Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models

Michael J. Hanmer University of Maryland
Kerem Ozan Kalkan Middle East Technical University

Models designed for limited dependent variables are increasingly common in political science. Researchers estimating such models often give little attention to the coefficient estimates and instead focus on marginal effects, predicted probabilities, predicted counts, etc. Since the models are nonlinear, the estimated effects are sensitive to how one generates the predictions. The most common approach involves estimating the effect for the "average case." But this approach creates a weaker connection between the results and the larger goals of the research enterprise and is thus less preferable than the observed-value approach. That is, rather than seeking to understand the effect for the average case, the goal is to obtain an estimate of the average effect in the population. In addition to the theoretical argument in favor of the observed-value approach, we illustrate via an empirical example and Monte Carlo simulations that the two approaches can produce substantively different results.

Models for binary dependent variables are now part of the standard toolkit for empirical political scientists.[1] Typically, in the second semester course on quantitative methods, after learning the shortcomings of the linear probability model, students learn the virtues of logit and probit models. Closely related models for dependent variables with ordered categories, multiple nominal categories, and counts follow shortly thereafter.

Although such models are increasingly used in practice, they are still fairly new in political science (for a brief history of these models, see McCulloch 2000). As recently as the mid-1970s, an article comparing ordinary least squares regression (OLS), probit, and discriminant analysis by Aldrich and Cnudde (1975) described a camp of "purists" in political science who argued that only "less powerful techniques, such as cross-tabulation or ordinal measures of association and correlation" (Aldrich and Cnudde 1975, 572) were appropriate for noninterval-level dependent variables. And it was not long ago that King (1998), King, Tomz, and Wittenberg (2000), and Long (1997) moved the discipline toward the production of results that are more interesting and intuitive than odds ratios and the coefficients themselves, representing a giant leap forward. Around that same time, Herron (1999) emphasized the importance of computing

Michael J. Hanmer is Associate Professor of Government and Politics, and Research Director at the Center for American Politics and Citizenship, University of Maryland, 3140 Tydings Hall, College Park, MD 20742 (mhanmer@umd.edu). Kerem Ozan Kalkan is Assistant Professor of Political Science and Public Administration, Middle East Technical University, Dumlupinar Bulvari No: 1, Cankaya 06800 Ankara Turkey (okalkan@metu.edu.tr).

Kalkan acknowledges the postdoctoral fellowship in quantitative methods he received from University of Oxford, Department of Politics and International Relations, between 2009 and 2010. For thoughtful and constructive comments on earlier versions of this work, we would like to thank Chris Achen, Mike Alvarez, Dave Armstrong, Brandon Bartels, Dan Biggers, Tom Carsey, Anne Cizmar, Dan Corstange, Sarah Croco, Jill Curry, Jim Curry, Ray Duch, Justin Esarey, Rob Franzese, Bill Greene, Kosuke Imai, Bill Jacoby, Karen Long Jusko, Gary King, Jon Ladd, Eric Lawrence, Geoff Layman, Scott Long, Irwin Morris, Irfan Nooruddin, Bill Reed, John Sides, Jeff Smith, Piotr Swistak, Ric Uslaner, Nick Valentino, and Chris Zorn. For research assistance, we thank Christina Heshmatpour and Ilya Kopysitsky. We also wish to thank the anonymous reviewers and the Editor for their close reading of the article, thoughtful suggestions, and guidance. All errors are our own.

[1]The data used in this study can be found at http://dvn.iq.harvard.edu/dvn/dv/ajps.

American Journal of Political Science, Vol. 57, No. 1, January 2013, Pp. 263–277

 DOI: 10.1111/j.1540-5907.2012.00602.x

measures of uncertainty for predicted probabilities from limited dependent variable models (aka qualitative response or categorical), demonstrated approaches for doing so, and recommended a better measure of fit than the percentage of cases correctly predicted. More recently still, Ai and Norton (2003) provided a valuable statement on the proper way to interpret interaction effects in limited dependent variable models (see also Berry, Esarey, and Demeritt 2010), an issue that has caused problems in the discipline even with OLS models (Brambor, Clark, and Golder 2006; Kam and Franzese 2007).

Notwithstanding these and other important advances, a basic issue remains unresolved. That is, even in the simplest limited dependent variable models with multiple independent variables, the literature has not provided clear guidance regarding the best way to calculate marginal effects or discrete differences, which are the necessary tools for substantive interpretation. The aim of this article is to provide such clarity by highlighting an approach that provides results that are more in line with the goals of theoretically driven empirical research.

Background

Empirical researchers using limited dependent variable models often give little attention to the coefficient estimates, which cannot be interpreted as straightforwardly as OLS coefficients, and instead focus on predictions based on these coefficients. The reason is clear—the marginal effects and predicted quantities (e.g., probabilities, counts) are the keys to understanding the relationships of interest in the population. For example, in a logit or probit model, either the marginal effect of a change in the independent variable of interest on the probability of success or the discrete difference in the probability of success due to a change in the independent variable of interest is more informative. Since these models are nonlinear and inherently interactive in all of the variables, the size of the effect of a change in the independent variable of interest depends on the values of the other independent variables. As a result, the estimated effects, and thus the conclusions one can draw regarding the substantive significance of the variables of interest, are sensitive to which values for the other variables are chosen. Following Achen (1982), Long (1997), and King, Tomz, and Wittenberg (2000), we take it as a given that understanding whether or not the relationship of interest is substantively significant, rather than just statistically significant, is the ultimate goal, as it is a necessary part of evaluating one's theory. As a result, the algorithm researchers follow to calculate predicted quantities out of limited dependent variable models requires serious attention. As students learn when introduced to limited dependent variable models, this is not an issue when using an OLS model without interaction terms; i.e., in OLS models the coefficient on the variable of interest represents the marginal effect, which is constant across its full range as well as that of all of the other independent variables.

In presenting predictions for a change in an independent variable of interest in limited dependent variable models, there are two general approaches for dealing with the other independent variables in the model. The first involves creating an example case by selecting a set of specific values for the other variables and calculating the relevant predicted probabilities or marginal effect for that case. With country-level data, this might be the specific values taken on by the primary case of interest, e.g., the United States. One might set up several example cases, but researchers usually just use one case. Typically, the values of each of the other independent variables are set to their respective sample means; we will refer to this approach as the "average case" approach. The second approach involves holding each of the other independent variables at the observed values for each case in the sample, calculating the relevant predicted probabilities or marginal effect for each case, and then averaging over all of the cases; we will refer to this approach as the "observed value" approach. In the language of the burgeoning literature on causal inference (see, e.g., Ho et al. 2007 and Imai, King, and Stuart 2008), under the usual assumptions (see, e.g., Morgan and Winship 2007), the observed-value approach is used to compute the quantities of interest, such as the average treatment effect (ATE).

Although there are prominent studies using the observed-value approach, such as Wolfinger and Rosenstone's (1980) *Who Votes?*, our reading of the literature led us to suspect the average-case approach to be the dominant approach in political science. A content analysis of the *American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics* from 2006 supports that expectation; 68% of the articles using a limited dependent variable employed the average-case approach, and just 1% used the observed-value approach! Unfortunately, for 15% of the relevant articles the authors did not state clearly how they generated the reported predicted probabilities/marginal effects. The remaining articles did not report any results beyond the coefficients (11%) or reported odds ratios (4%).

