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

Christopher H. Achen

Department of Politics
Princeton University
Princeton, NJ 08544
achen@princeton.edu

Prepared for presentation at the
annual meeting of the Peace Science Society,
Rice University, Houston, Texas,
November 12-14, 2004

November 11, 2004

Abstract

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 statistical models with more than a few independent variables are likely to be inaccurate. Instead, a quite different research methodology is needed, one that integrates contemporary powerful statistical methods with classic data–analytic techniques of creative engagement with the data.

1 Introduction[1]

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 international politics that appear deeper than those in many other parts of political science. Game theory has exposed these counter–intuitive aspects of reality. Methodologies like game theory, more analytically powerful than human experience and wisdom on their own, seem 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

[1]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 Peace Science meetings. 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, David Collier, Rob Franzese, David Freedman, John Jackson, Warren Miller, Bob Powell, Bruce Russett, Anne Sartori, Merrill Shanks, John Zaller, and many others. Anne Sartori suggested the adjective "garbage–can." I apologize for not citing the many articles each of these notable scholars has written that have contributed to my understanding and argument.

math, 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 discovery 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 discovery and validation to be carried out? 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 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 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. In the great majority of applied work, a particular statistical distribution is assumed for the dependent variable, conditional on the independent variables, and these statistical distributions differ from one application to another. However, each statistical setup has a common structure. The explanatory factors are assumed to exert their influence through one or more parameters, usually just the mean of the statistical distribution. The function that connects the independent variables to the mean is known as the "link function."

In each of these statistical frameworks, researchers nearly always use one or another version of a linear specification for the link function. Most often, as with ordinary regression, probit, and logit, the mean of the distribution is assumed linear in the independent variables.[2] In other cases, such as most duration or count models, as well as heteroskedastic probit, the *log* of some parameter is typically assumed linear in the explanatory factors. But in all such cases, the key parameter is linked to the explanatory variables in an essentially linear way. Computer packages make this easy: One just enters the variables into the specification, and linearity is automatically applied. Then we carry out least squares or maximum likelihood estimation 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 linear computing routines and let 'er rip.

James Ray (2003a, 2003b) has discussed several ways in which this research procedure can go wrong. First, researchers may be operating in a

[2] In the logit and probit cases, of course, it is the mean on the underlying scale that is modeled, not the mean of the dichotomous or polychotomous observed variable. See any introductory econometrics text for an explanation of this distinction.

multi–equation system, perhaps with a triangular causal structure. For example, we may have three endogenous variables, with this causal structure: $y_1 \rightarrow y_2 \rightarrow y_3$. 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 (or *total*) effects, but confuse them with direct effects and so run regressions such as:

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

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.[3] 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. As Ray says, just because a variable is correlated with the dependent variable does not mean that it belongs in a regression as a control factor.

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.[4]

Ray (2003a, 2003b) also cautions sensibly against using multiple mea-

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

[4] The distinction here is between *multiple* regression, which means "more than one independent variable," 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. "Multivariate regression" applies when regression methods are used to estimate an entire system of equations, as in seemingly unrelated regressions (SUR) and related techniques.

sures of the same causal factor in regressions, against 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.

2 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 non–monotonic 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 parameter being modeled, controlled for everything else. That is, when the variable goes up, the parameter to which it is linked always rises, too; or else when the variable goes up, the parameter always goes down. That is what is meant by (strict) positive and negative monotonicity, respectively.

For concreteness, let us suppose that the relevant parameter is the mean of the distribution, as in regression, probit, and logit, so that we can talk about monotonic relationships *to the dependent variable*. Then we will discuss the case in which, no matter what the values of other variables, 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.[5] 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?

[5] "Expected," of course, because statistical relationships are stochastic. In the regression case, this reduces to saying that the regression has a disturbance term, and so no forecast is perfect.

Does more democracy lead to fewer militarized interstate disputes? All these are questions about *conditional monotonic* relationships: Conditional on the other explanatory variables, the 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 relationship is conditionally *linear,* a stronger statement than conditional monotonicity. Linearity requires that a one unit increase in each independent variable lead to the *same* expected change in the dependent variable, no matter what the values of any of the independent variables, including the one whose value is being changed. 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 any of the independent variables including the one whose value is being changed.

