# Interpreting Probit Analysis
## Jonathan Nagler
## Revised - March 3, 1994

### Problems of OLS

Regression via ordinary least squares (OLS) is a commonly applied statistical technique in political science. However, when the dependent variable is dichotomous (0-1) rather than continuous, ordinary least squares becomes an inefficient estimation technique, and the underlying linear probability model (LPM) that is being estimated represents a poor apriori choice of model specification (Aldrich and Nelson, 1984). The linear probability model assumes that the expected value of the dependent variable, or the probability that the dependent variable takes the value 1, is a linear combination of some set of independent variables. Or,

$$E[Y_i] \;=\; \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i \tag{1}$$

The well known property of the OLS estimators is that they are BLUE (Best Linear Unbiased Estimators) if and only if the Gauss-Markov Assumptions are met. One of the GM assumptions is that the variance $\sigma^2$ of the disturbance term $u$ is constant (i.e., $\sigma_i^2 = \sigma^2$ for all $i$). It is easy to show that this condition is not met if the dependent variable is dichotomous (Aldrich and Nelson, p. 13). Under this condition of heteroscedasticity the OLS estimates themselves will be unbiased, but estimates of their standard errors will be invalid. Since it is impossible to know if the computed standard errors are too large or too small, any statistical tests, and hence any inferences about the possible range of population parameters, will be meaningless.

However, there is a second problem of using OLS with dichotomous variables that is particularly troubling for the analysis produced here. When the dependent variable is dichotomous, the prediction of Y offered by OLS $(\hat{Y})$ is interpreted as a prediction of the probability that Y takes on the value 1 (i.e., $\hat{Y} = E[Pr(Y_i = 1)]$). Since this is a probability, it should be bounded by 0 and 1. However, nothing constrains the predictions of the LPM model offered by OLS from being either less than 0 or greater than 1.

This lack of boundedness is related to a fundamental feature of the linear probability model, namely, it is linear in the independent variables. Changes in the independent variables are assumed to have a constant affect on the dependent variable. If $x_1$ has a coefficient of 2, then any unit change of $x_1$ will cause a change in $Y$ of 2, whether $x_1$ goes from 0 to 1, or whether $x_1$ goes from 1000 to 1001. This naturally allows the probability to be unconstrained. Arbitrarily large or small probabilities can be predicted for arbitrarily large or small values of $x_1$. The practical interpretation of this is that individuals who are initially almost certain not to vote, and individuals who are undecided, undergo the same change in probability of voting for a given stimulus.

The following topics regarding the probit model are covered below:

- Development of the Basic Model for Dichotomous Variables
- Estimation (briefly)
- Inference about the Parameters (i.e., t-tests for $\hat{\beta}$)
- Interpreting goodness of fit of the model based on 'Guess-Rate'
- Predicted Values
- The Effects of the Coefficients on Parameters
    1. Effect of Unit Change from Mean Values
    2. The .5 Baseline Method
    3. Predictions at Sample Values
    4. Direct Estimation of the Derivative

This paper does *not* cover the actual maximum likelihood technique used to generate the estimates. [Note that Estimation is listed as being *briefly* covered above!]

**The Probit Model**

The common solution to the deficiencies of the LPM model as estimated via OLS is to adopt a different model specification. The Probit model constrains the estimated probabilities to be between 0 and 1, and relaxes the constraint that the effect of independent variables is constant across different predicted values of the dependent variable. In common parlance, the probit model assumes an S-shaped response curve such that in each tail of the curve the dependent variable, $Pr(Y_i = 1)$, responds slowly to changes in the independent variables, while towards the middle of the curve, i.e., towards the point where $Pr(Y_i = 1)$ is closest to .5, the dependent variable responds more swiftly to changes in the independent variables (Figure 1).

The probit model assumes that while we only observe the values of 0 and 1 for the variable $Y$, there is a latent, unobserved continuous variable $Y^*$ that determines the value of $Y$. We assume that $Y^*$ can be specified as follows:

$$Y_i^* = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i \tag{1}$$

and that:

$$\begin{aligned} Y_i &= 1 \quad if \quad Y_i^* > 0 \\ Y_i &= 0 \;\; otherwise. \end{aligned}$$

where $x_1, x_2, ..., x_k$ represent vectors of random variables, and $u$ represents a random disturbance term.