The prevalence of the average-case approach is perhaps influenced by the wide availability of easy-to-use software such as Tomz, Wittenberg, and King's (2001)

Clarify (see also King, Tomz, and Wittenberg 2000)[2] and Long and Freese's (2005) *SPost* that set the default to the average-case approach.[3] Though these scholars made important contributions, we argue that the observed-value approach is preferable to the more common average-case approach on theoretical grounds. The thrust of our argument begins by noting that our theories and data collection efforts are not designed around making inferences about the average case. Since the research process does not start with a special concern for the average case, we should not conclude our empirical investigations by discussing results just for that case. Instead, we should conclude with results that allow us to make inferences about the population we have theorized about and collected data to represent. Moreover, as we show through an empirical example, using one of the most prominent dependent variables in the discipline and Monte Carlo simulations, the effect sizes generated from the two approaches can be quite different. Additionally, our Monte Carlo simulations demonstrate that the results from the observed-value approach are more robust to model misspecification than the average-case approach. While we view the argument as rather simple and straightforward, textbook treatments offer little guidance, and political scientists have gravitated to an approach that does not fully connect the interpretation of the results back to the ultimate goal of their research. Given the pervasiveness of limited dependent variable models, it is essential that we develop a set of best practices for the interpretation of the results from these models.

This article is organized as follows. First, we provide a brief review of the important features of limited dependent variable models, using binary response models as an illustration. Though we focus on binary response models, the logic applies straightforwardly to models for ordered categories, multiple nominal categories, counts, and durations (see below). Next, we lay out our argument, focusing on connecting the results to the goals of the empirical research endeavor. We then illustrate the differences between the average-case and observed-value approaches with an empirical example using data from the American National Election Studies (ANES). After that we use Monte Carlo simulations to provide additional insights. Before concluding, we discuss extensions and show how one might report the results.

[2]In the Supporting Information (SI) Section C, we provide sample Stata code to calculate marginal effects and discrete differences using the observed-value approach via the simulation method. With both the average-case and observed-value approach, one can compute confidence intervals via the delta method, bootstrapping, or simulation (see Herron 1999; King, Tomz, and Wittenberg 2000).

[3]It is important to note that the evidence does not support a claim that computational or programming issues are the reason why these programs and others implemented the average-case approach. For example, Rosenstone and Wolfinger (1978) implemented the observed-value approach with the technology that existed over 30 years ago.

Brief Review

Nonlinearity is the key attribute that makes interpretation of limited dependent variable model results less straightforward than OLS estimates. A brief review of the basic setup used for a binary dependent variable will help illustrate this point and motivate our argument. We begin with an unobserved continuous latent variable, y^*, such that $y^* = x\beta + \varepsilon$, where x is a matrix of independent variables, β a column vector of coefficients, and ε the errors.[4] Although we do not observe y^*, we do observe a realization in the form of a binary dependent variable, y, such that $y = 1$ if $y^* \geq \tau$ (where τ is a threshold usually set to 0) and $y = 0$ if $y^* < \tau$. The task is to estimate the probability of success, p, which can be written as:

$$p = \Pr(y = 1|x) = \Pr(y^* \geq 0|x) = \Pr(x\beta + \varepsilon \geq 0|x), \quad (1)$$

where Pr stands for probability, and τ is set to 0. Subtracting $x\beta$, we get:

$$p = \Pr(\varepsilon \geq -x\beta|x), \quad (2)$$

which, provided the distribution is symmetric, can then be expressed as:

$$p = 1 - F(-x\beta) = F(x\beta), \quad (3)$$

where F is the cumulative distribution function (cdf) of the errors. When dealing with binary dependent variables, F is usually assumed to follow either the logistic or normal distribution, both of which are S-shaped curves. For logit, which results from the logistic distribution, equation (3) can be written as: $p = 1 - \frac{1}{1+e^{x\beta}} = \frac{1}{1+e^{-x\beta}}$; and for probit, which results from the normal distribution, equation (3) can be written as: $p = 1 - \Phi(-x\beta) = \Phi(x\beta)$.

Although there are a variety of ways to describe the effect of the independent variables, we focus on the calculation of marginal effects and discrete differences. The issues discussed here do not arise if one interprets the results as odds ratios. However, we find odds ratios to be less informative because on their own they do not reveal anything about the initial probability of success, a necessary element for determining substantive significance.

[4]In the interest of simplicity, we subscript for each observation, i, only when calling attention to predictions that set each value of x to each observation's actual value observed in the sample.

To find the marginal effect of a continuous variable, x_k, we take the derivative of equation (3) with respect to x_k to get:

$$\frac{\partial p}{\partial(x_k)} = f(x\beta)\beta_k, \tag{4}$$

where f represents the probability density function (pdf). This result differs in obvious ways from the more simple result from OLS, where the effect, β_k, is constant regardless of the values of x_k or any other x. Equation (4) reveals that in addition to the value of the coefficient on the variable of interest (β_k), the effect will depend on the distribution (F) that is chosen, the values of all of the other coefficients (β), and *the values of all of the independent variables* (x).[5] In other words, due to the inherent nonlinearity, the effect is not constant—instead it depends on where one evaluates the value of the curve and which curve one uses. We discuss the implications of this after a review of the discrete difference method (aka first difference method), which is the appropriate way to calculate the effects of binary independent variables and can also be used for continuous independent variables.

With the discrete difference method, to calculate the effect of a change in x_k for two different values of x_k, say, from $x_k = \mathrm{d}$ to $x_k = \mathrm{c}$, we can simply estimate the probability of success when x_k is at each of these values and then compute the difference. For example, if x_k is a binary independent variable, one could set d to 0 and c to 1 (or vice versa). This can be expressed as:

$$\Pr(y = 1|x_{\sim k}, x_k = \mathrm{c}) - \Pr(y = 1|x_{\sim k}, x_k = \mathrm{d}), \tag{5}$$

where $x_{\sim k}$ represents all of the independent variables except x_k. Going back to the general expression for the probability of success, we can rewrite this in terms of where we evaluate the cumulative distribution function:

$$F(x_{\sim k}, x_k = \mathrm{c}; \beta) - F(x_{\sim k}, x_k = \mathrm{d}; \beta). \tag{6}$$

Again, the effect depends on the values of all of the coefficients, the values of all of the independent variables, and the choice of distribution. For our purposes, when calculating the effect of x_k, we are most concerned with the consequences associated with the choice of values to which the other independent variables should be set.

As noted above, the two general choices are to create an example case by setting the other independent variables to some set of values (usually the means of each independent variable are used), or to set the other independent variables to the values observed for each observation and then take the average. It is easy to see that when f is a nonlinear function, evaluating the marginal effects at the mean of x is not equivalent to calculating the marginal effect for each observation, using the values actually taken on by each observation, and then computing the average (i.e., the average marginal effect). Taking equation (4) and applying each approach, we see that the average-case approach (left-hand side) is not equivalent to the observed-value approach (right-hand side):

$$f(\bar{x}\beta)\beta_k \neq \frac{1}{n}\sum_{i=1}^{n} f(x_i\beta)\beta_k, \tag{7a}$$

where n is the sample size.

