Postestimation Uncertainty in Limited Dependent Variable Models

Michael C. Herron
Northwestern University

Many political science research articles that use limited dependent variable models report estimated quantities, in particular, fitted probabilities, predicted probabilities, and functions of such probabilities, without indicating that such estimates are subject to uncertainty. This practice, along with the reporting of "percentage correctly predicted," can overstate the precision of reported results. In light of this, the present article describes a variety of measures of uncertainty that authors can include alongside estimates generated by limited dependent variable models. It also proposes an alternative to "percentage correctly predicted" and illustrates its calculations with congressional cosponsorship data from Krehbiel (1995).

1 Introduction

It is the contention of this article that estimates from limited dependent variable models should never be reported without some indication that they are subject to uncertainty. It is common, however, for political science research articles that use limited dependent variable models to report estimated quantities, in particular fitted probabilities, predicted probabilities, and functions of such probabilities, without measures of uncertainty. Consequently, these estimates appear to be more precise than they actually are. At best, model estimates reported without measures of uncertainty leave readers agnostic over matters of interpretation; at worst, such estimates overstate precision and can unintentionally mislead.

This article consists of two parts, each of which is motivated by the importance of measures of uncertainty insofar as they pertain to estimates produced by limited dependent variable models. The first part describes how to calculate standard errors and confidence intervals for fitted probabilities, predicted probabilities, and functions of these probabilities. The second argues that a common statistic, "percentage correctly predicted," that is frequently used to assess the fit of limited dependent variable models is almost always reported in a way that makes it appear more precise than it actually is. In place of "percentage correctly predicted," this article proposes and justifies an alternative statistic which is theoretically grounded and does not overstate precision.

Limited dependent variable models, such as probit, logit, multinomial probit, and so forth, are used by quantitative researchers when a dependent variable of interest can take on

Author's note: The author thanks Keith Krehbiel for comments and data and thanks Gary King, Jonathan Nagler, Jasjeet Sekhon, Wendy Tam, Craig Volden, and three anonymous referees for helpful comments on early drafts of this article. Software code is available at the *Political Analysis* website.

Copyright 1999 by the Society for Political Methodology

only one of several possible values. There is a vast body of literature on different types of limited dependent variable models, also known as categorical data models (Maddala 1983; Greene 1997; Agresti 1996). Within the political science literature, the most common limited dependent variable model is probably the logit although the probit is becoming increasingly popular. While political science research has also used other models in the limited dependent variable class (e.g., Alvarez and Brehm 1995, 1998; Alvarez and Nagler 1997; Bailey and Brady 1998; Lacy and Burden 1999), for ease of exposition this article's technical discussion focuses solely on the probit model. Similarly, the article's discussion of, critique of, and suggested replacement for "percentage correctly predicted" is based on the probit framework. However, the article's general points about uncertainty, estimate precision, and problems associated with "percentage correctly predicted" apply to all types of limited dependent variable models.[1]

The remainder of this article is organized as follows. The next section describes the basic probit model and highlights probit-based quantities that are typically of interest to quantitative political researchers. The article then discusses fitted and predicted probabilities in probit models, explains how they should be reported, and describes how standard errors and confidence intervals for such probabilities can be constructed. The subsequent section elaborates on the model fit statistic known as "percentage correctly predicted," explains how the statistic can overstate the accuracy of probit results, and proposes an alternative statistic that does not overstate precision. After briefly describing how the probit results presented in this article can be generalized to multinomial models and commenting on the difference between in-sample and out-of-sample estimates, the calculations discussed here are illustrated using data from Krehbiel (1995). The last section summarizes and concludes.

2 The Basic Probit Model

Let y_i^* denote an unobserved, latent variable indexed by observation i, and let y_i represent an observed, dichotomous dependent variable whose value, either 0 or 1, depends on y_i^*. In accordance with standard usage, assume that $y_i = 1$ if and only if $y_i^* \geq 0$ and $y_i = 0$ otherwise. Suppose that $y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$, where $\mathbf{x}_i$ is a covariate vector unique to observation i, $\boldsymbol{\beta}$ is a parameter vector common to all observations, and ϵ_i is a standard normal disturbance.[2] Given the set of assumptions, the probability p_i that $y_i = 1$ is

$$p_i = P(y_i^* \geq 0) = P(\mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i \geq 0) = P(\epsilon_i \geq -\mathbf{x}_i'\boldsymbol{\beta}) = 1 - \Phi(-\mathbf{x}_i'\boldsymbol{\beta}) \tag{1}$$

where $\Phi(\cdot)$ denotes the standard normal distribution function.

Based on the fact that y_i is a Bernoulli random variable, the following loglikelihood function can be used to estimate β:

$$L(\boldsymbol{\beta}) = \sum_i ((1 - y_i)\ \ln[\Phi(-\mathbf{x}_i'\boldsymbol{\beta})] + y_i \ln[1 - \Phi(-\mathbf{x}_i'\boldsymbol{\beta})]) \tag{2}$$

$\hat{\boldsymbol{\beta}}$ is chosen as the unique maximizer of the expression in Eq. (2), and let $\hat{\boldsymbol{\Sigma}}$ denote the

[1] The article's comments about precision and measures of uncertainty apply to statistical models in general. However, for reasons that are unclear, political science research that relies on limited dependent variable models is particularly problematic with respect to its frequent reporting of estimates without measures of uncertainty.

[2] It is the standard normal distributional assumption on ϵ_i that characterizes a probit model. Had it been assumed that ϵ_i had a logistic distribution, the resulting model would be a logit.

resulting covariance matrix.[3] Under standard regularity conditions, it follows that

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \overset{a}{\sim} N(\mathbf{0}, \hat{\boldsymbol{\Sigma}}) \tag{3}$$

There are two interpretations of the asymptotic result in Eq. (3). The frequentist interpretation is that $\boldsymbol{\beta}$ is fixed, the estimate $\hat{\boldsymbol{\beta}}$ is random, and $\hat{\boldsymbol{\Sigma}}$ reflects uncertainty about $\hat{\boldsymbol{\beta}}$. To Bayesians and what Tanner (1993, p. 16) calls likelihoodists, $\hat{\boldsymbol{\beta}}$ is fixed and the posterior distribution of $\boldsymbol{\beta}$ is approximately normal with mean vector $\hat{\boldsymbol{\beta}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$. This article adopts the latter perspective and considers $\boldsymbol{\beta}$ to be a random variable whose posterior distribution is characterized by $\boldsymbol{\beta} \overset{a}{\sim} N(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}})$. The importance of this expression becomes apparent when it is explained how functions of probit estimates can be calculated by sampling from the normal distribution with mean vector $\hat{\boldsymbol{\beta}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$. See King et al. (1998) on the interpretation of Eq. (3).

Substituting the estimate $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ in Eq. (1) yields fitted probabilities, $\hat{p}_i$. This "plug-in" notion is the same idea used to calculate fitted values in ordinary least squares models. In particular,

$$\hat{p}_i = 1 - \Phi(-\mathbf{x}_i' \hat{\boldsymbol{\beta}}) \tag{4}$$

At this point, the basic probit model and its likelihood function have been set up, and three estimated quantities, $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\Sigma}}$, and $\hat{p}_i$, have been described. In the following sections of the article, it is argued that the way in which many political science research articles report values of $\hat{p}_i$ and functions of these values can be misleading with respect to estimate precision.