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 Gigantic Pseudo–Theorem is true:

First Gigantic 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 Gigantic Pseudo–Theorem is approximately correct:

Second Gigantic 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.

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.

3 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 dataset with 15 observations, for example, that we get a fairly large, statistically significant coefficient of the wrong sign?

The dataset I will use here is given in Table 1. The variables x_1 and x_2 are the true explanatory factors. To avoid outliers and to eliminate any possibility of stochastic accidents, no disturbance term is used. The dependent variable y is constructed in a very simple linear way to ensure that its relationship to the independent variables involves nothing unusual:

$$y = x_1 \ + 0.1x_2 \tag{2}$$

Thus the fit of this regression equation is perfect. If these variable are used in a regression equation of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u \tag{3}$$

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.

Table 1: Data

Obs.	x_1	$\hat{x}_1$	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

Now as Table 1 shows, the variable x_1 takes on just five values: 0, 1, 2, 8, 12. These might have come to the researcher in the form Conciliate, Warn, 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. The researcher might have chosen equally spaced values such as 0, 1, 2, 3, 4, or transforming linearly to an equivalent equally-spaced scale with the same range as the original variable, the researcher might have used 0, 3, 6, 9, 12. We call this latter recoded variable $\hat{x}_1$. As Figure 1 shows, the difference between the true original variable and its coded version is quite mild, and the monotonic relationship between them should ensure (and does in the dataset at hand) that the regression equation remains conditionally monotonic in each independent variable, as desired.

Another way the relationship in Equation 1 might have occurred is not

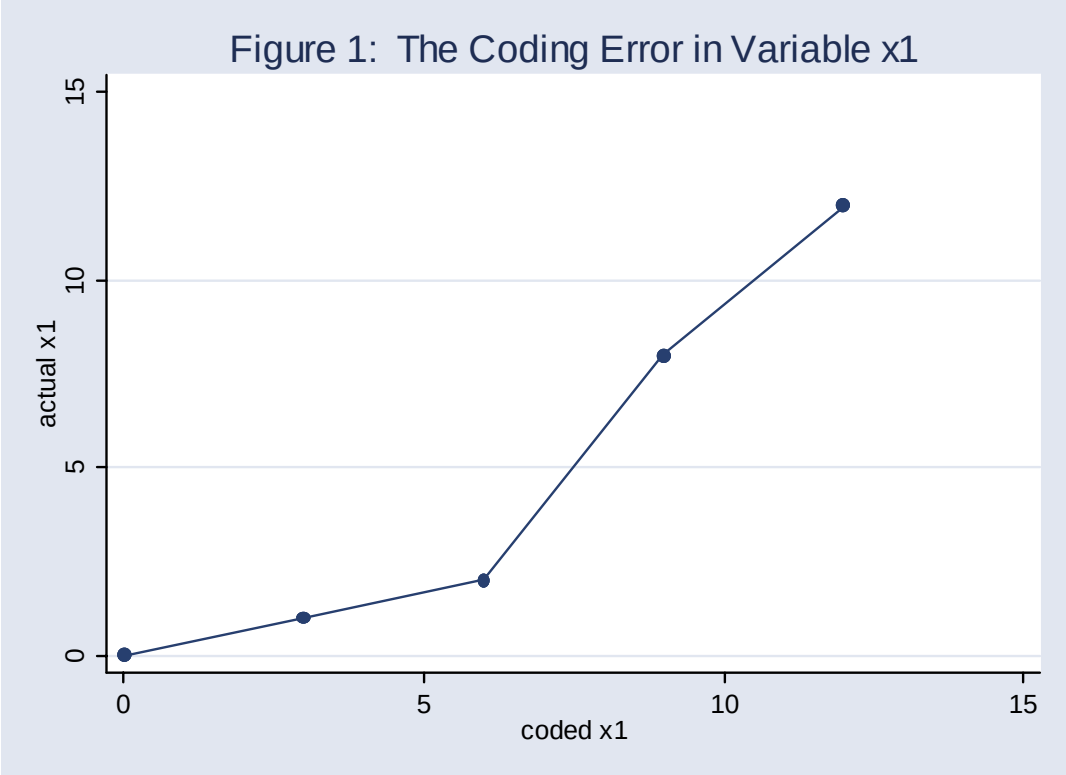
Figure 1: The Coding Error in Variable x1
actual x1
0
5
10
15
coded x1
0
5
10
15