The same logic applies to the calculation of discrete differences. Taking equation (5) and applying each approach, we again see that for nonlinear functions the average-case approach is not equivalent to the observed-value approach:

$$\begin{aligned}&\Pr(y = 1|\bar{x}_{\sim k}, x_k = c) - \Pr(y = 1|\bar{x}_{\sim k}, x_k = d)\\ &\quad \neq \frac{1}{n}\sum_{i=1}^{n}[\Pr(y_i = 1|x_{i\sim k}, x_{ik} = c)\\ &\qquad - \Pr(y_i = 1|x_{i\sim k}, x_{ik} = d)],\end{aligned} \tag{7b1}$$

where $\bar{x}_{\sim k}$ is the mean for each of the x's that are not x_k. Alternatively, using equation (6), we can write this as:

$$\begin{aligned}&F(\bar{x}_{\sim k}, x_k = c; \beta) - F(\bar{x}_{\sim k}, x_k = d; \beta)\\ &\quad \neq \frac{1}{n}\sum_{i=1}^{n}[F(x_{i\sim k}, x_k = c; \beta) - F(x_{i\sim k}, x_k = d; \beta)].\end{aligned} \tag{7b2}$$

If the functions used in (7a) or (7b2) were strictly concave or strictly convex, then we could simply apply Jensen's inequality to determine whether the effect for the average case was larger than or smaller than the average effect calculated using the observed-value approach (we return to this in the Discussion section). Since the probit and logit distributions have portions that are concave and portions that are convex, we use Taylor series expansions to approximate the range over which the effect for the average case will be larger than or smaller than the average effect calculated via the observed-value approach. The results presented in Supporting Information (SI) Section A suggest that over most of the range of predicted probabilities, the effects for the average case will be substantively larger than the average effect calculated via the observed-value approach.[6] We explore the implications of this result in subsequent sections.

[5]For logit, equation (4) can be written as: $\frac{\partial p}{\partial(x_k)} = \frac{e^{x\beta}}{(1+e^{x\beta})^2}\beta_k$, and for probit: $\frac{\partial p}{\partial(x_k)} = \phi(x\,\beta)\beta_k$.

[6]Though Greene (2003) suggested that the two approaches would be equivalent asymptotically, he revised this position in Greene (2008). We discuss this further in Supporting Information Section A (see footnote 15) and provide Monte Carlo simulation results in Section D.

Having reviewed the basic implications of the nonlinearity in the models and shown that the average-case and observed-value approaches are not mathematically equivalent, we ask: at what values should the other independent variables be set?

Argument: Back to Basics

The above question seems like it should have an easy answer; however, someone looking to textbook treatments for an answer is likely to be left confused. Although textbooks cover the fact that the effects in nonlinear models are not constant and depend in part on the values for the other independent variables, advice on which values to choose is often lacking, incomplete, or inconsistent.

Hanushek and Jackson (1977) and Aldrich and Nelson (1984) stand out as important introductions to models for limited dependent variables in political science, especially in light of the earlier discussion of the "purist" camp. Both works contributed by emphasizing the distinctions between linear and nonlinear models and illustrating the distinctions through an empirical example showing the predictions for a range of values that the other independent variables might take on. Aldrich and Nelson (1984) go on to suggest that the effects might be summarized by picking "various interesting values" for the other independent variables. Given how little was known about these models at the time, by demonstrating how one might efficiently examine the results, these early efforts represent considerable advances. Unfortunately, they stop short of discussing alternative approaches and do not take up the issue of how one might determine which values are interesting.

In his seminal work, King ([1989] 1998) also discusses the importance of where one evaluates the curve and is careful to argue for studying the effect of the predictions at various combinations of the independent variables. He then suggests: "Probably the most obvious value to use is the most typical, the sample mean of each variable" (King [1989] 1998, 108). But here one is left wondering why the sample mean of each variable is the most obvious and how probable it is that the sample mean across all of the variables represents the most typical case. That is, although the sample mean of a single variable is an obvious representation of that variable, it is not clear that the case created by taking the mean of each independent variable represents a meaningful case, or a case that actually exists, let alone the most typical case.

As is the case with King ([1989] 1998), Long (1997) has become a vital resource for empirical political scientists. While one of the virtues of Long's text is the discussion of a variety of approaches to calculating meaningful outcomes, the text does not offer clear advice on how best to calculate marginal effects and predicted probabilities.

In his treatment of marginal effects, for instance, Long notes that while it is a popular approach, setting the other independent variables to their means (the average-case approach) is potentially problematic. The challenge leveled against the average-case approach is an appropriate one—the average case might not exist in the population. Long mentions that the average-case and observed-value approaches might yield different results, but he does not take a firm view on which is best, saying only that the observed-value approach "might be preferred" (1997, 74). Since Long's examples rely almost exclusively on the average-case approach and his *SPost* software (Long and Freese 2005) uses the average values as the default and does not provide estimates using the observed-value approach, it seems reasonable to conclude that his preference is for the average-case approach.

At the heart of the problem with textbook coverage is that the research question and data-collection process are not completely connected to the implications of estimating a model that is inherently, fully interactive in all of the variables. We show that the linkages between the various stages of the research enterprise are the strongest when researchers use the observed-value approach.

Why Use the Observed-Value Approach?

Evaluating one's theoretical expectations regarding the effect of changes in the independent variable(s) of interest on the dependent variable is the primary goal of observational studies.[7] After defining the population of interest, using standard probability sampling techniques, a sample is then drawn so as to be representative of that population. The population of interest obviously varies based on the nature of the research question and reflecting the theory that is specified, but it is usually defined broadly (e.g., those eligible to vote in the United States, the set of established democracies, citizens in European Union nations, and so on). After estimating the statistical model, one then seeks to make inferences from the sample to the population of interest.

Political science theorizing has not developed, nor has it sought to develop, theories about fine-grained categories created from whatever combination happens to result from taking the sample mean across all of the independent variables. Rather than seeking to understand

[7] The logic applies straightforwardly to experimental research that uses nonlinear models.

the effect for the average case, the goal is to obtain an estimate of the average effect in the population. We do not know of any studies that begin by stating what the average case is in their sample, explaining its wider significance, and then testing hypotheses about this case. Consider a typical scenario that arises with survey data from the United States: we are not aware of any theories that are specifically concerned with 48-year-old white women who are independent politically and have an income of $40,000–$45,000. Or, in a typical scenario with aggregate data, we are not aware of theories about countries with a 9% unemployment rate and a population of 18 million people, 17% of whom are over the age of 65, etc. With a population so narrowly defined, descriptive statistics would likely suffice, thus avoiding the issue of how best to obtain results from a limited dependent variable model.

As noted above, with OLS the effect of the independent variable of interest is constant for all (assuming a simple model without interaction terms), so both approaches produce identical outcomes. As a result, concerns with focusing on a narrowly defined subgroup, about which a specific theory has not been formulated, do not arise. However, the fully interactive nature of limited dependent variables necessitates greater care when interpreting the results. Picking a single case as the basis against which the effects of the independent variable(s) of interest will be examined can be problematic. That is, in assigning specific values for just a single case, one risks not only obtaining a case that is not in the population, rare, illogical, or not especially interesting, but also that the effects for this particular case are not representative of the overall effect or the effect for other classifications one might reasonably select. In other words, the interpretation of effects for just the average case provides a less thorough test of the theory and an inefficient use of the data, thus potentially limiting the ability to make inferences about the population of interest.