3 Consequences of Postestimation Uncertainty

After probit estimation, $\boldsymbol{\beta}$ is a random variable. This is noted in Eq. (3), which indicates that residual or postestimation uncertainty surrounding $\boldsymbol{\beta}$ is captured by the posterior mean vector $\hat{\boldsymbol{\beta}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$. Consequently, a statement such as $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ would overstate the precision with which $\boldsymbol{\beta}$ is known. Moreover, if $\boldsymbol{\beta}$ is a random variable after estimation, then functions of $\boldsymbol{\beta}$ are random as well. In particular, randomness in $\boldsymbol{\beta}$ implies that the values of such functions cannot be known with complete certainty even after probit estimation. Therefore, when researchers estimate probit models and report functions of estimated $\boldsymbol{\beta}$ vectors, it is incumbent on them to identify residual uncertainty by also reporting standard errors and/or confidence intervals for the estimated function values.

From Eq. (4), it can be seen that $\hat{p}_i$, the estimate of p_i, is a function of $\boldsymbol{\beta}$ where the estimate $\hat{\boldsymbol{\beta}}$ has been substituted for the parameter vector $\boldsymbol{\beta}$. Perhaps the most common way to derive a confidence interval for $\hat{p}_i$ is, first, to use the delta method to estimate its standard error and, second, to specify the confidence interval based on the standard error. The delta method is a general approximation procedure for calculating the standard error of a possibly nonlinear function of a vector of random variables, and from Eq. (4) it can be seen that $\hat{p}_i$ is indeed a nonlinear function of $\boldsymbol{\beta}$ (when substituted for by $\hat{\boldsymbol{\beta}}$). The mathematics of the delta method can be seen via the following simple example.[4]

[3] $\hat{\boldsymbol{\Sigma}}$ can be estimated in a variety of fashions, and the details of these calculations are beyond the scope of this article. See, for example, Greene (1997, Chap. 4).

[4] The theory behind the delta method is explained in most textbooks on statistics. See, for example, Rice (1988, pp. 142–147) and Greene (1997, p. 124).

Let γ denote a scalar random variable with mean μ and variance σ^2 and let $f(\cdot)$ be a differentiable function. A first-order Taylor-series approximation of $f(\gamma)$ about μ yields

$$f(\gamma) \approx f(\mu) + f'(\mu)(\gamma - \mu) \tag{5}$$

Therefore,

$$\begin{aligned} \text{Var}[f(\gamma)] &\approx \text{Var}[f'(\mu)(\gamma - \mu)] \\ &\approx [f'(\mu)]^2 \sigma^2 \end{aligned} \tag{6}$$

Eq. (6) illustrates the well-known result that the variance of a function $f(\cdot)$ of a random variable γ can be approximated based on its first derivative $f'(\cdot)$ and the variance σ^2 of the underlying random variable. Extensions of Eq. (6) to vector-valued random variables are straightforward. Rather than containing a single first derivative as in Eq. (6), as illustrated below such vector-valued extensions include a matrix of first partial derivatives.

Suppose that there are k covariates in each observation's $\mathbf{x}_i$ vector so that $\hat{\boldsymbol{\beta}}$, the posterior mean of $\boldsymbol{\beta}$ (or the estimate of $\boldsymbol{\beta}$), is a k vector. This means that $\hat{p}_i$ is a function of a k-dimensional random vector. Let x_{ij} denote the (scalar) value of the jth covariate for observation i, where $1 \leq j \leq k$. Similarly, let $\hat{\beta}_j$ denote the jth element of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{G}_i$ the row vector of first derivatives of $\hat{p}_i$ with respect to $\hat{\boldsymbol{\beta}}$. Since $\hat{\boldsymbol{\beta}}$ has k elements, $\boldsymbol{G}_i$ has k elements as well, and let G_{ij} denote the jth element of $\boldsymbol{G}_i$. Then,

$$G_{ij} = \frac{\partial \hat{p}_i}{\partial \hat{\beta}_j} = \frac{\partial [1 - \Phi(-\mathbf{x}_i' \hat{\boldsymbol{\beta}})]}{\partial \hat{\beta}_j} = x_{ij} \phi(-\mathbf{x}_i' \hat{\boldsymbol{\beta}}) \tag{7}$$

where $\phi(\cdot)$ is the standard normal density function and the right-hand side of Eq. (4) substitutes for $\hat{p}_i$. According to the delta method as generalized to a k vector, the approximate variance of $\hat{p}_i$ is $\boldsymbol{G}_i \hat{\boldsymbol{\Sigma}} \mathbf{G}_i'$. When $k = 1$, $\mathbf{G}_i$ and $\hat{\boldsymbol{\Sigma}}$ are scalars, and this expression simplifies to that in Eq. (6). In general, though, $\mathbf{G}_i$ is a $1 \times k$ vector, $\hat{\boldsymbol{\Sigma}}$ is a $k \times k$ matrix, and $\mathbf{G}_i'$ is a $k \times 1$ vector, so $\mathbf{G}_i \hat{\boldsymbol{\Sigma}} \mathbf{G}_i'$ is a scalar.

Note the presence in $\mathbf{G}_i \hat{\boldsymbol{\Sigma}} \mathbf{G}_i'$ of $\hat{\boldsymbol{\Sigma}}$, the covariance matrix generated from maximizing the probit loglikelihood function in Eq. (2). The presence of $\hat{\boldsymbol{\Sigma}}$ implies that covariances between the individual elements of the $\hat{\boldsymbol{\beta}}$ vector are internalized into the delta method's variance calculation.

A researcher seeking a 95% confidence interval for $\hat{p}_i$ would use $(\hat{p}_i - 1.96\,\hat{\sigma}_i, \hat{p}_i + 1.96\,\hat{\sigma}_i)$, where $\hat{\sigma}_i = (\mathbf{G}_i \hat{\boldsymbol{\Sigma}} \mathbf{G}_i')^{\frac{1}{2}}$. If, say, such a confidence interval did *not* contain 0.5, the null hypothesis $p_i = 0.5$ could be rejected at the 0.05 level. Similarly, a t statistic for $\hat{p}_i$ could be calculated by dividing $\hat{p}_i$ by its estimated standard error $\hat{\sigma}_i$. If the t statistic were greater in magnitude than 1.96, one could argue that p_i was significantly different from zero at the 0.05 level. See Bailey and Brady (1998) for an example of these calculations.[5]

[5]Many computer packages automatically implement the delta method when asked to evaluate the variance of a function of estimated parameters, and these packages allow users to avoid the potentially tedious derivative calculations akin to those in Eq. (7). In Stata version 6, the command *testnl* implements the delta method. In TSP, the appropriate command is *analyz*. For example, to use *testnl* in Stata a researcher would estimate a probit regression using the *probit* command, define an equation using *eq*, and then use *testnl* on the equation. The Stata command *dprobit* automatically uses the delta method to calculate the standard errors of fitted probabilities. To use *analyz* from within TSP, a researcher would estimate a probit model using either the TSP command *probit* or

A confidence interval for $\hat{p}_i$ can also be generated with a simulation where such an approach involves drawing random vectors from the normal distribution in Eq. (3) and then computing $\hat{p}_i$ repeatedly.[6] The main reason that a researcher might want to use a simulation as opposed to the delta method is because derivatives, for example, those in Eq. (7), often necessitate tedious calculations. Furthermore, the delta method requires a linearization [see Eq. (5)] which can result in inaccuracies. On the other hand, a drawback to simulations is that replication becomes troublesome; namely, simulating the same confidence interval twice will generate two slightly different intervals. Moreover, accurate simulations often require a great number of iterations. Choosing between the delta method and a simulation therefore involves a trade-off among accuracy, computing speed, and importance of replicability.