Figure 1:

due to coding error at all. Suppose instead that the equally-space variable $\hat{x}_1$ is the true explanatory variable, but that its effect on y is slightly nonlinear. In this case, y would have been generated in this way:

$$y = f(\hat{x}_1) \ + 0.1x_2 \tag{4}$$

where the function $f : \hat{x}_1 \rightarrow x_1$ is given in Figure 1. On this interpretation, then there is nothing wrong with the coding of the variables. The problem instead is that the true relationship is very slightly nonlinear. These two interpretations have precisely the same statistical consequences, and so the discussion below applies to them both. That is, we are considering both the case of linear relationships with unknown, slightly nonlinear coding schemes for the independent variables, and we are also considering the case of perfectly coded variables in unknown slightly nonlinear functional forms.

In summary, the true variable x_1 is unknown and unavailable. What the researcher actually has in hand are the two explanatory variables $\hat{x}_1$ and x_2. Both have nice monotonic relationships to the dependent variable y, as Table 1 confirms. Moreover, each has a conditional strictly monotonic relationship to the dependent variable, which is the case relevant to the two pseudo–theorems under consideration.

The bivariate regressions (with standard errors in parentheses) show the expected positive relationships, too, with R^2 of .89 and .53 respectively, and all slopes highly statistically significant with t-ratios above 4:

$$\begin{aligned} y &= -1.524 + 1.047\hat{x}_1 \\ &\quad (.7069) \ (.0962) \end{aligned} \tag{5}$$

$$\begin{aligned} y &= -3.680 + 5.341x_2 \\ &\quad (2.233) \ \ (1.305) \end{aligned} \tag{6}$$

Thus nothing seems amiss except a bit of noise—actually, quite a bit less noise than in many international politics datasets! Everything is conditionally (positively) monotonic, 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, we ought to get two positive, statistically significant coefficients when the dependent variable is regressed on $\hat{x}_1$ and x_2. Indeed, when this regression is carried out, statistical significance does hold comfortably for both slopes, and the adjusted R^2 rises to .92,. Alas, however, 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:

$$\begin{aligned} y &= 0.5888 + 1.427\hat{x}_1 - 2.780x_2 \\ &\quad (1.034)\ (.1722)\quad (1.111) \end{aligned} \tag{7}$$

What is particularly odd here is that the messed–up coefficient is the one we measured perfectly.

Nothing in this finding depends on doing one regression or having a small sample. If one prefers the 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 $\hat{x}_1$, 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. 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.

4 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 dataset. 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) has made similar arguments for the statistical study of crisis bargaining behavior, and Sartori (2002) has proposed an entirely new statistical estimator for data subject to selection bias, based on her formal model of crisis bargaining. 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 required. And when no formal model is available, the most common situation, then very careful inspection of the data is needed.

Consider, for example, the small empirical problem already analyzed in this paper. Figures 2 and 3 show the simple bivariate plots of y against $\hat{x}_1$ and x_2. Both cases show evidence of slight nonlinearity, the usual sign