The observed-value approach, however, provides estimates that speak directly to the quantities of interest in the population. For example, if one is interested in the effect of electoral reform on turnout in the American electorate, the observed-value approach provides the average effect in the sample that one can then infer to the population—the American electorate. By contrast, the average-case approach considers the effect for a single case, e.g., 48-year-old white women with incomes of $40,000–$45,000, etc., and thus only represents the effect for that particular case within the larger population. Due to the fully interactive nature of the models, where the observations are placed on the curve will determine the size of the effect. As such, the average effect across the sample and the effects for other categorizations of the observations could be smaller (including no effect), the same, or larger than the effect for the average case (see SI Section A). Thus, unless the hypothesis relates to the small subset that makes up the average case, an investigation of just the average case is insufficient to evaluate fully the hypothesis of interest regarding the substantive and statistical significance of the effect. That is, the effect for the average case might not be generalizable to the wider population, especially if the average case does not exist in the population or is rare. Though researchers do not usually specify in the statement of their hypothesis the threshold beyond which an effect is substantively significant (though one could argue we should),[8] implicit in the statement of any hypothesis is that the effect is substantively important for the population of interest. Ultimately, the research community will evaluate the contribution of the work in large part based on how substantively important the effects are. Additionally, if one has taken the trouble to obtain a representative sample of a broad and meaningful population, describing a single case whose meaning may not be theoretically motivated is an inefficient way to use the data.

With large n studies, such as those based on survey data, the point seems rather clear as any individual case is anonymous and does not carry any special meaning. With studies based on a legislative body, a set of countries, and so on, the individual cases are entities in which we might have a special interest. For example, we might want to predict what could have influenced the probability that Hillary Clinton (in her time as a senator) would have supported an immigration bill, the number of environmental regulations Ireland will enforce, which countries are likely to fall into a civil war, or how historical events might have turned out differently. Signorino and Tarar (2006) provide a nice example of the last scenario as they analyze how the immediate and short-term balance of forces might have altered the 1948 Berlin Blockade and the 1937–38 Soviet-Japanese conflict over Manchukuo. For each of these examples, calculations focused on these specific observations are part of the observed-value approach, as a prediction is first obtained for each observation and then the average is taken to get the average effect across the sample. So, if one centers the interpretation upon just a single known case, we contend there must be theoretical justification for doing so and recommend that authors state this clearly.

While there might be good reason to study a particular legislator, country, etc., in depth, as discussed above, it seems unlikely that a theoretical justification exists for

[8]See Achen (1982, 44–51) on the importance of substantive significance, including tests of substantive significance.

studying the average case. For example, examining the effect of an increase in the Hispanic population on the probability of Senator Clinton's support for an immigration bill would be of greater theoretical interest than the effect for the average senator, who might not exist in the population or might never reasonably be thought to enter the population. And even if the average senator does exist in the population, he or she might be exceptionally rare, unexciting, or not electable in a place that might experience the change of interest. Theoretical reasons might exist for focusing on a single case in situations such as this, but we think political scientists should not stop there. Consider again the Clinton example. While there might be reasons to focus on Senator Clinton, doing so ignores the set of questions having to do with broader concerns about public policy outcomes and/or the way the Senate operates more generally. As such, any detailed investigation of a single case (or set of cases) should be accompanied by the overall effect obtained via the observed-value approach.

A concern with the average-case approach worth further discussion involves situations in which naive use leads to establishing rare or meaningless cases as the baseline from which one attempts to generalize. Surprisingly, nearly 20% of the articles using the average-case approach in our content analysis admitted to setting all of the other independent variables to their means even when those variables included dummy variables and/or squared terms. Naively setting all other independent variables to their means will set dummy variables to their sample proportions. However, assigning a "central value" (e.g., the mean) to binary or bimodal variables can make "the concept of a 'central value' . . . less meaningful" (Gelman and Hill 2007, 467). A dummy variable for blacks (vs. other race) might be set to 0.12, which, literally translated, suggests calculating a prediction for someone who is 12% black. While there are people who fit some descriptions of this nature, our surveys generally do not search for such detailed information; if they did, the type of sample drawn and the coding of the variables would necessarily have to change. Consider also a dummy variable indicating whether one has at least a college degree or not. Here, setting the value to the sample proportion simply does not make sense as calculating a prediction with this variable set to any value other than 0 or 1 does not make sense—one either has the degree or does not. So, while Wooldridge suggests choosing a value for dummy variables "is really based on taste" (2002, 466), we disagree.[9]

[9]Of course, the problem of odd or nonsensical cases can be resolved by selecting the modal value for dummy variables, though the effects for each group might differ substantially.

A similar problem might arise with more complex specifications that include squared terms or interaction terms. Take squared terms, for instance. In the 2004 ANES, the average of the age variable is 47, but the average of age squared is 2,528, the square root of which is 50. So, in a model that includes age and age squared, examining the effects for all of the independent variables at their sample mean produces a nonsensical result, as the average case cannot be simultaneously 47 and 50 years old. Although greater attention to the nature of the data provides an easy fix, our content analysis revealed that many were not particularly attentive to this detail.

In sum, the virtue of using the observed-value approach is simple: it better serves the goal of theory-driven empirical research—making inferences about the population of interest from the sample. As Hanushek and Jackson elegantly stated: "Let us reemphasize that meaningful empirical work must be based upon explicit hypotheses and statements about predicted behavior" (1977, 3). Estimates from the observed-value approach connect directly to the original hypotheses and more easily allow researchers to evaluate the substantive implications of their theory.

But it is crucial to note that the observed-value approach is not foolproof. Though our focus is on setting the variables not being manipulated, setting the values of the variable being manipulated is at least as important. That is, researchers still have to take care in defining their counterfactuals so that they are realistic for the population of interest (Gelman and Pardoe 2007; King and Zeng 2006). Commonly, researchers set the variable of interest to its minimum and maximum values and compute the difference in predicted probabilities from this change. However, this may be problematic when there are very few cases at the minimum and maximum values or when such changes are not likely to actually occur at all, or for some subgroups. For example, moving from strong Democrat to strong Republican on party identification is a very rare change. In circumstances like this, researchers should examine a more modest change in the independent variable of interest, report only the average effect using the observed-value approach for the subpopulations for which the changes are reasonable, or as Gelman and Pardoe (2007) offer, weight the predicted probabilities with respect to the closeness of counterfactual scenarios in the data. Researchers might even employ these strategies in combination with one another.

An Empirical Example

We now demonstrate that the two approaches to estimating predicted probabilities and marginal effects can produce substantively different results.[10] Using the 2004 ANES, we estimate a probit model of the vote choice between George W. Bush (coded as 1) and John Kerry (coded as 0).[11] The model includes a standard set of demographic and attitudinal variables (see SI Section B for coding and model results). After estimating the model, we determined that the average case (setting dummy variables to their mode and rounding to the nearest whole number for categorical variables) is a white 48-year-old female who identifies as an independent, has an associates degree, is politically moderate, believes economic performance has been the same, disapproves of the Iraq war but not strongly, and has income between $45,000 and $50,000. There is nothing that stands out as odd about this combination of characteristics, and we imagine such people do exist in the population. Critically, however, such a person does *not* exist in the 2004 ANES, and we are unaware of any theories specific to this particular type of person.

We begin with an investigation of the baseline-predicted probability of voting for Bush calculated using the average-case and observed-value approaches. While the predicted probability of voting for Bush for the average case is 60%, the average probability of voting for Bush in the sample using the observed-value approach is 51%. Since the result from the observed-value approach can also be thought of as an estimate of the aggregate proportion of votes going to Bush, an appropriate way to evaluate this result is to compare it to the official proportion of the popular vote that went to Bush in the 2004 election. The official proportion of the popular vote going to Bush was also 51%. Thus, this simple analysis indicates that the probability of voting for Bush for the average case does not provide a good representation of the probability of voting for Bush among American voters.