Steps for a simulation that can be used to estimate a confidence interval for $\hat{p}_i$ are as follows.

1. Estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ using the loglikelihood function in Eq. (2).
2. Repeat the following steps S times (e.g., $S = 5000$), indexing each iteration by l.
   (a) Draw a vector $\tilde{\boldsymbol{\beta}}$ from a multivariate normal distribution with mean vector $\hat{\boldsymbol{\beta}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$.
   (b) Calculate $\hat{p}_i^l = 1 - \Phi(-\mathbf{x}_i'\tilde{\boldsymbol{\beta}})$
3. Use the percentiles of the histogram of $\hat{p}_i^l, l = 1, \ldots, S$, to form a confidence interval for $\hat{p}_i$.

For example, to construct a 95% confidence interval for $\hat{p}_i$, a researcher would determine the 0.025 and the 0.975 percentiles of the distribution of $\hat{p}_i^l$. More generally, for $\alpha \in (0, 1)$, a $(1 - \alpha)$ confidence interval could be derived using the $\alpha/2$ and $(1 - \alpha)/2$ percentiles of the distribution of $\hat{p}_i^l, l = 1, \ldots, S$.[7]

In the brief discussion of the delta method presented earlier, it was pointed out that the delta method variance calculation internalizes the covariances of the individual elements of the $\boldsymbol{\beta}$ vector. The same is true of the aforementioned simulation procedure. In particular, in each iteration of the simulation procedure, $\tilde{\boldsymbol{\beta}}$ is drawn from a multivariate normal distribution with covariance matrix $\hat{\boldsymbol{\Sigma}}$. This implies that covariance between the elements of the estimated $\boldsymbol{\beta}$ vector are not ignored when using the simulation procedure to derive a variance estimate for $\hat{p}_i$.

Consequences for $\hat{p}_i$ of changes in the covariate vector $\mathbf{x}_i$ should also be accompanied by standard errors or confidence intervals. For example, when studying legislative voting, a researcher might consider the impact on $\hat{p}_i$ of a change in i's political party membership. Whereas $\mathbf{x}_i$ denotes observation i's original covariate vector, let $\mathbf{x}_i^*$ be a modification of $\mathbf{x}_i$. To ascertain the impact on $\hat{p}_i$ of a change from $\mathbf{x}_i$ to $\mathbf{x}_i^*$, the following simulation could be used.

1. Estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ using the loglikelihood function in Eq. (2).
2. Repeat the following procedure S times, indexing each iteration by l.

ml and would then define a function with the command *frml*. The function would be the probability of interest, that is, a function of estimated parameters. Then *analyz* would be used on the defined formula and a standard error for the probability would be produced.

[6] Similar simulations and a discussion of them are given by Katz and King (1997) and King et al. (1998).

[7] Increasing the number of simulations beyond $S = 5000$ did not appreciably change the results in the latter section of this article. There may be situations, though, in which a greater number of simulations is necessary.

(a) Draw a vector $\tilde{\boldsymbol{\beta}}$ from a multivariate normal distribution with mean vector $\hat{\boldsymbol{\beta}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$.
(b) Calculate $\hat{p}_i^l = 1 - \Phi(-\mathbf{x}_i'\tilde{\boldsymbol{\beta}})$ and $\hat{p}_i^{l,*} = 1 - \Phi(-\mathbf{x}^{*\prime}{}_i\tilde{\boldsymbol{\beta}})$.
(c) Let $\delta^l = \hat{p}_i^l - \hat{p}_i^{l,*}$, where δ^l is the difference between the probability based on $\tilde{\boldsymbol{\beta}}$ and $\mathbf{x}_i$ and the probability based on $\tilde{\boldsymbol{\beta}}$ and the modified covariate vector $\mathbf{x}_i^*$.

3. After calculating S values of δ^l, percentiles of its distribution can be used to form a confidence interval for the impact on $\hat{p}_i$ of a change from $\mathbf{x}_i$ to $\mathbf{x}_i^*$.

For example, does a percentile-based 95% confidence interval for δ^l contain 0? If so, then it follows that a researcher could not reject the null hypothesis that a change from $\mathbf{x}_i$ to $\mathbf{x}_i^*$ has no impact on $\hat{p}_i$.

Finally, researchers often wish to estimate $\hat{p}_i$ for the average observation. What is frequently done is to calculate a covariate vector $\bar{\mathbf{x}}$ based on averaging $\mathbf{x}_i$ across all observations i. Then, define

$$\tilde{\bar{p}} = 1 - \Phi(-\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}) \tag{8}$$

Reports of $\tilde{\bar{p}}$, an estimate that depends on $\boldsymbol{\beta}$ (whose estimated or posterior mean is $\hat{\boldsymbol{\beta}}$), should always be accompanied by standard errors or confidence intervals.

A 95% confidence interval for $\tilde{\bar{p}}$ can be generated using the same simulation process described for $\hat{p}_i$. Similarly, to assess the impact on $\tilde{\bar{p}}$ of changes in $\bar{\mathbf{x}}$, a researcher would create a modified average covariate vector, $\bar{\mathbf{x}}^*$, sample S times from a multivariate normal distribution with mean vector $\hat{\boldsymbol{\beta}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$, repeatedly calculate $\tilde{\bar{p}} - \tilde{\bar{p}}^*$ based on S random draws, and then form a confidence interval for $\tilde{\bar{p}} - \tilde{\bar{p}}^*$ using the percentiles of the simulated distribution. If, for example, the confidence interval did not contain 0, then the null hypothesis $\tilde{\bar{p}} = \tilde{\bar{p}}^*$ could be rejected.

In addition, the delta method can be used to approximate the standard error of $\tilde{\bar{p}} - \tilde{\bar{p}}^*$, and then a researcher could calculate a t statistic based on $\tilde{\bar{p}} - \tilde{\bar{p}}^*$ divided by its approximate standard error. If, say, the magnitude of the t statistic were greater than the critical value of 1.96, then the researcher could reject at the 0.05 level the null hypothesis that $\tilde{\bar{p}} = \tilde{\bar{p}}^*$.

