Disagree, Not Sure, Agree, Strongly Agree. The point is that the appropriate numerical codes for these categories would not be obvious. Not knowing the true values, the researcher might have chosen equally spaced values for the variable such as 0, 1, 2, 3, 4, or an equivalent linearly transformed, interval-level scale such as 0, 3, 6, 9, 12. To keep close to the original range of z, suppose that we adopt the latter scale and call the recoded variable x_1. Figure 1 shows the difference between the true original variable z and its coded version x_1. Thus on this view, we have small coding errors and thus a little measurement error, with consequences well known from econometrics texts. But minor errors in coding ordinal variables are customarily thought to have small effects on regression coefficients (for example, Abelson & Tukey, 1970), and thus our optimism under this interpretation parallels our expectations when we deal with small amounts of nonlinearity. The two viewpoints are mathematically equivalent. Doug Rivers has mentioned to me that working from this measurement-error perspective, David Grether found results similar to this paper's in the early 1970s, but I have not been able to find the reference.

Some of the time, perhaps even most of the time, our results are "kinda, sorta" right. But there are no guarantees. Even with small amounts of unrecognized nonlinearity, as in this problem, violently incorrect inferences can occur.

What to Do?

Small nonlinearities creep into everything we do. So do big nonlinearities. No one can be certain that they are not present in a given data set. If these nonlinearities are as dangerous as I believe they are, what can be done about avoiding the threats to scientific accuracy that they present?

Part of the answer is formal theory. Its power to direct one's attention to the right statistical model remains less recognized in political science than it should be. Knowing how Bayesian theory works, for example, allowed Bartels (2002) to discover errors in the seemingly persuasive informal logic with which public opinion researchers have treated opinion change. Signorino (1999) and Signorino and Yilmaz (2003) have shown how surprising statistical nonlinearities and nonmonotonicities are implied by rational choice models of international crisis bargaining behavior, and Sartori (2003) has proposed an entirely new statistical estimator for data subject to selection bias, based on her formal model of crisis bargaining. Thus when formal models are available, the analyst is not free to dump bags of variables into some garbage-can statistical setup. Instead, nonlinearities are expected, the analyst knows where to look for them, and so they are detected and modeled statistically.

Even with a formal model, however, careful data analysis is required. Few formal models specify the precise functional form for statistical analysis, so that some empirical investigation is needed. And when no formal model is available, the most common situation, then very careful inspection of the data is essential.

Consider, for example, the small empirical problem already analyzed in this paper. Figures 2 and 3 show the simple bivariate plots of y against x_1 and x_2. Both cases show evidence of slight nonlinearity, the usual sign that carelessly dropping variables into a canned regression program would be dangerous.

Figuring out what is wrong here without knowing the true function $f(x_1)$ would be no trivial matter. However, since so much of what we know empirically, in international

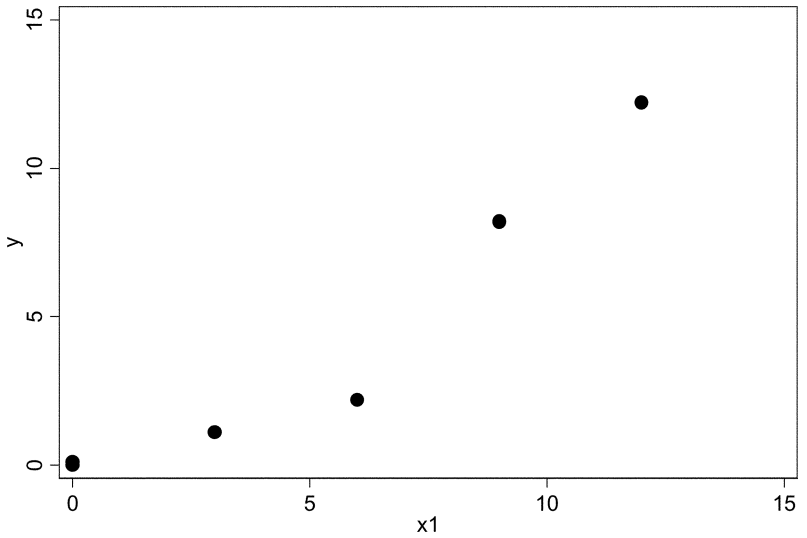

FIGURE 2 y vs. $x1$.