Next, we examine the predicted probability of voting for Bush across the range of values taken on by the independent variables. That is, for each independent variable, we examine the predicted probability at each value (or a set of values across the range of the age and income variables) and set all of the other variables to either their sample average/mode or observed values. For selected independent variables, Figures 1a–1d show that the probability of voting for Bush at some values of the independent variable of interest might differ substantially depending on which approach is chosen, with a gap sometimes in excess of 20 percentage points (for the rest of the results, see SI Section B Table 2). For example, someone who believes the economy is much better but otherwise has the average values for all of the other variables has an 87% chance of voting for Bush. By contrast, when all of the other variables are set to their observed values, the probability of voting for Bush among those who believe the economy is much better is 64%, 23 points lower. While the differences are sometimes small, for the variables that have the strongest influence on vote choice, differences of 10 percentage points or more are common.

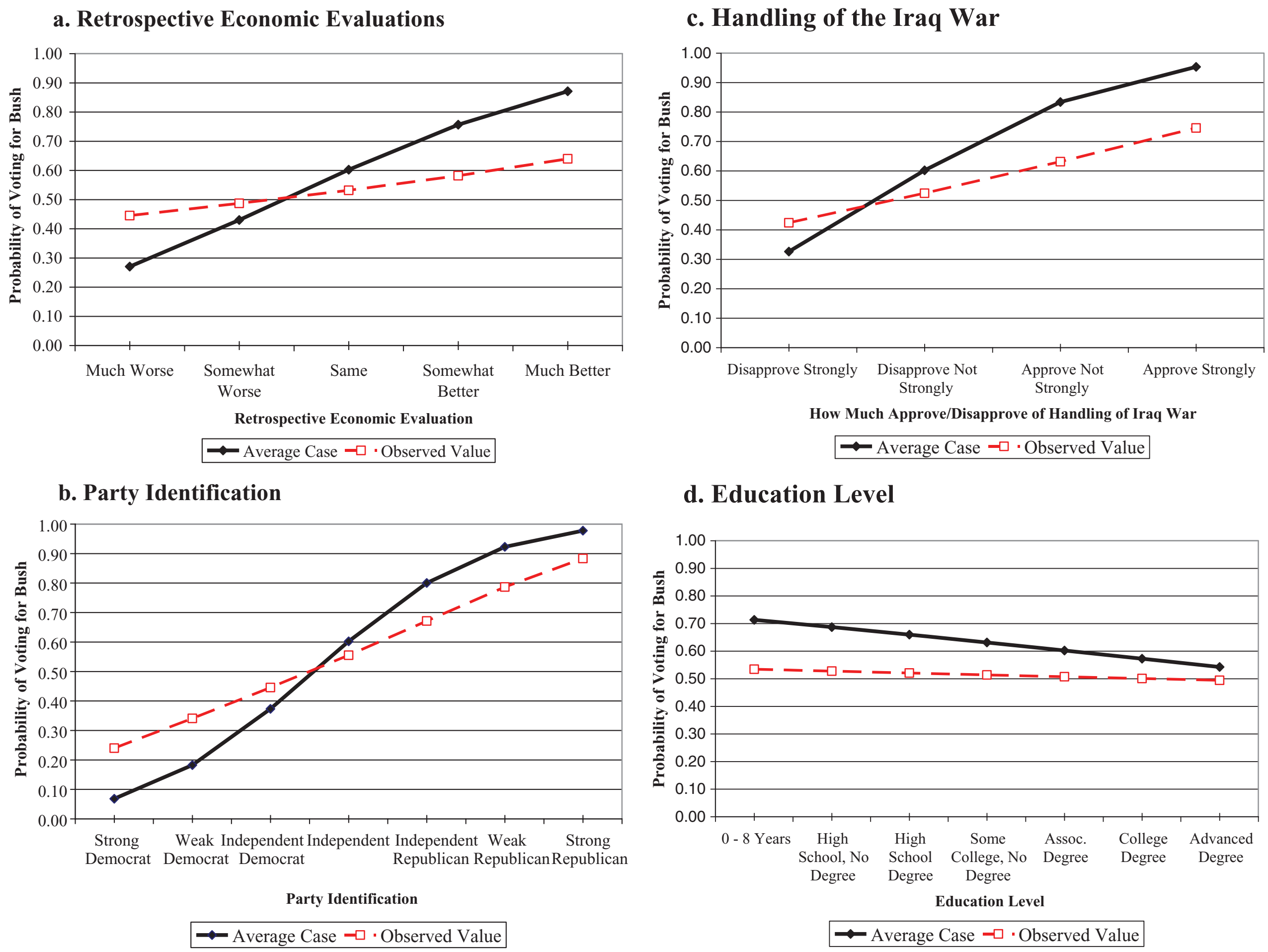
We conclude this section by comparing the effect of changes in the independent variables on the predicted probability of voting for Bush generated under the two approaches. In so doing, we use the predicted probabilities displayed in Figures 1a–1d and SI Section B Table 2. For each of the independent variables, we began by establishing the sample mean of that variable as the baseline against which changes would be made. Next, we calculated a predicted probability for each of the other values of that independent variable (or for age and income a set of values) and calculated the absolute value of the difference between those predictions and the baseline. For example, using the average-case approach, the predicted probability of voting for Bush at the average value of retrospective economic evaluations ("the same"), and the average (mode for dummy variables) of all of the other independent variables is calculated as 60%, and the predicted probability of voting for Bush when changing just the value of retrospective economic evaluations to "somewhat worse" is 43%, yielding an absolute difference of 17 percentage points. We repeated this process for the observed-value approach where the baseline was established as the predicted probability when the value for retrospective economic evaluations was set to "the same," and all other variables were set to their observed values; the effects were then calculated by taking the absolute difference between that prediction and the values when retrospective economic evaluations took on each of the other possible values, respectively.

For selected independent variables, Figures 2a–2d show the absolute differences in the sizes of the effects for the average-case and observed-value approaches. Each figure has two V-shaped lines that are both set to 0 for the baseline value. As would be expected from a close inspection of Figures 1a–1d, the effect for the average case is often substantially larger than the average effect calculated using the observed-value approach. For example, while the effect of moving from "the same" to "somewhat worse" on retrospective economic evaluations was 17 percentage points when using the average-case approach, the

[10]We find it most productive to illustrate the differences between the two approaches in a generic example; we leave it to others to examine the sensitivity of the conclusions to the method used to calculate the predictions in their respective substantive areas.

[11]We use the sample that was asked the traditional turnout question. The general conclusions do not change when we use those asked the experimental question or the pooled sample.

FIGURE 1 Predicted Probability of Voting for George W. Bush vs. John Kerry in 2004, Using the Average-Case and Observed-Value Approaches, for Selected Variables



Notes: Data are from the 2004 ANES, using respondents who first answered the standard turnout question. Results are based on estimates from the model reported in SI Section B Table 1.