In particular, let $\mathbf{G}$ represent the k vector of first derivatives of $\tilde{\bar{p}} - \tilde{\bar{p}}^*$ with respect to $\hat{\boldsymbol{\beta}}$. The appropriate delta method equation is

$$\mathbf{G}_j = \frac{\partial(\tilde{\bar{p}} - \tilde{\bar{p}}^*)}{\partial\hat{\beta}_j} = \frac{\partial[-\Phi(-\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}) + \Phi(-\bar{\mathbf{x}}^{*\prime}\hat{\boldsymbol{\beta}})]}{\partial\hat{\beta}_j} = \bar{\mathbf{x}}_j\phi(-\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}) - \bar{x}_j^*\phi(-\bar{\mathbf{x}}^{*\prime}\hat{\boldsymbol{\beta}}) \tag{9}$$

where $\bar{\mathbf{x}}_j$ is the jth element of $\bar{\mathbf{x}}$, $\bar{x}_j^*$ is the jth element of $\bar{\mathbf{x}}^*$, $\mathbf{G}_j$ is the jth element of G, and $1 \le j \le k$. The approximate variance $\hat{\sigma}^2$ of $\tilde{\bar{p}} - \tilde{\bar{p}}^*$ is $\mathbf{G}\hat{\boldsymbol{\Sigma}}\mathbf{G}'$, and the relevant t statistic is $(\tilde{\bar{p}} - \tilde{\bar{p}}^*)/\hat{\sigma}$. Note that Eq. (9) is very similar to Eq. (7): the only difference between the two is the function being differentiated.

4 Fitted Probabilities and "Percentage Correctly Predicted"

Researchers who estimate probit models often seek a one-number summary of model fit, something akin to the R^2 value frequently reported alongside linear regression results. Within political science, probably the most common and popular one-number summary of probit model fit is the statistic known as "percentage correctly predicted" (PCP). This article now addresses PCP, critiques the statistic on precision grounds, and then presents an alternative to it.

The use of one-number model fit statistics like PCP has advantages as well as drawbacks. The most advantageous feature of such succinct summaries is simply that they reduce the complexity of a statistical model to a single number. This can lead to one-number model fit statistics that are comparable across different models, although not all such statistics can be compared in this fashion. On the other hand, the primary drawback of one-number model fit summaries like PCP is that their concise nature leads them to ignore the many facets of model fit. Namely, as illustrated in the following discussion of PCP, any statistic that boils down a model based on a large number of observations and covariates to a single number is practically bound to discard useful information. Beyond PCP, various other one-number fit statistics for limited dependent variable models are discussed by Train (1986) and Menard (1995). The use of model deviance as a fit statistic is discussed by McCullagh and Nelder (1989, Chap. 4).

In a probit model with excellent fit, $\hat{p}_i$ values close to 1 should have associated y_i values of 1. This is because $\hat{p}_i$ is an estimate of p_i, the probability that $y_i = 1$. Similarly, $\hat{p}_i$ values close to 0 should be associated with $y_i = 0$. In a loose sense, when there are many observations that have either $\hat{p}_i$ close to 1 and $y_i = 1$ or $\hat{p}_i$ close to 0 and $y_i = 0$, it is logical to conclude that the associated probit model provides a good fit for the data being studied.

Let $\hat{y}_i$ denote observation i's fitted category. As y_i is 0 or 1, the same applies to $\hat{y}_i$. Assuming for the moment that $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ (this strong assumption is relaxed shortly), the probability that $\hat{y}_i = 1$ is $\hat{p}_i$ and the probability that $\hat{y}_i = 0$ is $1 - \hat{p}_i$. The expectation of $\hat{y}_i$ is

$$\mathrm{E}(\hat{y}_i) = 1 \times \hat{p}_i + 0 \times (1 - \hat{p}_i) = \hat{p}_i \tag{10}$$

Note that $\hat{y}_i$ is the realization of a random variable whose distribution depends on the fitted probability $\hat{p}_i$. In other words, while $\hat{p}_i$ describes the distribution of $\hat{y}_i$, there is no way to calculate $\hat{y}_i$ directly without additional assumptions that relate fitted probabilities $\hat{p}_i$ to fitted categories $\hat{y}_i$. And, as can be seen from Eq. (10), the mean of $\hat{y}_i$ is not a permissible value of $\hat{y}_i$ since $\hat{p}_i$ must be 0 or 1.[8]

As defined in traditional usage, PCP for a probit model is the percentage of cases for which $\hat{y}_i = y_i$. Broadly speaking, the reason that researchers report PCP is because, *ceteris paribus*, models with high PCP are preferable to those with low PCP.[9] Since the calculation of PCP requires assessing whether $y_i = \hat{y}_i$ for all observations i, it follows that the value of PCP cannot be calculated without specifying $\hat{y}_i$. This is a problem because $\hat{y}_i$ remains unknown even with knowledge of $\hat{\boldsymbol{\beta}}$. Articles seeking to report PCP must somehow manipulate knowledge of the distribution of $\hat{y}_i$ into knowledge of $\hat{y}_i$ itself.

PCP is typically defined by the following procedure.

1. Estimate $\hat{\boldsymbol{\beta}}$ using the probit loglikelihood function in Eq. (2), and for each observation i, calculate $\hat{p}_i$ using Eq. (4).
2. For those observations with $\hat{p}_i \geq 0.5$, set $\hat{y}_i = 1$; otherwise set $\hat{y}_i = 0$.

[8] In ordinary least squares models, fitted values are calculated as expectations. The permissibility problem that arises in probit models is not an issue in ordinary least squares models since in the latter there are no restrictions on dependent variable values.

[9] The argument here is not that researchers use PCP to discriminate between competing models. Typically, likelihood-ratio or Wald tests would be employed in this capacity. Researchers do, however, use PCP as a measure of model accuracy or fit, and they have been known to trumpet high values of PCP as evidence of a good model fit.

3. Call each observation i with $y_i = \hat{y}_i$ a correct prediction. PCP is defined as the percentage of observations that are correctly predicted.

PCP is a classification algorithm that depends on the estimate $\hat{\boldsymbol{\beta}}$ and also depends on a classifying rule based on $\hat{p}_i \geq 0.5$.[10] However, there are two problems associated with the statistic. The first problem is somewhat minor and it pertains to PCP's use of the word "predicted." In particular, PCP is an in-sample statistic because it is based on observations with known values of y_i. Therefore, use of the word "predicted" in PCP is slightly inappropriate, as PCP does not attempt to assess the quality of out-of-sample predictions. While it is arguable that PCP should actually be known as "percentage correctly classified", for compatibility with the existing statistical terminology, this article retains the PCP acronym.

The second problem with PCP is that its assignment rule for $\hat{y}_i$ can make PCP classifications appear more precise than they actually are. Again, consider the situation in which $\boldsymbol{\beta}$ is assumed to be known exactly. If $\hat{p}_i \geq 0.5$, it is appropriate to conclude that it is *more likely* that $\hat{y}_i = 1$ than it is that $\hat{y}_i = 0$. But it is potentially misleading to say that $\hat{p}_i \geq 0.5$ *implies* $\hat{y}_i = 1$. In fact, such an implication can be false. Even if $\hat{p}_i = 0.99$, there is still a 1% chance that $\hat{y}_i = 0$. This subtlety is hidden by PCP, which makes it appear that probit models are capable of assigning fitted categories, when, based on the estimate or posterior mean $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, they can only assign distributions over such categories.[11]

Broadly speaking, the reason that PCP can overstate probit model precision is that it conflates a *best guess* of $\hat{y}_i$ with *knowledge* of $\hat{y}_i$. PCP disregards the possibility of errors in best guesses when it assumes that $\hat{p}_i \geq 0.5$ implies $\hat{y}_i = 1$. Furthermore, PCP treats $\hat{p}_i = 0.51$ and $\hat{p}_i = 0.99$ in the same fashion, this despite the fact that the former value of $\hat{p}_i$ says much less than the latter. Since a value of $\hat{p}_i$ close to $\frac{1}{2}$ indicates greater uncertainty about $\hat{y}_i$ compared to a value of $\hat{p}_i$ that is close to 1, it follows that PCP's classification rule can overstate precision.

