3 Binary Outcomes: The Linear Probability, Probit, and Logit Models

Binary dependent variables are extremely common in the social sciences. Maddala and Trost (1982) studied the decisions by a bank to accept loan applications. Domencich and McFadden (1975) analyzed factors affecting the use of public versus private transportation for commuting. Aldrich and Cnudde (1975) considered the decision to vote for McGovern in the 1972 presidential election; Allen (1991) examined contributions by the corporate elite to the Democratic Party; while Ragsdale (1984) studied the president's decision to make a discretionary speech to the nation. Other outcomes include whether fraud was committed by a savings and loan institution (Tillman & Pantell, 1995); if a trainee decided to remain with the sponsoring employer (Gunderson, 1974); and whether a student collaborated with his or her mentor during graduate study (Long, 1990). Even a cursory glance at recent journals in the social sciences turns up dozens of additional examples, ranging from having intercourse before marriage, dropping out of high school, joining a union, to enlisting in the military.

In this chapter, I present four models for the analysis of binary outcomes: the linear probability model (LPM), the binary probit model, the binary logit model, and, briefly, the complementary log-log model. The LPM is the linear regression model applied to a binary dependent variable. While I do not recommend the LPM, the model illustrates the problems resulting from a binary dependent variable, and motivates our discussion of the logit and probit models. The probit and logit models are developed first in terms of the regression of a latent variable. The latent variable is related to the observed, binary variable in a simple way: if the latent variable is greater than some value, the observed variable is 1; otherwise it is 0. This model is linear in the latent variable, but results in a nonlinear, S-shaped model relating the independent variables to the probability that an event has occurred. Given the great similarity between the logit and probit models, I refer to them jointly as the *binary response model*, abbreviated as BRM. The BRM is also developed as a nonlinear probability model. Within this context, the complementary log-log model is introduced as an asymmetric alternative to the logit and probit models.

3.1. The Linear Probability Model

The *linear probability model* is the regression model applied to a binary dependent variable. The structural model is

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$$

where $\mathbf{x}_i$ is a vector of values for the ith observation, $\boldsymbol{\beta}$ is a vector of parameters, and ε is the error term. $y = 1$ when some event occurs, and $y = 0$ if the event does not occur. For example, $y = 1$ if a woman is in the paid labor force, and $y = 0$ if she is not. If we have a single independent variable, the model can be written as

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

which is plotted in Figure 3.1. The conditional expectation of y given x, $E(y \mid x) = \alpha + \beta x$, is shown as a solid line. Observations are plotted as circles at $y = 0$ and $y = 1$.

To understand the LPM, we must consider the meaning of $E(y \mid \mathbf{x})$. When y is a binary random variable, the unconditional expectation of y is the probability that the event occurs:

$$E(y_i) = [1 \times \Pr(y_i = 1)] + [0 \times \Pr(y_i = 0)] = \Pr(y_i = 1)$$

For the regression model, we are taking conditional expectations:

$$E(y_i \mid \mathbf{x}_i) = [1 \times \Pr(y_i = 1 \mid \mathbf{x}_i)] + [0 \times \Pr(y_i = 0 \mid \mathbf{x}_i)] = \Pr(y_i = 1 \mid \mathbf{x}_i)$$

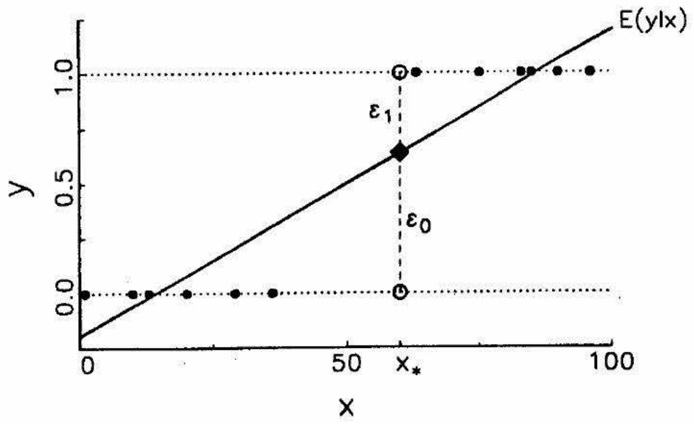


Figure 3.1. Linear Probability Model for a Single Independent Variable

Therefore, the expected value of y given $\mathbf{x}$ is the probability that $y = 1$ given $\mathbf{x}$. This allows us to rewrite the LPM as