Now from equation 1,

$$Pr(Y_i = 1) = Pr(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i > 0) \tag{2}$$

Rearranging terms,

$$\begin{aligned} Pr(Y_i = 1) &= Pr(u_i > -(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki})) \\ &= 1 - Pr(u_i < -(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki})) \\ &= 1 - F(-(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki})) \end{aligned} \tag{3}$$

where F is the cumulative density function of the variable $u$. If we make the usual assumption that $u$ is normally distributed, we have:

$$\begin{aligned} Pr(Y_i = 1) &= 1 - \Phi(-(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki})) \\ &= 1 - \Phi(-X_i\beta) \\ &= \Phi(X_i\beta) \end{aligned} \tag{4}$$

where $\Phi$ represents the cumulative normal distribution function.

Using maximum likelihood techniques we can compute estimates of the coefficients ($\beta s$) and their corresponding standard errors that are asymptotically efficient. However, these estimates cannot be interpreted in the same manner that normal regression coefficients are. These coefficients give the impact of the independent variables on the latent variable $Y^*$, not $Y$ itself. To transfer $Y^*$ into a probability estimate for $Y$ we compute the cumulative normal of $Y^*$. Because of this transformation there is no linear relationship between the coefficients and $Pr(Y_i = 1)$. Hence the change in $Pr(Y_i = 1)$ caused by a given change in $x_{ji}$ will depend upon the value of all of the other $x$s and their corresponding coefficients, or more precisely on the value of the sum $X_i\beta$, as well as the change in $x_{ji}$. To see this, look at the shape of the cumulative normal function (Figure 1). It is steepest in the middle, and flatter at the tails. To simplify things, assume there is only one independent variable, and that $\beta_0 = 0$, and $\beta_1 = 1$. Hence the expression in equation 6 simplifies to $Pr(Y_i = 1) = \Phi(x_1)$. As the points on Figure 1 indicate, a change in $x_1$ from 0 to 1 causes a change in $Pr(Y_i = 1)$ of .34 (.84 -.50); while a change in $x_1$ from -1.65 to -.65 (also a change of only 1 unit) causes a change in $Pr(Y_i = 1)$ of .21 (.26 - .05). Hence simply knowing the change in $x$ can not tell us the predicted change in $Pr(Y_i = 1)$, that change depends upon where on the curve we start.

**[Figure 1 Here]**

Imagine now that Figure 1 represents a model with two independent variables, $x_1$ and $x_2$. Now the effect of a given change in $x_1$ upon $Pr(Y_i = 1)$ will depend upon the sum $(x_1\beta_1 + x_2\beta_2)$. The effect of $x_1$ will be greatest when the sum $(x_1\beta_1 + x_2\beta_2)$ is closest to 0, and weakest as $(x_1\beta_1 + x_2\beta_2)$ approaches $\pm\infty$. But this means that the effect of $x_1$ depends upon the value of $x_2$. Or, without making any substantive inference about the relationship between $x_1$ and $x_2$, we have been able to conclude that they 'interactively' affect $Pr(Y_i = 1)$. This interactive effect is assumed in the model specification. And the interaction is assumed to be greatest when $(x_1\beta_1 + x_2\beta_2)$ is closest to 0, or, when $\Phi(x_1\beta_1 + x_2\beta_2) = Pr(Y_i = 1)$ is closest to 0.5. This suggests the hazards of drawing inferences about relationships among independent variables from probit estimates of predicted probabilities. Below I attempt to clearly illustrate this problem with a sample analysis. I briefly review several common means of presenting and interpreting probit estimates to show how these techniques can suggest interactive effects even without the use of explicit interactive terms in the underlying model.[1]

Before proceeding to the data analysis it is important to realize that the above need not dissuade us from attempting to identify true interactive effects when using the probit model. Our estimates of $Pr(Y_i = 1)$ are based on a transformation of an underlying model generating the latent variable $Y^*$. $Y^*$ is the variable of interest, though it is unobserved. Since the model generating $Y^*$ is linear, and thus not contaminated by assumed interactive effects, we can use this underlying model to perform tests for substantive interactive relationships among independent variables. To do so requires adding explicit interactive terms to the model specification.

## Section II: Presenting and Interpreting Probit Estimates

To help elucidate the analysis of interactive terms in the probit model I first review interpretations of individual probit coefficients. An understanding of proper techniques for individual inferences is both a useful preliminary to understanding group analysis, and is essential for attempts at individual level confirmation of group phenomena.