Given PCP's problem with precision, should research articles that use probit models even report PCP? Because the classification rule that defines PCP can cover up model uncertainty by treating fitted categories as known without error, it is the contention of this article that researchers should not report the statistic. As described shortly, this article proposes a modification of PCP that takes into consideration the randomness inherent in $\hat{y}_i$. This modification should be reported in lieu of standard PCP.

Before delving into the article's proposed alternative to PCP, first consider the model fit statistic known as "percentage reduction in error" (PRE). See Hagle and Mitchell (1992) and Menard (1995) for details. Given a dichotomous dependent variable and a probit analysis of it, PRE is based on a comparison of PCP with PMC, the percentage of observations in the modal category of the observed data. For example, if a probit data set has 100 observations and $y_i = 1$ for 60 of them, then PMC = 0.6. PMC is a known quantity since it does not depend on any estimated quantities.

Again, assuming that $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, PRE is defined as

$$\text{PRE} = \frac{\text{PCP} - \text{PMC}}{1 - \text{PMC}} \tag{11}$$

[10] Maximizing the probit loglikelihood in Eq. (2) will not necessarily lead to an estimate $\hat{\boldsymbol{\beta}}$ that maximizes correct classifications or predictions.

[11] This problem is even worse for multinomial models such as ordered probits. In such models, a fitted category that comes from a process similar to PCP may not even reflect the category with the highest fitted probability.

PRE seeks to compare the information provided by probit fitted categories with the classification errors a researcher would make if she naively assigned all fitted categories to the modal category. If PCP is less than PMC, then the PCP-based classification errors are actually greater than the classification errors a researcher would generate if she eschewed probit estimation and classified observations based only on modal category.

However, since PRE is a function of PCP, the former carries with it the same precision problems that accompany the latter statistic. In particular, PRE relies on calculations of fitted categories that ignore randomness in $\hat{y}_i$, and PRE can therefore understate the extent of uncertainty surrounding probit results. Nonetheless, if a researcher reports PCP contrary to the recommendations of this article, PMC or PRE should always be reported as well.

In lieu of reporting PCP, this article proposes that researchers report ePCP where "e" stands for expected. ePCP is calculated in a manner which reflects the fact that $\hat{p}_i$ is a probability and that researchers never know the value of $\hat{y}_i$ exactly. In addition, the derivation of ePCP recognizes that best guesses about $\hat{y}_i$ can be mistaken. Therefore, rather than assuming that $\hat{p}_i \geq 0.5$ implies $\hat{y}_i = 1$, ePCP is defined as the sum over all observations i of the probability that $\hat{y}_i = y_i$. Given a data set of size N,

$$\text{ePCP} = \frac{1}{N}\left(\sum_{y_i=1} \hat{p}_i + \sum_{y_i=0}(1 - \hat{p}_i)\right) \tag{12}$$

Suppose that $N = 3$, $y_1 = 0$, $y_2 = 1$, and $y_3 = 1$, and in addition, suppose that $\hat{p}_1 = 0.6$, $\hat{p}_2 = 0.6$, and $\hat{p}_3 = 0.8$. Then, ePCP is $\frac{1}{3}(0.4 + 0.6 + 0.8) = 0.6$. PCP in this example has a value of two-thirds, which, clearly, is greater than ePCP. This is not a general result, however. It is possible that a given data set may have ePCP greater than PCP.

ePCP can be justified in two ways. First, suppose that one were to assign values of $\hat{y}_i$ for all observations i by assuming that $\hat{y}_i = 1$ if $\hat{p}_i \geq 0.5$ and $\hat{y}_i = 0$ otherwise. As noted earlier, this is a logical way to proceed, given that $\hat{p}_i \geq 0.5$ implies that the probability that $\hat{y}_i = 1$ is at least as great as the probability that $\hat{y}_i = 0$. Were one to assign fitted values $\hat{y}_i$ in such a manner, ePCP is the *expected percentage of correct predictions*. In other words, ePCP is the expected percentage of correct predictions given that one uses the assignment rule that defines PCP.[12]

A second justification for ePCP is based on long-run frequencies. Suppose that a researcher were to iterate repeatedly the assignment process for fitted categories in the following manner. On iteration j, for each observation i assume that $\hat{y}_i = 1$ with probability $\hat{p}_i$ and $\hat{y}_i = 0$ with probability $1 - \hat{p}_i$. Then, let CP_j be the percentage of correct predictions on iteration j, and let $\overline{\text{CP}}_n = 1/n \sum_{j=1}^{n} \text{CP}_j$ be the average number of correct predictions over n iterations ("CP" is used here to distinguish it from PCP). Since $\text{E}(\text{CP}_j) = \text{ePCP}$ for each j and since CP_j is bounded, it follows from the strong law of large numbers that $\overline{\text{CP}}_n \rightarrow \text{ePCP}$ almost surely as $n \rightarrow \infty$. Importantly, the average number of correct predictions $\overline{\text{CP}}_n$ will not in the long run equal PCP. By itself, this suggests that PCP may be a problematic statistic with respect to model fit.

Note that ePCP, in contrast to PCP, distinguishes between large and small values of $\hat{p}_i$. Suppose for a given observation i that $y_i = 1$ and that $\hat{p}_i = 0.99$. In this scenario, guessing that $\hat{y}_i = 1$ will be correct 99% of the time. Therefore, i's contribution to ePCP would be

[12] This is based on holding β fixed at its estimate of $\hat{\beta}$. This is why the expectation of PCP, need not be concerned with the nonlinearity of the normal distribution function.

0.99, a number close to 1. However, were $\hat{p}_i = 0.51$, only 51% of the time would $\hat{y}_i$ be correct. In this case, i's contribution to ePCP would be 0.51.[13]

While the formula for ePCP in Eq. (12) relies on the estimate or posterior mean $\hat{\boldsymbol{\beta}}$ through $\hat{p}_i$, it does not allow for the fact that uncertainty over $\boldsymbol{\beta}$ remains after probit estimation. Recall the numerous caveats about $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ in the previous section. These caveats should be taken into consideration when calculating a statistic based on fitted categories because fitted categories are themselves a function of $\boldsymbol{\beta}$, a random variable. Rather than calculating and reporting ePCP as a single number between 0 and 1, a researcher could use the following procedure to report a simulation-based confidence interval around ePCP.

1. Estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ using the loglikelihood function in Eq. (2).
2. Repeat the following steps S times indexing each iteration by l.
   (a) Draw a vector $\tilde{\boldsymbol{\beta}}$ from a multivariate normal distribution with mean vector $\hat{\boldsymbol{\beta}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$.
   (b) For each observation i, define $\tilde{p}_i$ as the value of p_i where $\tilde{\boldsymbol{\beta}}$ is used in place of $\boldsymbol{\beta}$. See Eq. (1).
   (c) Using Eq. (12), calculate ePCP based on $\tilde{p}_i$.
   (d) Set ePCP_l as the value of ePCP given the draw of $\tilde{\boldsymbol{\beta}}$.
3. Examine a histogram of ePCP_l, $l = 1, \ldots, S$, and form a confidence interval for ePCP based on the percentiles of this distribution.

After estimating a probit model, a valid way to describe the relation between observed categories and fitted categories is with a confidence interval around ePCP. This is because analysis of fitted categories requires consideration of two distinct levels of randomness. At the top, $\hat{y}_i$ is random and its distribution is determined by fitted probabilities $\hat{p}_i$. But, due to the fact that $\boldsymbol{\beta}$ is a random variable even after probit estimation, $\hat{p}_i$ remains uncertain as well. Both these sources of randomness are internalized in the ePCP confidence interval, but both are ignored in PCP.[14]

5 Extending ePCP Beyond Probit

To maintain a sense of continuity throughout the article, the discussion of fitted probabilities $\hat{p}_i$, PCP, and ePCP has focused on the two-category probit model. This section briefly describes how to extend the article's calculations beyond two-category probit models. Extension to logit models is straightforward, as the only difference between probit and logit is the presence of a logistically distributed error term.

For a multinomial limited dependent variable model, one would calculate ePCP based on the number of categories in one's model. For example, suppose that a researcher has a three-category model, an estimated parameter vector $\hat{\boldsymbol{\beta}}$, and three relevant fitted probabilities, $\hat{p}_{i,0}$, $\hat{p}_{i,1}$, and $\hat{p}_{i,2}$ such that $\hat{p}_{i,0} + \hat{p}_{i,1} + \hat{p}_{i,2} = 1$. Note the additional subscripts on fitted

[13] In a manner akin to PCP, PRE could be adjusted so that it reflects the fact that fitted categories are not known without error. Furthermore, the denominator of PRE [see Eq. (11)] can be similarly adjusted. If PMC $= \alpha$, then with probability α each observation would be classified as falling in the modal category. This would provide an ePCP-like framework for PRE and PMC.

[14] A similar one-number model fit statistic is pseudo-R^2 as defined by McFadden (1974). The latter statistic is automatically calculated by Stata when a user estimates a probit model. The prefix "pseudo" indicates that the pseudo-R^2 statistic used to evaluate probit model fit is in some sense similar to the familiar R^2 common in ordinary least squares regressions. Pseudo-R^2 is by no means the only formulation of an R^2-like fit statistic for probit models, and various other measures are discussed by Maddala (1983, Chap. 2).

probabilities when dealing with a model with more than two categories. If, for instance, observation i were truly in category c for $c \in \{0, 1, 2\}$, the contribution to the multinomial ePCP from observation i would be the value of $\hat{p}_{i,c}$. A confidence interval for the multinomial ePCP could be generated by sampling from a multivariate normal distribution and repeatedly calculating multinomial ePCP.

Similarly, extending the delta method to multinomial models requires only that the function being differentiated in Eq. (7) or (9) be replaced by whatever probability expression is appropriate. Using simulations to generate confidence intervals for estimated probabilities or changes in estimated probabilities is likewise straightforward. One would repeatedly draw $\tilde{\boldsymbol{\beta}}$ from a multivariate normal distribution and compute probabilities based on $\tilde{\boldsymbol{\beta}}$. These steps follow logically and directly from the probit instructions described earlier.

6 Out-of-Sample Predictions and Postestimation Uncertainty

Suppose that a researcher estimates a probit model for a group of N observations based on covariate data $\mathbf{x}_i$ and a categorical dependent variable y_i. This produces $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\Sigma}}$, and $\hat{p}_i, i = 1, \ldots, N$. Now consider how this researcher can predict or forecast the categories for T new observations for which neither y_i nor y_i^* is observed. In particular, suppose that the researcher wants to predict the fraction F of the T observations that have $y_i = 1$, $i = N + 1, \ldots, N + T$. This example, in contrast to the standard nomenclature for PCP, reflects a logical use of the word "prediction." The fraction F of interest pertains to an out-of-sample group of observations that was not used to derive $\hat{\boldsymbol{\beta}}$.

There is a variety of situations in which out-of-sample predictions may be useful. One such situation occurs when a researcher seeks to predict the voting behavior of individuals who abstained in an election. Or a researcher might want to predict how members of Congress who abstained from voting on a given bill would have cast their ballots conditional on voting.

Another situation where out-of-sample predictions are used is cross-validation. A researcher using cross-validation in a probit problem would divide a sample in half, estimate $\hat{\boldsymbol{\beta}}$ based on the first half of the data set, and then "predict" the categories of the second half of the observations based on $\hat{\boldsymbol{\beta}}$ from the first half. In this case, "predict" is in quotes because, in truth, the researcher actually knows the categories for the second half of the observations. However, the notion of cross-validation is that the predicted categories for the second half of the dataset should be as accurate as the correctly classified categories in the first half.

For each of the T additional observations not used in estimating $\hat{\boldsymbol{\beta}}$, let $\tilde{y}_i^*$ denote the predicted value of y_i^* based on the estimate $\hat{\boldsymbol{\beta}}$. That is, $\tilde{y}_i^* = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$, $i = N + 1, \ldots, N + T$. Based on Eq. (1), it might be tempting to posit that $\tilde{y}_i^* \geq 0$ implies $\tilde{y}_i = 1$. Then F would be defined as the fraction of the T observations that have $\tilde{y}_i^* \geq 0$.

Defining and reporting F using the $\tilde{y}_i^* \geq 0$ classification rule, however, would lead to a prediction that overstates precision. Namely, this type of definition would disguise the fact that $\hat{\boldsymbol{\beta}}$ is only an estimate of $\boldsymbol{\beta}$ and that $\tilde{y}_i$ is the realization of a random variable whose distribution is characterized by $\tilde{y}_i^*$. That is, $P(\tilde{y}_i = 1) = \Phi(\tilde{y}_i^*)$, but it is not true that $\tilde{y}_i^* \geq 0$ implies $\tilde{y}_i = 1$. Consequently, F should be estimated and reported using a simulation similar to that used to characterize the ePCP confidence interval.

In particular, after deriving $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ using $\mathbf{x}_i$ and $y_i, i = 1, \ldots, N$, a researcher would repeatedly draw a vector $\tilde{\boldsymbol{\beta}}$ from a multivariate normal distribution with mean vector $\hat{\boldsymbol{\beta}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$. For each observation i, $i = N + 1, \ldots, N + T$, define $\tilde{p}_i$ as the value

of p_i where $\tilde{\beta}$ is used in place of β. Then

$$F_j = \frac{1}{T} \sum_{i=N+1}^{N+T} \tilde{p}_i \tag{13}$$

where F_j is the expected fraction of observations in category one. A histogram of F_j, $j = 1, \ldots, S$ can then be used to form a confidence interval for F based on the percentiles of this distribution. The long-run interpretation of ePCP also applies to calculations of F_j.