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = \mathbf{x}_i \boldsymbol{\beta}$$

Having a binary outcome does not affect the interpretation of the parameters that was presented in Chapter 2: for a unit increase in x_k, the expected change in the probability of an event occurring is β_k, holding all other variables constant. Since the model is linear, a unit change in x_k always results in the same change in the probability. That is, the model is linear in the probability, and hence the name *linear probability model.*

Example of the LPM: Labor Force Participation

Many authors have presented models in which the dependent variable is whether a married woman was in the paid labor force. For example, Gunderson (1974) compares the use of logit, probit, and LPM models. Nakamura and Nakamura (1981, pp. 464–468) use a probit model to compare labor force participation in the United States and Canada. While Mroz (1987) focuses on models for a woman's hours of paid labor, he uses a probit model to correct for sample selection bias. Berndt (1991, pp. 618–619) reviews the research in this area.

TABLE 3.1 Descriptive Statistics for the Labor Force Participation Example

Name	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*	*Description*
LFP	0.57	0.50	0.00	1.00	1 if wife is in the paid labor force; else 0
K5	0.24	0.52	0.00	3.00	Number of children ages 5 and younger
K618	1.35	1.32	0.00	8.00	Number of children ages 6 to 18
AGE	42.54	8.07	30.00	60.00	Wife's age in years
WC	0.28	0.45	0.00	1.00	1 if wife attended college; else 0
HC	0.39	0.49	0.00	1.00	1 if husband attended college; else 0
LWG	1.10	0.59	−2.05	3.22	Log of wife's estimated wage rate
INC	20.13	11.63	−0.03	96.00	Family income excluding wife's wages

NOTE: $N = 753$.

Our analysis is based on data extracted by Mroz (1987) from the 1976 Panel Study of Income Dynamics.[1] The sample consists of 753 white, married women between the ages of 30 and 60. The dependent variable *LFP* is 1 if a woman is employed and is 0 otherwise. The independent variables, which are similar to those used by Nakamura and Nakamura (1981), Mroz (1987), and Berndt (1991), are listed in Table 3.1. Our measures of educational attainment are dummy variables indicating whether the husband or wife spent at least one year in college, rather than the more commonly used measures of the number of years of education. This was done to illustrate the interpretation of dummy independent variables.

The model being estimated is

$$LFP = \beta_0 + \beta_1 K5 + \beta_2 K618 + \beta_3 AGE + \beta_4 WC + \beta_5 HC + \beta_6 LWG + \beta_7 INC + \varepsilon$$

with estimates presented in Table 3.2. Interpretation is straightforward. For example:

- *Unstandardized coefficients for continuous variables.* For every additional child under 6, the predicted probability of a woman being employed decreases by .30, holding all other variables constant.
- *x-standardized coefficients for continuous variables.* For a standard deviation increase in family income, the predicted probability of being employed decreases by .08, holding all other variables constant.
- *Unstandardized coefficients for dummy variables.* If the wife attended college, the predicted probability of being in the labor force increases by .16, holding all other variables constant.

[1] These data were generously made available by Thomas Mroz.

TABLE 3.2 Linear Probability Model of Labor Force Participation

Variable	β	β^{S_x}	t
Constant	1.144	—	9.00
K5	−0.295	−0.154	−8.21
K618	−0.011	−0.115	−0.80
AGE	−0.013	−0.103	−5.02
WC	0.164	—	3.57
HC	0.019	—	0.45
LWG	0.123	0.072	4.07
INC	−0.007	−0.079	−4.30

NOTE: $N = 753$. β is an unstandardized coefficient; β^{S_x} is an x-standardized coefficient; t is a t-test of β.