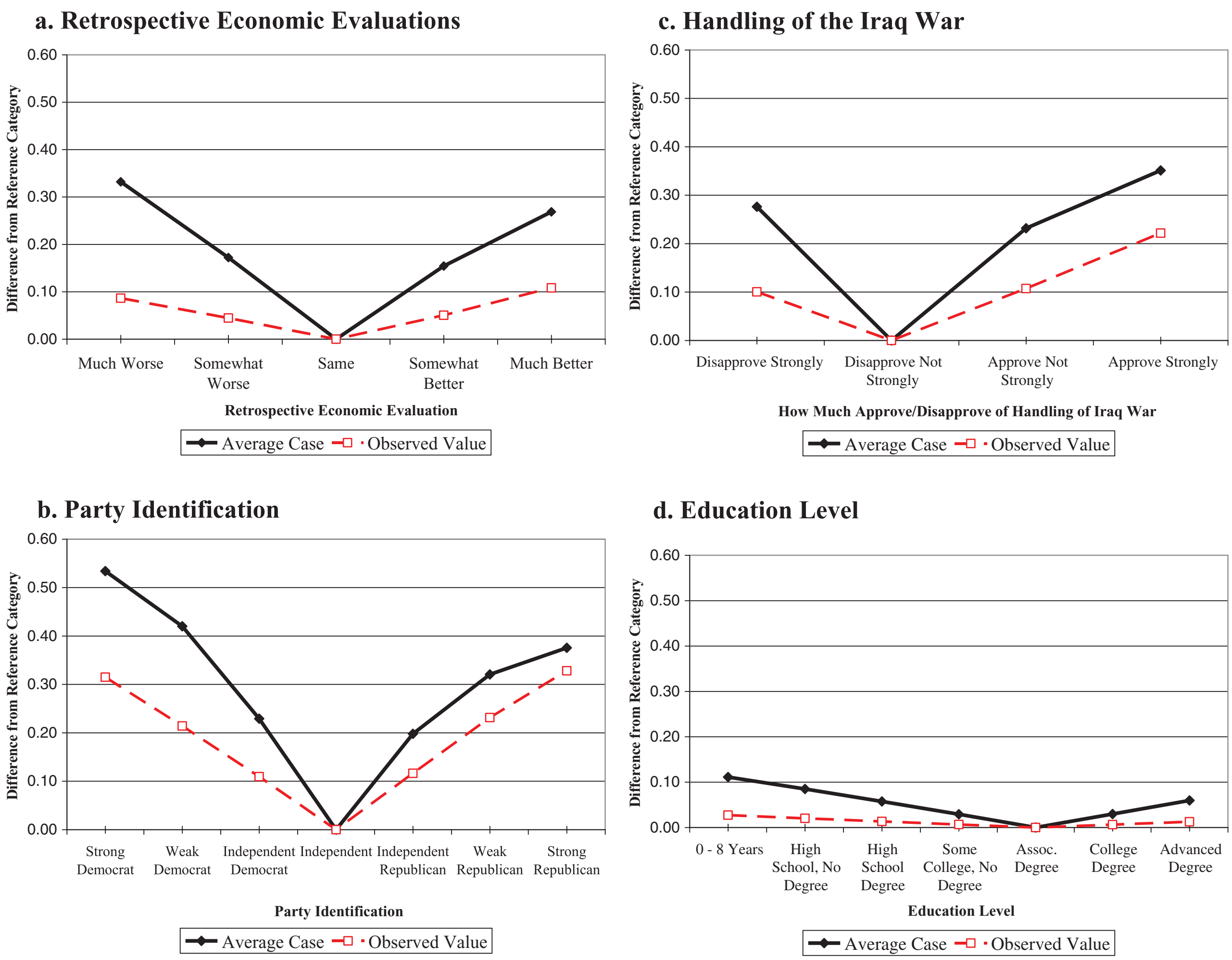
average effect of this movement in retrospective economic evaluations using the observed-value approach is just 4 percentage points, an effect four times smaller than the effect for the average case (see Figure 2a). In other words, in this example the effect for the average case does not represent the average effect of the independent variables on the vote choice of American voters.

Monte Carlo Simulations

Using Monte Carlo simulations, we demonstrate further the basic differences between the results from the average-case and observed-value approaches. In addition to different effect sizes, with these data the estimated predictions and effects are often more sensitive to model misspecification when using the average-case approach than when using the observed-value approach.[12] The intuition here is straightforward and derives from the important lesson provided by King and Zeng (2006; see their Appendix B for a proof)—the farther one moves from the support of the data, the more sensitive to model misspecification the results will be. By definition, with the observed-value approach, the values of all of the other independent variables will be on the support of the data; however, the mean of all of the other independent variables might not represent a case that is present or common in the population or sample.

[12]We thank Gary King for suggesting this line of inquiry.

FIGURE 2 Effect (Absolute Value) of a Discrete Change in the Independent Variable from Its Mean on the Probability of Voting for George W. Bush vs. John Kerry in 2004, Using the Average-Case and Observed-Value Approaches, for Selected Variables



Notes: Data are from the 2004 ANES, using respondents who first answered the standard turnout question. Results are based on estimates from the model reported in SI Section B Table 1.

The Monte Carlo simulations consisted of 1,000 trials using 1,000 observations generated with the following model:

$$y_i^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i,$$

where $y_i = 1$ if $y_i^* > 0, y_i = 0$ if $y_i^* \leq 0$, x_1 was drawn from a uniform distribution and recoded into a categorical variable that takes on one of three values (1, 2, or 3) with an equal likelihood, x_2 and x_3 were both drawn from the standard normal distribution, and ε_i was drawn from the standard normal distribution. We also varied the correlation between x_2 and x_3. We set the latent variable, y_i^*, as $y_i^* = 2 + -1x_1 + 1x_2 + .5x_3 + \varepsilon_i$. Based on the analysis presented in SI Section A, we expected this setup would show that the effects estimated using the average-case approach would be larger in absolute magnitude than the effects estimated via the observed-value approach.[13]

For both the average-case and observed-value approaches, we generated the following estimates: (1) the predicted probability of success for each value of x_1 and the effect of a one-unit increase in the value of x_1; (2) the marginal effect of x_2; and (3) the marginal effect of

[13]We also ran the simulations with β_0 set to 0.5. As expected based on the analysis presented in SI Section A, the effects estimated using the average-case approach were smaller in absolute magnitude than the effects estimated via the observed-value approach. The overall conclusions were substantively similar to those reported below. The results are available upon request.

x_3. For each of the values for the correlation between x_2 and x_3, we ran the following sets of 1,000 trials using the identical data: (1) a "true" probit model in which y was regressed on x_1, x_2, and x_3; (2) a probit model omitting x_1; (3) a probit model omitting x_2; and (4) a probit model omitting x_3. We also calculated the true values using both the average-case and observed-value approaches.

We report the results in Table 1. In the first two columns of Panel A, we report the marginal effects from the true model for x_2 and x_3 using the average-case approach (column 1) and the observed-value approach (column 2). The remaining columns show the average difference in the estimates of the marginal effects when the model is misspecified by dropping one of the independent variables. The first two columns of Panel B show the first differences for each one-unit change in the value of x_1 for the true model across the two approaches, while the remaining columns show the average difference in the estimates when the model is misspecified.