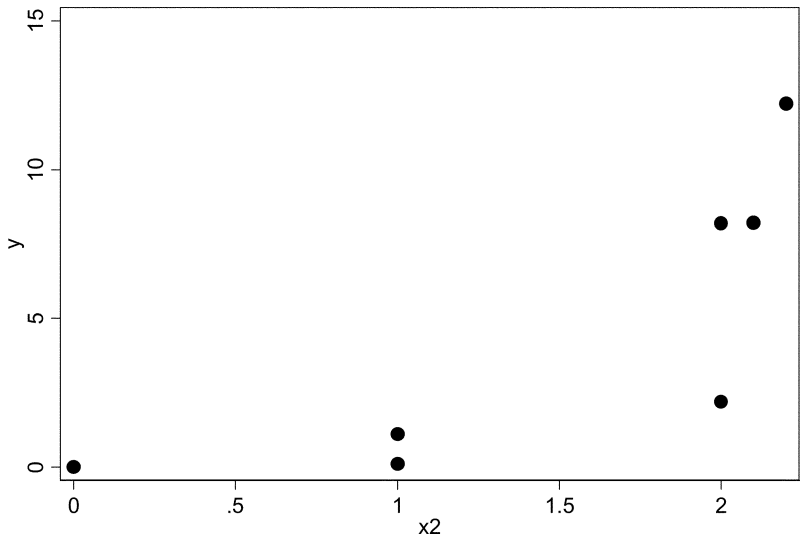

FIGURE 3 y vs. $x2$.

politics and elsewhere in political science, derives from cross-tabulations, a sensible first step might be to tabulate mean values of y by the two independent variables. Table 2 results.

Patient inspection of these data shows that changes in x_2 cause the same proportional increase in the dependent variable whatever the fixed value of x_1, while the reverse is not true. Hence the nonlinearity is confined to x_1. After discovering that the obvious nonlinear fixes (logs, exponentials, quadratics) do not eliminate the problem, the researcher might try a fully flexible transformation for x_1—a regression with added dummy variables for the three middle categories of x_1.[6] (The remaining two categories are left fixed to set the scale.)

The extra three dummies create a total of five independent variables (measuring just two actual independent factors) plus the constant. That regression fits perfectly—the decisive evidence that the two variables have separable effects with respect to each other (no interactions), but that the form of the x_1 effect is slightly nonlinear. We would be able to infer the correct nonlinear transformation f (namely, the true z) from the dummy coefficients. Finally, we could regress y on z and x_2. That regression would fit perfectly, all the variables would have the correct coefficients, and we would be done.

The central point of this paper is that the pseudo-theorems are incorrect, and that is what the empirical example is meant to show. However, another implication follows as well, one not immediately related to the two-variable example above. That implication follows from the complexity of the data analysis in the example. Getting right that simple little equation with its one small nonlinearity is perfectly possible. But it is no trivial matter of ten minutes' effort. Since the function f is not known, it would take time and patient effort to find it. Without that effort, the statistical results from the data are not just useless, but actively misleading. And this problem has just two independent variables! Three would create a serious burden for the data analyst, and four effectively puts the problem outside the range of what most of us have time to do right. Probit and logit are even harder because dichotomous dependent variables are so much more noisy than continuous variables and

[6]This table also shows how little real statistical information about the true specification exists in this dataset in spite of all the statistically significant coefficients that emerge from it. I have drawn the same inference looking at various international politics datasets. Peace science is simply a difficult subject for quantitative work.

TABLE 2 Mean value of y

	x_2				
	0	1	2	2.1	2.2
x_1	0	0	0.1		
	3	1.1			
	6		2.2		
	9		8.2	8.21	
	12				12.22

thus contain much less information. In short, as I have argued elsewhere, the kind of careful work needed to get the correct answer in most practical substantive problems in political science can only be carried out when one has three or fewer independent variables (Achen, 2002).[7]

Now it goes without saying that most of us use more than three explanatory variables and that virtually all of us have done so at some stage of our careers, the present author included. For example, inspection of the last two issues (August and October, 2004) of a premier quantitative international politics outlet, the *Journal of Conflict Resolution*, yields several articles with regressions and probits using eight, ten, or more independent variables, occasionally as many as fifteen or eighteen, and in two instances, more than twenty. It also goes without saying that we customarily put much less time than needed into graphical analysis, partial regression plots, nonlinear fitting, and all the other details that are needed to obtain reliable results. Indeed, it is nearly impossible to do so with large specifications. We believe the two pseudo-theorems will save us. Unfortunately, they will not. No wonder our coefficients zip around when we take a few variables in and out of our big regressions and probits. The truth is that in such cases, the coefficients are virtually all unreliable anyway.