What techniques for presenting and interpreting probit estimates have in common is that each attempts to illustrate the effect of changes in an independent variable on the *probability* that the dependent variable takes the value 1. To illustrate each of these techniques I use a multivariate model explaining an individual's decision whether to vote as a function of respondent's: education, income, age, race, student status, ability to work, environment (urban/rural), and the number of days before the election that registration in respondent's state closes. This model is a simplification of the Wolfinger-Rosenstone (1980) model. I present the simpler model here for heuristic reasons.

### Method 1: Effect of Unit Change from Mean Values

The simplest technique used to present probit estimates is to set each independent variable to its mean (or mode for discrete variables), and show the effect on $Pr(Y_i = 1)$ as the independent variables vary one at a time. This is done by first computing $Pr(Y_i = 1) = \Phi(X_i\beta)$ with $x_k = \bar{x}_k$ for all $k$. [For notational simplicity I use $\bar{x}_k$ to denote the mode of $x_k$ for discrete variables, as well as the mean of continuous variables.] To determine the impact of a unit change in $x_j$ on $Pr(Y_i = 1)$, $x_j$ is set equal to $\bar{x}_j + 1$, and $Pr(Y_i = 1) = \Phi(X_i^*\beta)$ is recomputed, where $X_i^* = [\bar{X} \mid \bar{x}_j + 1]$. The first probability is subtracted from the second probability, and this difference is the impact of a unit change in $x_j$ *when all other independent variables are held at their mean or mode.* A different value would be arrived at if all other $x$s were, for example, set to 0. Again, this is because of the shape of the cumulative normal function (Figure 2).

Setting each independent variable to its mean does *not* guarantee that you will be in the steepest part of the curve. In fact there is no reason at all to expect this unless you are dealing with a phenomena that occurs approximately 50% of the time. For instance, if one examined voting behavior in countries such as Sweden where turnout is 90% (i.e., the mean $Pr(Y_i = 1)$ is .9), setting the independent variables to their mean would put us on the flat *tail* of the S-shaped curve. Alternatively, setting each independent variable to its mean in an examination in the U.S., where turnout is approximately 50%, would put us on the middle of the curve.

**[Table 1]**

An example of this method is offered in Table 1. For each variable, the table presents both the probit coefficient, and the estimated change in the probability of an individual voting for a one unit change in the variable. Reading from the table, Family Income has a coefficient of 0.0686; and an additional 'unit' of family income - an increase from the mean value of 7.3 to 8.3, would result in an increased probability of voting of .0234. This change in probability is easily understandable. It is the increase in the likelihood of an individual voting who is a 'typical' white, non-student, i.e., the modal individual. Since the probability of voting is in fact our dependent variable, this value of .0234 is much more meaningful than a coefficient of .0686, which only tells us the effect on the unobserved variable $Y^*$.

There are two important caveats that go with this method. First, the estimated change is conditional upon the other independent variables having specific values (their mean or mode). Second, the effect of income on voting will not be linear. An additional two units of family income will *not* result in an increased probability of voting of .0468. (To examine effects of hypothetical changes in the independent variables see Section 3, below.)

Reporting the change in $Pr(Y_i = 1)$ is a superior means of reporting the relationship between changes in the independent and dependent variables. However, if the researcher is interested in comparing the relative effects of income and education on voting, then it may not be an appropriate technique because a one unit change in income and a one unit change in education are different things. They are functions of the way each independent variable is measured and scaled; and hence the changes in probability reported in Table 1 are results of this measurement and scaling as well. For comparisons among the relative impacts of different independent variables, as well as estimates of the magnitudes of impacts, Method 3 below is more appropriate.

### Method 2: The .5 Baseline Method

An alternative to examining the effects with all variables set to their mean or mode is to examine effects on individuals for whom $Pr(Y_i = 1)$ is closest to .5. This is accomplished by choosing values of the independent variables such that $Pr(Y_i = 1) = 0.5$. Then one can successively let each independent variable vary by one unit, or alternatively let each independent variable vary by a standard deviation (or some standardized unit), and observe the effect on the change in $Pr(Y_i = 1)$. There is no reason that this method is better or worse than method 1 above. However, since this method evaluates changes where the slope of the cumulative normal is greatest, it will necessarily *always* have the effect of maximizing the variation in $Pr(Y_i = 1)$ reported based on changes in $x$. For cases where the dependent variable occurs approximately 50% of the time, the two methods will produce almost identical answers. Since this is the case for voting in the United States, I do not report the estimates here in this manner. [2]