The first noteworthy result is that in the true model, as expected, the estimates from the average-case and observed-value approach differ. More specifically, the size of the effects estimated using the average-case approach exceeds the size of the effects estimated via the observed-value approach. For example, the marginal effects of x_2 and x_3, respectively, when calculated using the average-case approach, are nearly twice as large as when the calculations were done using the observed-value approach. Such discrepancies could be quite problematic in a variety of situations, especially in the context of public policy evaluation, where the costs of potential policy changes need to be carefully weighed against the benefits. While we maintain that adherence to the logic of the research process is sufficient reason to favor the observed-value approach, these results provide further evidence that the two approaches can yield different substantive understandings of the magnitude of the relationships between the independent and dependent variables.[14]

Turning to the issue of model dependence, the differences between the two approaches are again quite clear. When the correlation between x_2 and x_3 is 0, omitting from the probit model any one of the independent variables used in the data-generating process has virtually no effect on the predictions and marginal effects calculated via the observed-value approach; for example, when x_1 is dropped from the model, the marginal effects of both x_2 and x_3 differ from the estimates from the true model by less than 0.002. The same is not true for the estimates obtained with the average-case approach, though the discrepancies are not always large. The story changes somewhat when x_2 and x_3 are highly correlated; but the pattern of results is not a clear function of the correlation between x_2 and x_3. For example, the one time in which the average-case approach comes closer to the truth in the misspecified model (the marginal effect of x_3 when x_2 is dropped) is when the correlation is set to 0.5. But when the correlation is increased to 0.8, the observed-value approach again provides an estimate closer to the truth. Overall, the simulations are consistent with our expectation that the estimates from the observed-value approach tend to be more robust to model misspecification. However, due to the number of moving parts in nonlinear models, pinpointing all of the conditions under which the observed-value approach will be significantly less model dependent is difficult.

Discussion

Extensions

We focused our discussion thus far on binary response models due to their wide use and status as a building block for other nonlinear models. But the logic of our argument extends straightforwardly to all nonlinear models. Since S-shaped curves are common to ordered probit and logit and multinomial logit and probit models, the magnitude of the predicted effects for the average-case in comparison to the observed-value approach will depend on the same set of indicators that are relevant for logit or probit estimation (plus the cut-points for the ordered models). For count models (both Poisson and negative binomial regressions), since the formula for the expected count, $e^{x\beta}$, is strictly convex, by Jensen's inequality, the average effect calculated via the observed-value approach will always be substantively larger than the effect for the average-case approach. Here, we would like to add a note of caution. Although one might argue that when the results from the average-case approach are smaller they represent conservative estimates and are thus not problematic, this is not the case. First, as we have argued, it is best to use the average-case approach only when there is a theoretical justification for focusing on the effect for the average case, or some other specific case. Second, if the results from the average-case approach fall below the bar set for concluding there is a substantively significant effect, scholars and/or policy makers might incorrectly conclude that the effects are too small for further consideration, possibly hindering progress in both domains.

[14] At this point in the analysis, it is not the case that one approach performed better than the other in recovering the truth; rather, it is simply the case that there are two different sets of truths—one for the average case and one for the average effect calculated using the observed-value approach. That is, the simulations are consistent with the inequalities stated in equations (7a), (7b1), and (7b2).

TABLE 1 Monte Carlo Simulation Results Examining Model Dependence Using the Average-Case and Observed-Value Approaches for $y^* = 2 + -1x_1 + 1x_2 + 0.5x_3 + e$

Panel A: Marginal Effects of x_2 and x_3 for True Model and *Amount of Bias* Due to Model Misspecification

Correlation x_2 & $x_3 = 0$	Effects for True Model[1]		Bias in Model 1: Excludes x_1		Bias in Model 2: Excludes x_2		Bias in Model 3: Excludes x_3	
Variable of Interest	Average Case[2]	Observed Values[3]	Average Case[2]	Observed Values[3]	Average Case[2]	Observed Values[3]	Average Case[2]	Observed Values[3]
x_2	0.400	0.230	−0.097	0.001	–	–	−0.043	0.000
x_3	0.201	0.115	−0.049	0.000	−0.059	0.000	–	–
Correlation x_2 & $x_3 = .5$								
x_2	0.400	0.213	−0.095	0.001	–	–	0.059	0.054
x_3	0.201	0.107	−0.049	0.000	0.102	0.107	–	–
Correlation x_2 & $x_3 = .8$								
x_2	0.401	0.205	−0.095	0.001	–	–	0.136	0.082
x_3	0.200	0.102	−0.048	0.000	0.246	0.164	–	–

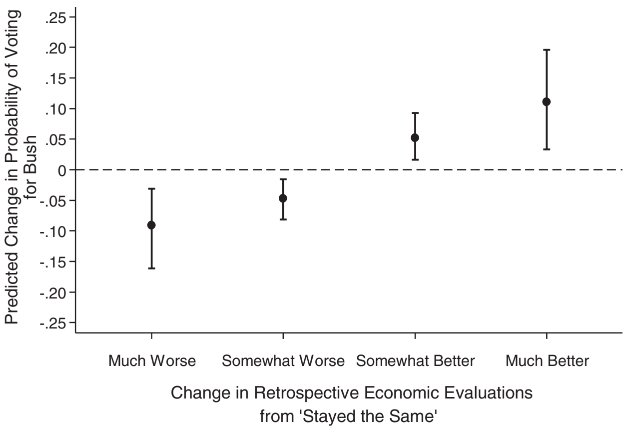
Panel B: Predicted Effects (First Differences) Across Values of x_1 (Going from $x_1 = 1$ to $x_1 = 2$ and $x_1 = 2$ to $x_1 = 3$) for True Model and *Amount of Bias* Due to Model Misspecification

Correlation x_2 & $x_3 = 0$	Effects for True Model[1]		Bias in Model 2: Excludes x_2		Bias in Model 3: Excludes x_3	
Change in Value of x_1	Average Case[2]	Observed Values[3]	Average Case[2]	Observed Values[3]	Average Case[2]	Observed Values[3]
1 to 2	−0.342	−0.248	0.081	0.000	0.027	0.000
2 to 3	−0.341	−0.247	0.081	0.000	0.027	0.000
Correlation x_2 & $x_3 = .5$						
1 to 2	−0.342	−0.227	0.066	0.000	0.021	0.000
2 to 3	−0.342	−0.227	0.066	0.000	0.021	0.000
Correlation x_2 & $x_3 = .8$						
1 to 2	−0.342	−0.217	0.037	0.000	0.011	0.000
2 to 3	−0.342	−0.217	0.037	0.000	0.011	0.000

Notes: Results for the misspecified models represent the average difference from the estimate of the truth.
[1] True model includes x_1, x_2, and x_3. The results represent the marginal effect (Panel A) or first difference (Panel B). For both approaches, using the true coefficients (and for the average-case approach, the true mean values) produces results that are nearly identical to those reported in the true model columns, respectively.
[2] Estimates computed by setting all other independent variables to their sample means.
[3] Estimates computed by setting all other independent variables to their observed values in the sample.

The logic of the observed-value approach also extends straightforwardly to more complex models. As we hinted at above, models with squared terms and/or interaction terms require some care but do not pose any problems. One can also implement this approach with designs that first preprocess the data via a matching algorithm and then run a nonlinear model (see, e.g., Herrnson, Hanmer, and Niemi, forthcoming), multilevel models (see, e.g.,

FIGURE 3 Predicted Effects (First Differences) of Changing Retrospective Economic Evaluations on the Probability of Voting for George W. Bush vs. John Kerry in 2004, Using the Observed-Value Approach, with 95% Confidence Intervals



Notes: Data are from the 2004 ANES, using respondents who first answered the standard turnout question. Results are from statistical simulation.

Biggers 2011), models with fixed effects (see, e.g., Hanmer 2009), and random effects panel data models (see, e.g., Bartels et al. 2011). Our point that those using the observed-value approach use their data more efficiently is especially relevant with data across time.