There are several things to note about these interpretations. First, the effect of a variable is the same regardless of the values of the other variables. Second, the effect of a unit change for a variable is the same regardless of the current value of that variable. For example, if a woman has four young children compared to no young children, her predicted probability of employment decreases by 1.18 (= 4 × −.295), which is obviously unrealistic. This problem is considered in the next section. Finally, fully standardized and y-standardized coefficients are inappropriate given the binary outcome, and x-standardized coefficients are inappropriate for binary independent variables.

3.1.1. Problems With the LPM

While the interpretation of the parameters is unaffected by having a binary outcome, several assumptions of the LRM are necessarily violated.

Heteroscedasticity. If a binary random variable has mean μ, then its variance is $\mu(1-\mu)$. (*Prove this.*) Since the expected value of y given $\mathbf{x}$ is $\mathbf{x}\boldsymbol{\beta}$, the conditional variance of y depends on $\mathbf{x}$ according to the equation:

$$\text{Var}(y\,|\,\mathbf{x}) = \Pr(y = 1\,|\,\mathbf{x})[1 - \Pr(y = 1\,|\,\mathbf{x})] = \mathbf{x}\boldsymbol{\beta}(1 - \mathbf{x}\boldsymbol{\beta})$$

which implies that the variance of the errors depends on the x's and is not constant. (*Plot the* $\text{Var}(y\,|\,\mathbf{x})$ *as* $\mathbf{x}\boldsymbol{\beta}$ *ranges from* −.2 *to* 1.2.) Since the LPM is heteroscedastic, the OLS estimator of $\boldsymbol{\beta}$ is inefficient and the standard errors are biased, resulting in incorrect test statistics.

Goldberger (1964, pp. 248–250) suggested that the LPM could be corrected for heteroscedasticity with a two-step estimator. In the first step, $\widehat{y}$ is estimated by OLS. In the second step, the model is estimated with generalized least squares using $\widehat{\text{Var}}(\widehat{\varepsilon}) = \widehat{y}(1-\widehat{y})$ to correct for heteroscedasticity. While this approach increases the efficiency of the estimates, it does not correct for other problems with the LPM. Further, for $\widehat{y} < 0$ or $\widehat{y} > 1$, the estimated variance is negative and ad hoc adjustments are required.

Normality. Consider a specific value of x, say x_*. In Figure 3.1, $E(y\,|\,x_*)$ is represented by a diamond on the regression line. ε is the distance from $E(y\,|\,x)$ to the observed value. Since y can only have the values 0 and 1, which are indicated by the open circles, the error must equal either $\varepsilon_1 = 1 - E(y\,|\,x_*)$ or $\varepsilon_0 = 0 - E(y\,|\,x_*)$. Clearly, the errors cannot be normally distributed. Recall that normality is not required for the OLS estimates to be unbiased.

Nonsensical Predictions. The LPM predicts values of y that are negative or greater than 1. Given our interpretation of $E(y\,|\,\mathbf{x})$ as $\Pr(y = 1\,|\,\mathbf{x})$, this leads to nonsensical predictions for the probabilities. For example, using the means in Table 3.1 and the LPM estimates in Table 3.2, we find that a 35-year-old woman with four young children, who did not attend college nor did her husband, and who is average on other variables, has a predicted probability of being employed of −.48. (*Verify this result.*) While unreasonable predictions are sometimes used to dismiss the LPM, such predications at extreme values of the independent variables are also common in regressions with continuous outcomes.

Functional Form. Since the model is linear, a unit increase in x_k results in a constant change of β_k in the probability of an event, holding all other variables constant. The increase is the same regardless of the current value of $\mathbf{x}$. In many applications, this is unrealistic. For example, with the LPM each additional young child decreases the probability of being employed by .295, which implies that a woman with four young children has a probability that is 1.18 less than that of a woman without young children, all other variables being held constant. More realistically, each additional child would have a diminishing effect on the probability. While the first child might decrease the probability by .3, the second child might only decrease the probability an additional .2, and so on. That is to say, the model should be nonlinear. In general, when the outcome is a probability, it is often *substantively* reasonable that the effects of independent variables will have diminishing returns as the predicted probability approaches 0 or 1. In my opinion, the most serious problem with the LPM is its functional form.