A simple conclusion follows: We need to stop believing much of the empirical work we've been doing. And we need to stop doing it that way.

Conclusion

The argument of this paper is that linear link functions are not self-justifying. Garbage-can lists of variables entered linearly into regression, probit, logit, and other statistical models have no explanatory power without further argument. Just dropping variables into SPSS, STATA, S, or R programs accomplishes nothing, no matter how high-powered or novel the estimators. In the absence of careful supporting argument, the results belong in the statistical rubbish bin.

What should the supporting argument for a statistical specification consist of? As I argued above, giving credibility to a statistical specification, linear or otherwise, requires at least one of two supports—either a formal model or detailed data analysis. In the first case, researchers can support their specifications by showing that they follow as a matter of rigorous mathematical inference from their formal model—not just from distributional assumptions, but from a legitimate game-theoretic model or other microfoundation. This is always the most impressive support that a statistical model can receive. Though one has to guard against the risk of compounding any limitations in the formal model by imposing

[7]Of course, atheoretical forecasting and testing methodologies such as autoregressive integrated moving average (ARIMA) modeling, vector autoregressions (VAR), or Granger causality testing are immune from this constraint.

them on the data, nonetheless, integrating formal theory and statistical model puts to rest a host of uncertainties about the specification, as Most and Starr (1984, 1989) emphasized long ago. As noted above, better tools for doing so in international relations are under rapid development.

When no formal theory is available, as is often the case, then the analyst needs to justify statistical specifications by showing that they fit the data. That means more than just "running things." It means careful graphical and cross-tabular analysis. As Anscombe (1973) demonstrated long ago, precisely the same regression output can occur from wildly different data sets with quite divergent causal patterns. Only graphical analysis, cross-tabulation, and regression diagnostics will uncover the truth.

Is the effect really there in all parts of the data? Does it actually work the same way for all the observations? Are there parts of the data in which the competing hypotheses imply opposite results, so that we can carry out a classic critical test? And if we intend to apply a linear model with constant coefficients, are the effects really linear and the same size in all the parts of the data? Show us! If we have not discussed and answered these questions in our articles, no one should believe our work. In other words, we have to think a little more like an experienced chef adjusting the broth as he cooks, and less like a beginner blindly following the recipe whether it suits the ingredients at hand or not.

Suitable texts at various levels have long been available that show us how to do scientifically serious data analysis for regression (for example, Berk, 2004; Cook & Weisberg, 1994; Cleveland, 1992; Hamilton, 1992), but we ignore them too often. And when our data sets have a variety of statistical regimes in them, as Ragin (1987) has argued is often the case, then a variety of new and old methodologies are available to analyze them properly (for instance, Braumoeller, 2003; Breiman et al., 1984). But these methods, too, often go unused, even when the researcher's verbal presentation of the argument makes it clear that they are necessary. We need to do better.

When I present this argument to political scientists, one or more scholars (sometimes even my former students) say, "But shouldn't I control for everything I can in my regressions? If not, aren't my coefficients biased due to excluded variables?" This argument is not as persuasive as it may seem initially. First of all, if what you are doing is misspecified already, then adding or excluding other variables has no tendency to make things consistently better or worse. (For rigorous and enlightening treatment of this point, see Clarke, 2005.) The excluded variable argument only works if you are sure your specification is precisely correct with all variables included. But no one can know that with more than a handful of explanatory variables.

Still more importantly, big, mushy linear regression and probit equations seem to need a great many control variables precisely because they are jamming together all sorts of observations that do not belong together. Countries, wars, racial categories, religious preferences, education levels, and other variables that change people's coefficients are "controlled" with dummy variables that are completely inadequate to modeling their effects. The result is a long list of independent variables, a jumbled bag of nearly unrelated observations, and often a hopelessly bad specification with meaningless (but statistically significant with several asterisks!) results.

A preferable approach is to separate the observations into meaningful subsets—internally compatible statistical regimes. That is, the data should be divided into categories in which theory or experience or data analysis suggests that the coefficients are similar. A great many dummies and control variables can then be discarded because they are not needed within each homogeneous regime, where the controls all have the same value. In effect, the subset of observations are "matched" in a substantively meaningful way. The

result is a small, simple, coherent regression, probit, or logit whose observations can be looked at with care, whose effects can be modeled with no more than a handful of independent variables and whose results can be believed. If this can't be done, then statistical analysis can't be done. A researcher claiming that nothing else but the big, messy regression is possible because, after all, *some* results have to be produced, is like a jury that says,"Well, the evidence was weak, but *somebody* had to be convicted."