### Method 3: Predictions at Sample Values

Method 1 answers a limited question: what is the effect on $Pr(Y_i = 1)$ of a *unit* change in a variable $x_j$ when all other variables are held at their mean or mode. But since as we have emphasized, the rate of change is not continuous, this method can not help us make accurate predictions when $x_j$ varies for *more* than one unit from its mean. An alternative then to showing the effect of a unit (or standardized-unit) change in $x_j$ is to evaluate $Pr(Y_i = 1)$ for *several* different values of $x_j$. For instance, one computes $Pr(Y_i = 1|[X^*| \ x_j = a1 \ \ ])$, $Pr(Y_i = 1|[X^*| \ x_j = a2])$, $Pr(Y_i = 1|[X^*| \ x_j = a3])$; where $a1$, $a2$, and $a3$ are arbitrary values with $a1 \ < \ a2 \ < \ a3$, and $X^* = \bar{x}_k$ for all $k \ \neq \ j$. This shows the effect on the probability of $Y$ being equal to 1 as the independent variable $x_j$ increases from $a1$ to $a3$, when all other independent variables are held at their mean. Since the researcher may choose $a_1$, $a_2$, and $a_3$, this is the most precise method available for showing the effects on the dependent variable of postulated changes in the independent variables.

The final column of Table 2 shows the estimated probability of an individual voting given 3 different levels of education: individuals with only an eighth grade education ($\widehat{Pr_i} =$ .53), high school graduates ($\widehat{Pr_i} = .69$), and those with four years of college ($\widehat{Pr_i} = .88$). The calculation is performed with all variables besides education held at their mean or mode. This is a clear way to show the effect of rising levels of education. If all other variables are held constant, an individual with an eighth grade education would be 35% more likely to vote had they gone on to complete college. According to these estimates, education has a potent effect on voting.

**[Table 2 Here]**

At this point it would be useful to compare the effect of changes in education to changes in other independent variables. Rather than redo the computation described above for each independent variable, the technique is frequently modified by letting one other independent variable, say $x_l$, also vary across several values. This requires computing 9 values of $Pr(Y_i = 1)$, 1 for each pair: $(x_j = a1,\ x_l = b1), (x_j = a1,\ x_l = b2), ..., (x_j = a3,\ x_l = b3)$. Again, all of the other independent variables are held at their mean or mode. This allows the researcher to show how various combinations of two different variables affect the probability of the dependent variable taking on the value 1.

The first four columns of Table 2 show the effect of changing levels of income on individuals with different levels of education. Reading across the first row of the table, as the income of an individual with an eighth grade education goes from less than \$1000 to over \$25,000 the probability of that individual's voting goes from .36 to .62, an increase of .26. Alternatively, if an individual with eight years of education earning less than \$1000 increased their education level to 4 years of college, the new estimate would indicate that they had a probability of voting of .76, an increase of .40. Since the dependent variable changed less in response to the maximum possible change in income than it did for a large change in education, a reasonable inference is that education is more important in determining voting than income is.

It is significant that the changes across rows (or down columns) are not identical. While individuals with an eighth grade education increase their probability of voting by 26% via increased income, individuals with 4 years of college increase their probability of voting only 16% for identical income increases.

### Method 4: Direct Estimation of the Derivative

Finally, another method that can be used to report the effects of probit coefficients is to compute the derivative of $Pr(Y_i = 1)$ with respect to each independent variable, and report these derivatives directly. The derivates, being of the form $\partial(\Phi(X_i\beta))/\partial(x_j) = \phi(X_i\beta)\beta_j$ will depend on the values of *all* of the independent variables when they are evaluated, and will represent a *continuous* rate of change. Hence it gives the slope of a line tangent to the surface. The final column of Table 1 reports these derivatives with respect to each independent variable.

These derivatives are probably the closest thing to OLS coefficients available as they show the response of the dependent variable ($Pr(Y_i = 1)$) in response to changes in the independent variable. However, using a tangent line to predict a rate of change for a curve that is postulated to be S-shaped is problematic. The derivative is continuous, and changes as any independent variable changes. Hence the derivative may have one value at the point where we 'start' our calculation, and a very different value where we finish it. Figure 2 shows the probability that $Y_i = 1$ for the values 0 and 1 of $x$. It also shows the tangent line drawn at the point $x = 0$. At $x = 0$, $\phi(x) = .4$. Hence the slope of the tangent line is .4, and it realizes a value of .9 at $x = 1$. By contrast, we can see that according to the curve, $Pr(Y_i = 1) = .84$ for $x = 1$. Thus following the tangent line, rather than the probit curve, generates overestimates of $Pr(Y_i = 1)$. [3]