Reporting Results

Before concluding, we return to our empirical example to show how one might report the predicted effects from the observed-value approach. Since our earlier goal was to compare the two approaches, we did not report the results as we would if our main goal was to understand vote choice in the 2004 election. Figure 3 shows the predicted effects and the confidence intervals around the effects of changing retrospective economic evaluations on the probability of voting for George W. Bush, using the observed-value approach. As we did earlier, we calculated the first differences of a movement on retrospective economic evaluations from the "same" to each of the other answer choices (other changes in the variable of interest could be similarly displayed). The average effect is shown as a dot, and the 95% confidence intervals are represented by the lines extending from the predicted effect. We also present the predicted probability of voting for Bush across each of the independent variables, along with 95% confidence intervals, in SI Section B Table 3 and the first differences for selected variables, with 95% confidence intervals, in SI Section B Table 4.

Conclusion

Political science has come a long way since the "purist" camp suggested retreat to descriptive statistics in the face of a limited dependent variable. A variety of models to deal with limited dependent variables have become commonplace. But as King and Zeng have noted: "As recently as a half decade ago, most quantitative political scientists were still presenting results in tables of hard-to-decipher coefficients from logit, probit, event count, duration, and other analyses" (2006, 131). Yet, a lack of clarity on how best to estimate substantively meaningful results from these models represents a missing link in the literature. Unfortunately, many studies present results for cases that are of potentially limited theoretical interest, thus limiting the existing base of knowledge and ability to generalize.

The argument and results presented here call for a shift in the dominant practice of calculating predicted probabilities for limited dependent variable models. Due to both theoretical and methodological reasons, researchers using limited dependent variable models should report predicted probabilities and marginal effects estimated via the observed-value approach. Doing so is consistent with recognizing that the core of the scientific endeavor is to "infer beyond the immediate data to something broader that is not directly observed" (King, Keohane, and Verba 1994, 8). The observed-value approach does a better job associating the theoretical framework and data-collection activities with the properties of limited dependent variable models. Moreover, researchers can implement the observed-value approach straightforwardly in any statistical package (see SI Section C for an example using Stata; code for SPSS and R are available upon request). Thus, the shift to the observed-value approach will serve to improve future empirical research.

References

Achen, Christopher H. 1982. *Interpreting and Using Regression*. Thousand Oaks, CA: Sage.

Ai, Chunrong, and Edward C. Norton. 2003. "Interaction Terms in Logit and Probit Models." *Economics Letters* 80: 123–29.

Aldrich, John H., and Charles F. Cnudde. 1975. "Probing the Bounds of Conventional Wisdom: A Comparison of Regression, Probit, and Discriminant Analysis." *American Journal of Political Science* 19(3): 571–608.

Aldrich, John H., and Forrest D. Nelson. 1984. *Linear Probability, Logit, and Probit Models*. Newbury Park, CA: Sage.

Bartels, Brandon L., Janet M. Box-Steffensmeier, Corwin D. Smidt, and Renee M. Smith. 2011. "The Dynamic Properties of Individual-Level Party Identification in the United States." *Electoral Studies* 30(1): 210–22.

Berry, William D., Justin Esarey, and Jacqueline Demeritt. 2010. "Testing for Interaction in Binary Logit and Probit Models." *American Journal of Political Science* 54(1): 248–66.

Biggers, Daniel R. 2011. "When Ballot Issues Matter: Social Issue Ballot Measures and Their Impact on Turnout." *Political Behavior* 33(1): 3–25.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14(1): 63–82.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Gelman, Andrew, and Iain Pardoe. 2007. "Average Predictive Comparisons for Models with Nonlinearity, Interactions, and Variance Components." *Sociological Methodology* 37(1): 23–51.

Greene, William H. 2003. *Econometric Analysis*. 5th ed. Upper Saddle River, NJ: Prentice Hall.

Greene, William H. 2008. *Econometric Analysis*. 6th ed. Upper Saddle River, NJ: Prentice Hall.

Hanmer, Michael J. 2009. *Discount Voting: Voter Registration Reforms and Their Effects*. New York: Cambridge University Press.

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

Herrnson, Paul S., Michael J. Hanmer, and Richard G. Niemi. 2012. "The Impact of Ballot Type on Voter Errors." *American Journal of Political Science*. Forthcoming.

Herron, Michael C. 1999. "Post-Estimation Uncertainty in Limited Dependent Variable Models." *Political Analysis* 8(1): 83–98.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3): 199–236.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. "Misunderstandings between Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(2): 481–502.

Kam, Cindy D., and Robert J. Franzese Jr. 2007. *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. Ann Arbor: University of Michigan Press.

King, Gary. 1998. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Ann Arbor: University of Michigan Press.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2): 341–55.

King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2): 131–59.

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

Long, J. Scott, and Jeremy Freese. 2005. *Regression Models for Categorical Outcomes Using Stata*. 2nd ed. College Station, TX: Stata Press.

McCulloch, Charles E. 2000. "Generalized Linear Models." *Journal of the American Statistical Association* 95(452): 1320–24.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Rosenstone, Steven J., and Raymond E. Wolfinger. 1978. "The Effect of Registration Laws on Voter Turnout." *American Political Science Review* 72(1): 22–45.

Signorino, Curtis S., and Ahmer Tarar. 2006. "A Unified Theory and Test of Extended Immediate Deterrence." *American Journal of Political Science* 50(3): 586–605.

Tomz, Michael, Jason Wittenberg, and Gary King. 2001. CLARIFY: Software for Interpreting and Presenting Statistical Results. Version 2.0. Cambridge, MA: Harvard University. http://gking.harvard.edu.

Wolfinger, Raymond E., and Steven J. Rosenstone. 1980. *Who Votes?* New Haven, CT: Yale University Press.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

- Section A: Comparison of the Average Case and Observed Value Approaches
  - Marginal Effect Analysis to Determine the Difference Between the Average Effect Using the Observed Value Approach and the Effect for the Average Case
    - Figure 1. Second Derivative of the Probit pdf by the Probability of Success for the Average Case
  - Discrete Difference Analysis
    - Figure A2. Difference in the Second Derivative of the Probit CDF Changing From $x_k = 0$ to $x_k = 1$ for Various Values of β_k by the Probability of Success for the Average Case When $x_k = 0$
- Section B: NES Results
  - Table 1. Probability of Voting for George W. Bush vs. John Kerry In 2004
  - Table 2. Predicted Probability of Voting for George W. Bush vs. John Kerry in 2004, Using

  - the Average Case and Observed Value Approaches, for Variables Not Shown in Figure 1
  - Table 3. Statistical Simulation Results for Predicted Probability of Voting for George W. Bush vs. John Kerry in 2004, Using the Observed Value Approach, with 95% Confidence Intervals
  - Table 4. Predicted Effect (First Difference) of Changes in Select Variables on the Probability of Voting for George W. Bush vs. John Kerry in 2004, Using the Observed Value Approach, with 95% Confidence Intervals

- Section C: Sample Stata Code (Version 10)
  - Stata Code for First Differences and Marginal Effects
  - Stata Code for Implementing the Observed Value Approach Via Simulation

- Tips for Other Models
- Section D: Additional Monte Carlo Simulation Results
  - Table 1. Monte Carlo Simulations Showing That the Average Case and Observed Value Approaches Do Not Converge as the Sample Size Increases (Using True Models)
    - Panel a. Marginal Effects for x_2 and x_3 for True Models
    - Panel b: Predicted Probability of Success Across Values of x_1 for True Models

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.