The binary response model has an S-shaped relationship between the independent variables and the probability of an event, which addresses the problem with the functional form in the LPM. In the following section I develop this model in terms of a latent dependent variable. Section 3.4 shows how the logit and probit models can also be thought of as nonlinear probability models without appealing to a latent variable. And, in Chapter 6, the models are derived as *discrete choice models* in which an individual chooses the option that maximizes her utility.

3.2. A Latent Variable Model for Binary Variables

As with the LPM, we have an observed binary variable y. Suppose that there is an unobserved or *latent* variable y^* ranging from $-\infty$ to ∞ that generates the observed y's. Those who have larger values of y^* are observed as $y = 1$, while those with smaller values of y^* are observed as $y = 0$.

Since the notion of a latent variable is central to this approach to deriving the BRM, it is important to understand what is meant by a latent variable. Consider a woman's labor force participation as the observed y. The variable y can only be observed in two states: a woman is in the labor force, or she is not. However, not all women in the labor force are there with the same certainty. One woman might be very close to the decision of leaving the labor force, while another woman could be very firm in her decision. In both cases, we observe the same $y = 1$. The idea of a latent y^* is that there is an underlying propensity to work that generates the observed state. While we cannot directly observe y^*, at some point a change in y^* results in a change in what we observe, namely, whether a woman is in the labor force. For example, as the number of young children in the family increases, it is reasonable that a woman's propensity to be in the labor force (as opposed to working at home) would decrease. At some point, the propensity would cross a threshold that would result in a decision to leave the labor force.

Can *all* binary outcomes be viewed as manifestations of a latent variable? Some researchers argue that invoking a latent variable is usually inappropriate, others believe that an underlying latent variable is perfectly reasonable in all cases, while most seem to take a middle ground. Regardless of your assessment of the use of a latent variable, it is important to realize that the derivation and application of the BRM is not dependent on your acceptance of the notion of a latent variable. Section 3.4 shows that the same BRM can be derived as a nonlinear probability model, without invoking the idea of a latent variable.

The latent y^* is assumed to be linearly related to the observed x's through the structural model:

$$y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$$

The latent variable y^* is linked to the observed binary variable y by the measurement equation:

$$y_i = \begin{cases} 1 & \text{if } y_i^* > \tau \\ 0 & \text{if } y_i^* \le \tau \end{cases} \qquad [3.1]$$

where τ is the *threshold* or *cutpoint*. If $y^* \le \tau$, then $y = 0$. If y^* crosses the threshold τ (i.e., $y^* > \tau$), then $y = 1$. For now, we assume that $\tau = 0$. Section 5.2 (p. 122) discusses this identifying assumption in detail.

The link between the latent y^* and the observed y is illustrated in Figure 3.2 for the model $y^* = \alpha + \beta x + \varepsilon$. In this figure, y^* is on the vertical axis, with the threshold τ indicated by a horizontal dashed line. The distribution of y^* is shown by the bell-shaped curves which should be thought of as coming out of the figure into a third dimension. When y^* is larger than τ, indicated by the shaded region, we observe $y = 1$.

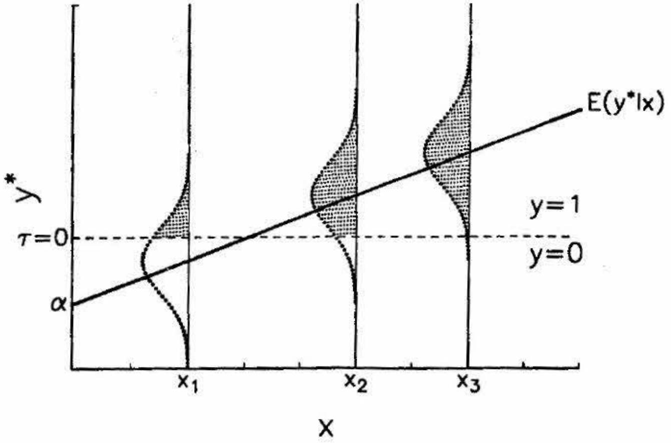


Figure 3.2. The Distribution of y^* Given x in the Binary Response Model

For example, at x_1 about 25% of the y's equal 1, at x_2 nearly 90% are 1's, and at x_3 nearly all cases are 1's.