**[Figure 2 Here]**

### Section III

Each of the methods of presenting probit results offered above shows the effect of changes in explanatory variables on the behavior of individuals, or more precisely on the *probability* of individuals choosing to vote. However, we may be interested in knowing the effect of changes in an explanatory variable on the overall population. Table 1 showed the effect of a change in 1 day of the closing requirement for registration on the probability of an *individual* voting, with all other variables held at their mean. What if we wanted to know how many additional persons would vote overall if the closing requirement for registration were relaxed by one day? What if we wanted to know the effect of changes in an explanatory variable on subgroups of the population?

In most work on the effects of registration laws on voting turnout it is not really the individual who is the center of attention. Rather, the substantive interest is in groups of individuals sharing a common trait: blacks, poor people, women, etc. In addition, a key explanatory variable of interest is a systemic, rather than individual, characteristic - the restrictiveness of the registration laws in a state. The question becomes how to show the effect of the registration laws on a *group* of individuals.

In *Who Votes?*, Wolfinger and Rosenstone used a complicated, but effective, technique to estimate such effects. First, they used a probit model to compute a predicted probability of voting, $\widehat{Pr_i} = \Phi(X_i\beta)$ for each individual $i$ in their sample. They then made the appropriate change in the explanatory variable of interest (they reset the number of days to registration closing to zero) and recomputed the probability of voting for each individual. Call this new hypothetical probability $\widetilde{Pr_i}$. $\widetilde{Pr_i}$ gives the estimated probability of the $i^{th}$ individual voting *if* there were no registration requirement. Each individual's actual values of the other independent variables are used to compute their own hypothetical probabilities; there is no reason to substitute values for any of the independent variables besides days to closing. In other words, $\widetilde{Pr_i} = \Phi(\tilde{X}_i\hat{\beta})$, where $\tilde{X}_i = [X_i|x_{ji} = 0]$, and $x_j$ is the variable of interest (closing) that is being hypothetically set to 0. The impact of the registration requirement on the $i^{th}$ individual is arrived at by subtracting the first number $(\widehat{Pr_i})$ from the second number $(\widetilde{Pr_i})$.

# Notes

[1] See King (1989) for an alternative discussion of presenting probit coefficients.

[2] For an example, see Jackson and King, 1989.

[3] The derivative will be a particularly inaccurate predictor of change if one of our independent variables is dichotomous.

# Table 1

## Effects of Changes in Explanatory Variables of Probit Model

## Dependent Variable: Probability of Voting

| Independent Variable | Est Coeff | t-Stat | Change in Pr(Vote) Per Unit Change in Indep Variable[a] | Deriv of Pr(Vote) w respect to each Indep Variable |
|---|---|---|---|---|
| Intercept | -2.1893 | 17.84 | – | – |
| Education Squared | 0.0272 | 21.45 | .0094 | .0097 |
| Family Income | 0.0686 | 10.20 | .0234 | .0245 |
| Black | 0.1020 | 1.92 | .0346 | .0365 |
| Age | 0.0621 | 13.26 | .0213 | .0222 |
| Age Squared | -0.0005 | -9.26 | -.0002 | -.0002 |
| Student | 0.2687 | 3.06 | .0865 | .0960 |
| Unable to Work | -0.5948 | -5.28 | -.2284 | -.2126 |
| Rural Nonfarm | -0.0493 | -1.38 | -.0174 | -.0176 |
| South | -0.1902 | -5.29 | -.0693 | -.0680 |
| Closing Days | -0.0076 | -3.99 | -.0027 | -.0027 |
| | | | | |
| N | | 8377 | | |
| Percent Voting | | 65.68 | | |
| Percent Correctly Predicted | | 71.09 | | |
| Log Likelihood | | -4784.6 | | |

[a] Changes calculated with all other independent variables set to their mean or mode value.

## Table 2

### Predicted Values of Probability of Voting: For Selected Values of Education and Family Income (All Other Variables Set to Their Mean or Mode Value)

| | Family Income | | | | |
|---|---|---|---|---|---|
| Years of Education | ≤ $1000 | $4000-4999 | $7500-9999 | $25,000 + | Total |
| 8 | .36 | .46 | .54 | .62 | .53 |
| 12 | .53 | .63 | .71 | .77 | .69 |
| 4 college | .76 | .84 | .88 | .92 | .88 |