The average of the S versions of F is the point estimate of the fraction of observations that are in category one, whereas the confidence interval around F recognizes that there will always be uncertainty as to the fraction's magnitude. Eq. (13) can be generalized in a straightforward way to models with polychotomous dependent variables.

Within the political science literature, published predictions for category totals—this is what F represents—are often not accompanied by measures of uncertainty. For example, Alvarez and Nagler's (1995) analysis of the 1992 American presidential election estimates how Perot voters would have allocated themselves had Perot not run in 1992. Alvarez and Nagler conclude that in such an instance 49.5% of actual Perot voters would have sided with Bush and 50.5% would have opted for Clinton. These two predictions—49.5 and 50.5%—are estimated quantities and therefore should be accompanied by confidence intervals or standard errors.

7 Postestimation Uncertainty: Another Look at "A to Z"

A number of the calculations described in this article are now illustrated using data from "Cosponsors and Wafflers from A to Z," an article by Keith Krehbiel (1995) on the "A to Z Spending Plan."[15] Table 2 of Krehbiel's article reports probit estimates for an analysis of the House members who chose to cosponsor this bill, formally known as H.R. 3266. Of the 434 House members analyzed in Krehbiel's article, 228 cosponsored the resolution. Therefore, since $228/434 \approx 0.525 > 0.5$, the modal category of House members consists of cosponsors (i.e., PMC = 52.5%).

The independent variables that were used in Krehbiel's analysis of H.R. 3266 cosponsors are a constant, a measure of House member liberalism published by Americans for Democratic Action (ADA), a measure of fiscal conservatism published by the National Taxpayers' Union (NTU), an indicator variable for Democratic Party membership, a measure of Congressional seniority, the previous electoral margin, an indicator variable for membership on the House Appropriations Committee, and an indicator variable for membership on the House Budget Committee. As specified in the Krehbiel article, the average House member has average seniority, average electoral margin, median ADA and NTU scores, is nonpartisan, and is a member of neither the Appropriations nor the Budget Committee.[16]

In his discussion of who cosponsored H.R. 3266, Krehbiel estimates five probit models using different combinations of independent variables. For each model Krehbiel's coefficient estimates were replicated, subject to a few minor coding errors, and this discussion

[15] Krehbiel was requested by the editor of the *American Journal of Political Science* to include PCP in his article. Krehbiel's (1995) data are also the subject of Binder et al. (1998).

[16] Nonpartisan means that the indicator variable for Democratic Party membership is set to the average value (the fraction of individuals who are Democrats) from the sample of 434 House members. See page 912 and, in particular, footnote 8, of Krehbiel's (1995) article.

Table 1 PCP, ePCP, and 95% ePCP confidence intervals

	Model				
	1	*2*	*3*	*4*	*5*
PCP	82.7	90.8	90.8	90.3	90.6
ePCP	77.6	85.4	85.4	85.9	86.9
ePCP interval	(75.5, 79.1)	(83.4, 86.7)	(83.3, 86.7)	(83.9, 87.1)	(84.8, 87.9)

focuses on values of PCP and ePCP rather than on coefficient estimates themselves.[17] Krehbiel used analytic second derivatives when estimating his five different covariance matrices, and this article therefore does the same.

Table 1 lists by model number PCP as reported in Table 2 of Krehbiel's article. In addition, the table reports for each model ePCP and a 95% confidence interval for ePCP.[18] The ePCP confidence intervals are based on the 0.025 and 0.975 percentiles of 5000-draw simulations. All figures in Table 1 are percentages.

As noted in Table 1, ePCP is less that PCP for all five models, and this indicates that PCP overstates the correct predictions of the models that appear in Krehbiel's article. In other words, once it is recognized that probit estimates generate only a distribution over fitted categories rather than fitted categories themselves, model classifications or predictions turn out to be less accurate than suggested by the traditional PCP statistic. The latter point is accentuated by the ePCP values in the table. Furthermore, it follows from the fact that the five 95% confidence intervals for ePCP in Table 1 do not include the five respective values for PCP.[19] The ePCP confidence intervals reported in Table 1 are based on a simulation, and Fig. 1 plots a histogram of the simulated values of ePCP for Krehbiel's Model 5.

Using the estimates from Krehbiel's Model 5 (which includes all independent variables used in Krehbiel's article), the probability of cosponsorship for the average House member is 0.318, with an estimated standard error of 0.0615. Therefore, using the delta method, a 95% confidence interval for this probability is (0.197, 0.439). Moreover, as described in the body of this article another useful way to calculate confidence intervals for estimated

Table 2 Estimated changes in cosponsorship probability for Model 5

Variable	*Probability change*	*SE*	*t*
ADA score	−0.0883	0.0589	−1.50
NTU score	0.646	0.0685	9.43
Party membership	0.315	0.131	2.40

[17] Four ADA scores using Krehbiel's (1995) original article were corrected by Binder et al. (1998). This article's use of data on the "A to Z Spending Plan" is for illustrative purposes only. In particular, the intent of the analysis here is not to enter into the debate on the House bill, a focus of Krehbiel (1995), Binder et al. (1999), and Krehbiel (1999).

[18] PMC is the same for each model in Table 1 because each model considers the same set of House members, 52.5% of whom cosponsored H.R. 3266.

[19] Based on Model 5, one might argue that 86.9% is reasonably close to 90.6% and therefore that PCP is just as good as ePCP. Such an assertion would be troubling for two reasons. First, 86.9% $<$ 90.6%, and if quantitative political science is going to be meaningful, then numbers matter. In other words, statistics that overstate precision should be considered dubious regardless of the extent to which they do so. Second, PCP of Model 5 is outside the 95% confidence interval for ePCP.

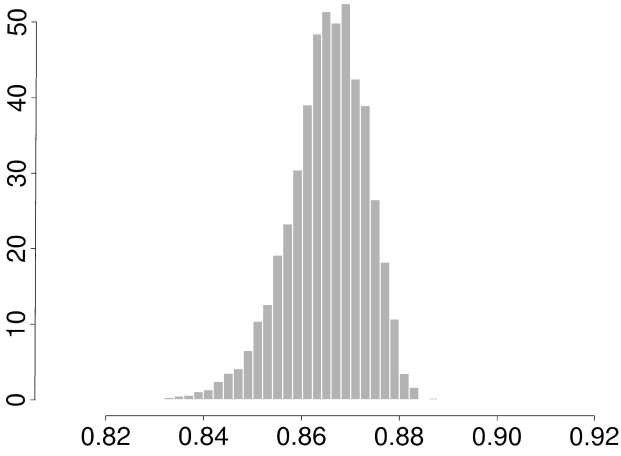

Fig. 1 Histogram of simulated ePCP (Krehbiel's Model 5).

probabilities is via simulation. Simulated cosponsorship probabilities are pictured in Fig. 2, and the 0.025 and 0.975 percentiles of the histogram in the figure lead to a 95% confidence interval of (0.206, 0.447). This is very close to the previously described confidence interval as calculated by the delta method.

The delta method of Eq. (9) can also be used to estimate the impact on cosponsorship probability of changes in ADA score, NTU score, and party membership. For each independent variable change, the new probability of cosponsorship is calculated and compared to a base probability. The base probability is 0.318 for the ADA and NTU changes, and for the party affiliation change the probability of cosponsorship for a Democratic House member with otherwise average characteristics is compared to the probability of cosponsorship for a Republican with otherwise average characteristics. Results are given in Table 2.