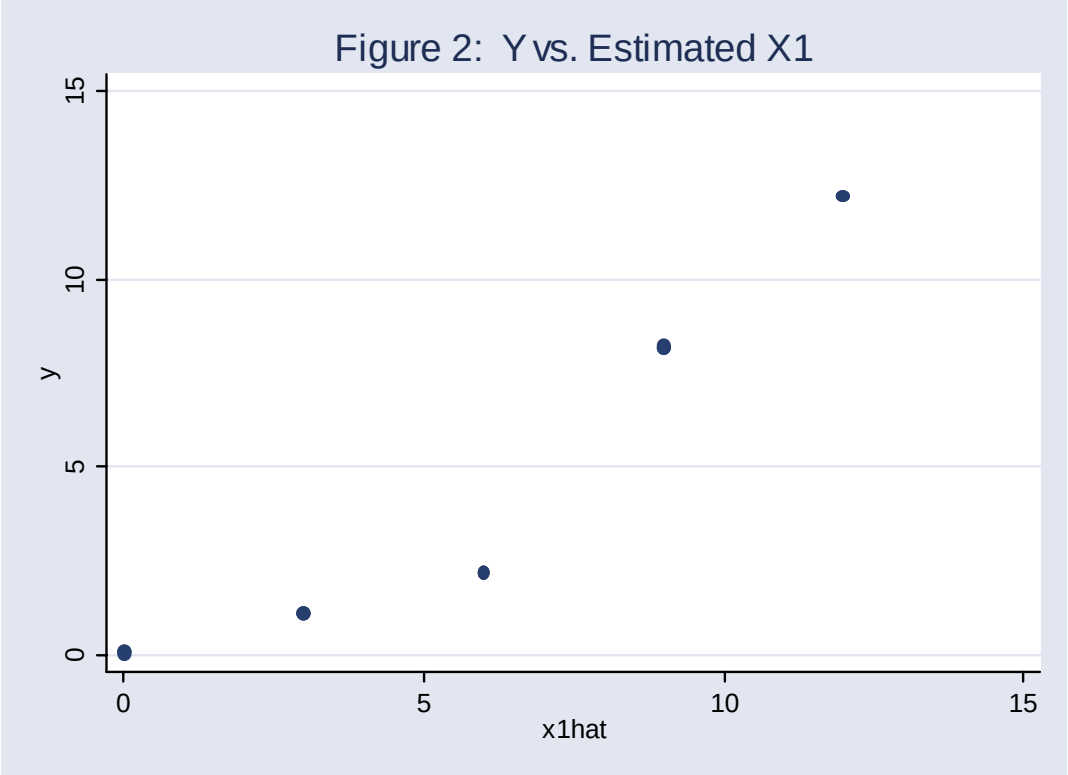

Figure 2:

that carelessly dropping variables into a canned regression programs would be dangerous.

Figuring out what is wrong here without knowing the true x_1 would be no trivial matter. However, since so much of what we know empirically, in international politics and elsewhere in political science, derives from crosstabulations, a sensible first step might be to tabulate mean values of y by the two independent variables. Table 2 results.