Oneal and Russett (2005) suggest that scholarship on the democratic peace has developed in this way. Naive linear specifications have been replaced by more sophisticated nonlinear and interactive models that eliminate more competing hypotheses. That is precisely the direction of research that the argument of this paper supports, when careful data analysis supports the nonlinearities. For the point of this paper is not that linear models are somehow wrong, nor that more elaborate nonlinear curve-fitting will save us. To the contrary, the more creative our simple models and the fewer elaborate, nonlinear mechanical canned statistical outputs we generate, the wiser we will be. But whatever our methods, as Oeal and Russett correctly insist, our first loyalty must be to detailed substantive knowledge of our observations. As a profession, that is what we are not doing very well now.

In sum, we need to abandon mechanical rules and procedures. "Throw in every possible variable" won't work; neither will "rigidly adhere to just three explanatory variables and don't worry about anything else." Instead, the research habits of the profession need greater emphasis on classic skills that generated so much of what we know in quantitative social science: plots, crosstabs, and just plain looking at data. Those methods are simple, but sophisticatedly simple. They often expose failures in the assumptions of the elaborate statistical tools we are using, and thus save us from inferential errors. Doing that kind of work is slow, and it requires limiting ourselves to situations in which the number of explanatory factors is small—typically no more than three. But restricting ourselves to subsets of our data where our assumptions make sense also typically limits us to cases in which we need only a handful of explanatory factors, and thus where our minds can do the creative thinking that science is all about. Far from being a limitation, therefore, small regression specifications limited to homogeneous subsets of the data (and their probit and logit equivalents) are exactly where our best chances of progress lie.

References

Abelson, R. P., and J. W. Tukey. 1970. Efficient conversion of non-metric information into metric information. *The quantitative analysis of social problems*, ed. E. R. Tufte, 407–417. Reading, MA: Addison-Wesley.

Achen, C. H. 2002. Toward a new political methodology: Microfoundations and ART. *Annual Review of Political Science* 5: 423–450.

Anscombe, F. J. 1973. Graphs in statistical analysis. *The American Statistician* 27: 17–21.

Bartels, L. M. 2002. Beyond the running tally. *Political Behavior* 24(2): 117–150.

Berk, R. A. 2004. *Regression analysis*. Thousand Oaks, CA: Sage.

Braumoeller, B. F. 2003. Causal complexity and the study of politics. *Political Analysis* 11: 209–233.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and regression trees*. Belmont, CA: Wadsworth.

Clarke, K. 2005. Return of the phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science* 22: this volume.

Cleveland, W. S. 1992. *The elements of graphing data*. Dordrecht, Netherlands: Kluwer.

Cook, R. D., and S. Weisberg. 1994. *An introduction to regression graphics.* New York: Wiley.

Hamilton, L. C. 1992. *Regression with graphics*. Pacific Grove, CA: Duxbury.

Kramer, G. H. 1986. Political science as science. In *Political science: The science of politics*, ed. H. F. Weisberg, 11–23. Washington, DC: American Political Science Association.

Most, B. A., and H. Starr. 1984. International relations theory, foreign policy substitutability, and "nice" laws. *World Politics* 36: 383–406.

Most, B. A., and H. Starr. 1989. *Inquiry, logic and international relations*. Columbia, SC: University of South Carolina Press.

Oneal, J. R., and B. Russett. 2005. Rule of three, let it be? When more really is better. *Conflict Management and Peace Science* 22: 293–310.

Ragin, C. C. 1987. *The comparative method*. Berkeley: University of California Press.

Ray, J. L. 2003a. Explaining interstate conflict and war: What should be controlled for? Presidential address to the Peace Science Society, University of Arizona, Tucson, November 2, 2002.

Ray, J. L. 2003b. Constructing multivariate analyses (of dangerous dyads). Paper prepared for the annual meeting of the Peace Science Society, University of Michigan, Ann Arbor, Michigan, November 13.

Sartori, A. E. 2003. An estimator for some binary outcome selection models without exclusion restrictions. *Political Analysis* 11(2): 111–138.

Signorino, C. S. 1999. Strategic interaction and the statistical analysis of international conflict. *American Political Science Review* 93: 279–297.

Signorino, C. S., and K. Yilmaz. 2003. Strategic misspecification in regression models. *American Journal of Political Science* 47: 551–566.