For the average House member, the hypothesis that a one standard deviation increase in the ADA score would not have impacted the cosponsorship probability cannot be rejected at

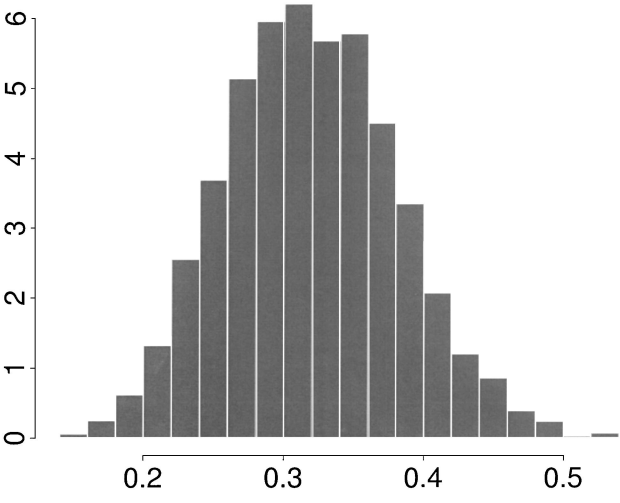

Fig. 2 Simulated cosponsorship probabilities for the average House member.

the 0.05 level. This is because the probability change's t statistic, -1.50, is less in magnitude than 1.96, the 0.05 critical value. Statistical insignificance of this probability change would not be apparent in a hypothetical results table that reported neither standard errors nor confidence intervals for estimated probability changes. In other words, this example illustrates precisely why estimated probability changes should always be reported with measures of uncertainty.

It is important to recognize that p values for significance tests on estimated probit coefficients do *not* inherently specify p values for tests on related probability changes. This is because the normal distribution function that constitutes part of the probit model is nonlinear in its parameters. In linear least squares models, however, significance tests on regression coefficients are equivalent to tests for whether changes in regressors have nonzero impacts on an outcome variable. Thus, results generated by probit models and limited dependent variable models in general require measures of estimate uncertainty that go beyond those that should be reported alongside results from ordinary least squares models.

Overall, this article's reanalysis of Krehbiel's study of the "A to Z Spending Plan" illustrates two main points. First, the extent to which the various cosponsorship models in the article fit the H.R. 3266 cosponsorship data was not as good as that described by the common and popular fit statistic PCP. Within Krehbiel's sequence of five models, the difference between PCP and ePCP hovered around 5 percentage points. Although in an arbitrary probit regression model the degree to which PCP will differ from ePCP is model-specific, the discussion of H.R. 3266 cosponsorship illustrates precisely the type of precision problems that can affect PCP.

Second, the use of standard errors and confidence intervals highlights the differences in various estimated probabilities associated with H.R. 3266 cosponsorship probabilities. Of the three probability changes reported in Table 2, only two are individually statistically significant at the 0.05 level. Had these estimated probability changes been reported in the fashion very common for political science research articles, this lack of statistical significance would not have been apparent.

8 Conclusion

This article has focused on estimated quantities produced by limited dependent variable models, a class of statistical models that is very common in quantitative political science research. The technical discussion revolved around probit models, but the points regarding precision and uncertainty apply to limited dependent variable models in general.

The article had three objectives. The first was to emphasize the importance of supplying standard errors and/or confidence intervals when reporting fitted and predicted probabilities generated by probit models. The second was to argue that a particular model fit statistic, "percentage correctly predicted" (PCP), that often accompanies limited dependent variable models is, first, inappropriately named—it really has little to do with prediction—and, second, can overstate the accuracy of reported results. The third objective of the article was to suggest a replacement for PCP. Thus, a measure denoted "expected percentage correctly predicted" (ePCP) was presented, the theory behind it was explained, and steps of a simulation for generating ePCP confidence intervals were detailed.

Accordingly, this article offers the following two recommendations. First, researchers should always include measures of uncertainty when reporting estimated probabilities and probability changes that are produced by limited dependent variable models. In particular, probabilities generated by such models should be accompanied by confidence intervals or

standard errors generated by either the delta method or a simulation. Second, researchers should not report the common and popular model fit statistic known as PCP. Instead, they should report what this paper has labeled ePCP and they should include with it a suitable confidence interval.

References

Agresti, Alan. 1996. *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons.

Alvarez, R. Michael, and John Brehm. 1995. "American Ambivalence Toward Abortion Policy: A Heteroskedastic Probit Method for Assessing Conflicting Values." *American Journal of Political Science* 39:1055–1082.

Alvarez, R. Michael, and John Brehm. 1998. "Speaking in Two Voices: American Equivocation about the Internal Revenue Service." *American Journal of Political Science* 42:418–452.

Alvarez, R. Michael, and Jonathan Nagler. 1995. "Economics, Issues, and the Perot Candidacy: Voter Choice in the 1992 Presidential Election." *American Journal of Political Science* 39:714–744.

Alvarez, R. Michael, and Jonathan Nagler. 1997. "When Politics and Models Collide: Estimating Models of Multiparty Elections." *American Journal of Political Science* 42:55–96.

Bailey, Michael, and David W. Brady. 1998. "Heterogeneity and Representation: The Senate and Free Trade." *American Journal of Political Science* 42:524–544.

Binder, Sarah A., Eric D. Lawrence, and Forrest Maltzman. 1999. "Uncovering the Hidden Effect of Party." *Journal of Politics* 61:815–831.

Greene, William H. 1997. *Econometric Analysis*, 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Hagle, Timothy M., and Glen E. Mitchell, II. 1992. "Goodness-of-Fit Measures for Probit and Logit." *American Journal of Political Science* 36:762–784.

Katz, Jonathan N., and Gary King. 1997. "A Statistical Model for Multiparty Electoral Data." California Institute of Technology, Social Science Working Paper 1005.

King, Gary, Michael Tomz, and Jason Wittenberg. 1998. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." Paper presented at the annual meeting of the American Political Science Association, Boston.

Krehbiel, Keith. 1995. "Cosponsors and Wafflers from A to Z." *American Journal of Political Science* 39:906–923.

Krehbiel, Keith. 1999. "The Party Effect from A-to-Z and Beyond." *Journal of Politics* 61:832–840.

Lacy, Dean, and Barry C. Burden. 1999. "The Vote-Stealing and Turnout Effects of Ross Perot in the 1992 U.S. Presidential Election." *American Journal of Political Science* 43:233–255.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. Cambridge: Chapman and Hall.

McFadden, Daniel. 1974. "The Measurement of Urban Travel Demand." *Journal of Public Economics* 3:303–328.

Menard, Scott. 1995. *Applied Logistic Regression Analysis*, Quantitative Applications in the Social Sciences, No. 106. Thousand Oaks, CA: Sage.

Rice, John A. 1988. *Mathematical Statistics and Data Analysis*. Belmont, CA: Wadsworth & Brooks.

Tanner, Martin A. 1993. *Tools for Statistical Inference*, 2nd ed. New York: Springer-Verlag.

Train, Kenneth E. 1986. *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand*. Boston: MIT Press.