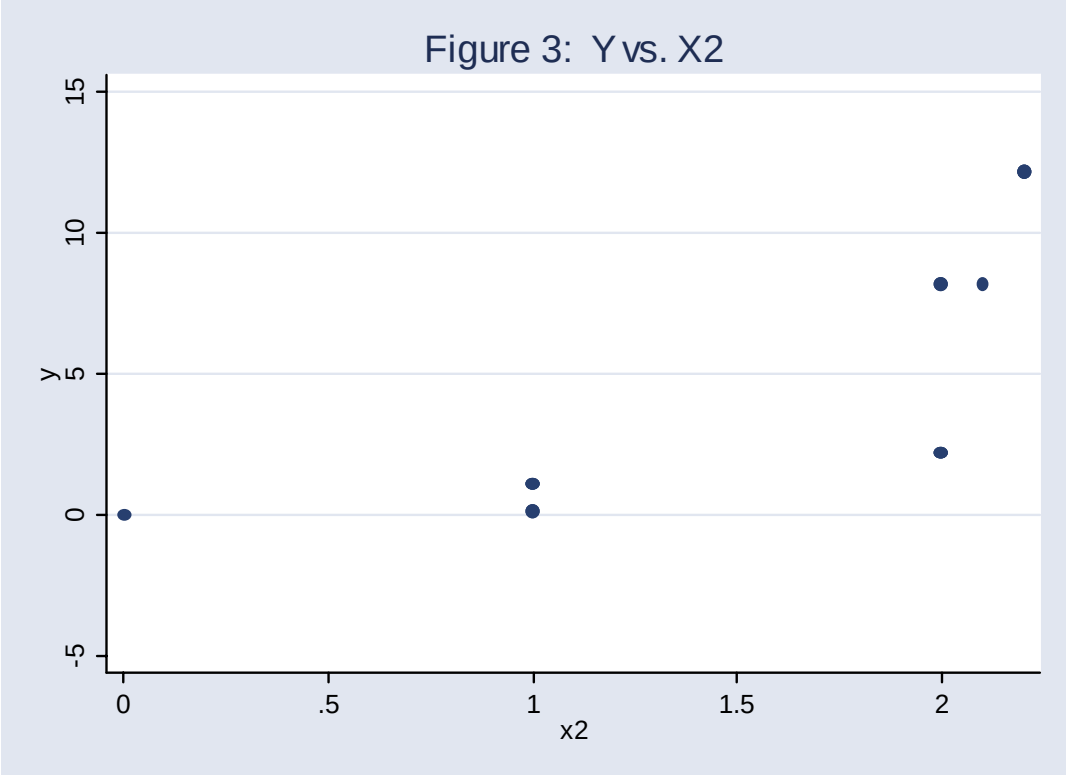

Figure 3:

Table 2: Mean Value of y

		x_2				
		0	1	2	2.1	2.2
	0	0	0.1			
$\hat{x}_1$	3		1.1			
	6			2.2		
	9			8.2	8.21	
	12					12.22

Patient inspection of these data shows that changes in x_2 cause the same 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 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 are additive but that x_1 is mis-coded (or equivalently, that its effect is nonlinear). We would be able to infer the correct coding (namely the true x_1) from the dummy coefficients. Finally, we could regress y on x_1 and x_2. That regression would fit perfectly, all the variables would have the correct coefficients, and we would be done.

The point I wish to make here is that getting right this simple little equation with its one small nonlinearity is perfectly possible. But it is no trivial matter of ten minutes' effort. Without knowing the exact problem in advance, it would take real time and patient effort to find it. Without that effort, the statistical results from the data are not just useless, but actively

[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.

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

Now it goes without saying that most of us use more than three explanatory variables, and that virtually all of 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 the premier quantitative international politics outlet, the *Journal of Conflict Resolution*, yields several articles with regressions and probits with 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 than are needed to obtain reliable results, and 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.

5 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. In the absence of careful supporting argument, they belong in the statistical rubbish bin.

What should the supporting argument for a statistical specification consist of? As I argued above, giving credibility to statistical specification, linear or otherwise, requires at least one of these 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. 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, nonetheless, integrating formal theory and statistical model puts to rest a host of uncertainties about the specification.

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 crosstabular analysis. 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 the 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.

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? If not, aren't my regression coefficients biased due to excluded variables?" But this argument is not as persuasive as it may seem initially.

First of all, if what you are doing is mis–specified already, then adding or excluding other variables has no tendency to make things consistently better or worse. 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 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, 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—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 regime. 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."

In their paper for this conference, ONeal and Russett (2004) 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.

The more creative testing and the fewer canned statistical outputs, the wiser we will be.

In sum, 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. These 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 are exactly where our best chances of progress lie.

References

[1] Achen, Christopher H. 2002. Toward a New Political Methodology: Microfoundations and ART. *Annual Review of Political Science* 5: 423-450.

[2] Bartels, Larry M. 2002. Beyond the Running Tally. *Political Behavior* 24, 2: 117-150.

[3] Kramer, Gerald H. 1986. Political Science as Science. In Herbert F. Weisberg, ed., *Political Science: The Science of Politics*. Washington, D.C.: American Political Science Association. Pp. 11-23.

[4] Oneal, John R., and Bruce Russett. 2004. Rule of Three, Let it Be? When More Really Is Better. Paper presented at the annual meeting of the Peace Science Society, Rice University, Houston, Texas, November 12-14, 2004.

[5] Ray, James Lee. 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.

[6] Ray, James Lee. 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, 2003.

[7] Sartori, Anne E. 2002. An Estimator for Some Binary Outcome Selection Models without Exclusion Restrictions. *Political Analysis* :ZZ

[8] Signorino, Curt S. 1999. Strategic Interaction and the Statistical Analysis of International Conflict. *American Political Science Review* 93: 279-297.