Since y^* is continuous, the model avoids the problems encountered with the LPM. However, since the dependent variable is unobserved, the model cannot be estimated with OLS. Instead, we use ML estimation, which requires assumptions about the distribution of the errors. Most often, the choice is between normal errors which result in the *probit* model, and logistic errors which result in the *logit* model. As with the LRM, we assume that $E(\varepsilon \mid \mathbf{x}) = 0$.

Since y^* is unobserved, we cannot estimate the variance of the errors as we did with the LRM. In the probit model, we assume that $\text{Var}(\varepsilon \mid \mathbf{x}) = 1$ and in the logit model that $\text{Var}(\varepsilon \mid \mathbf{x}) = \pi^2/3 \approx 3.29$. (The symbol "≈" means "is approximately equal to.") The specific value assumed for the variance is arbitrary in the sense that it cannot be disconfirmed by the data. We choose a value that results in the simplest formula for the distribution of ε.

The logistic and normal distributions are used so frequently for models with CLDVs that it is worth examining these distributions in detail. The probability density functions and cumulative distribution functions for the normal and logistic distributions are shown in Figure 3.3. The normal distribution is drawn with a solid line. When ε is normal with $E(\varepsilon \mid \mathbf{x}) = 0$ and $\text{Var}(\varepsilon \mid \mathbf{x}) = 1$, the pdf is

$$\phi(\varepsilon) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\varepsilon^2}{2}\right)$$

and the cumulative distribution function (hereafter, cdf) is

$$\Phi(\varepsilon) = \int_{-\infty}^{\varepsilon} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

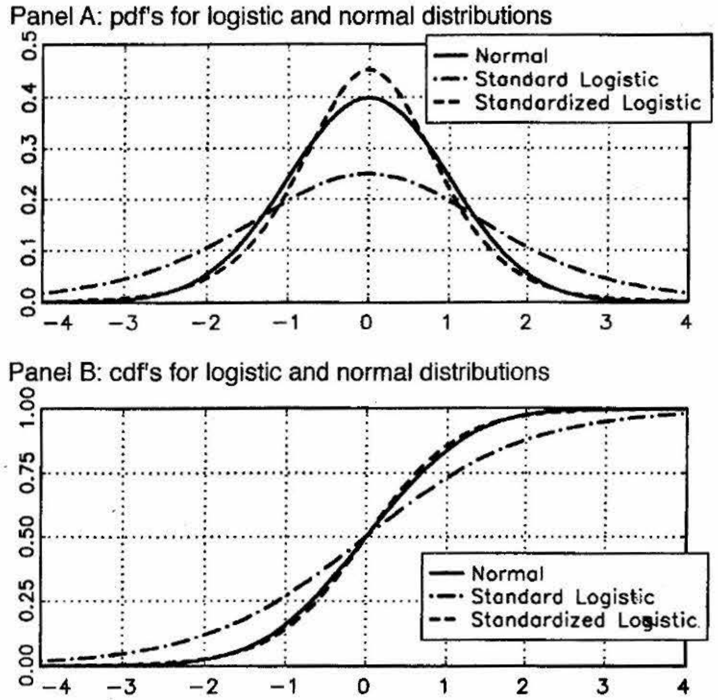
The cdf indicates the probability that a random variable is less than or equal to a given value. For example, $\Phi(0) = \Pr(\varepsilon \leq 0) = .5$. (*Find this point in panel B of Figure 3.3.*)

In the logit model, the errors are assumed to have a *standard logistic distribution* with mean 0 and variance $\pi^2/3$. This unusual variance is chosen because it results in a particularly simple equation for the pdf:

$$\lambda(\varepsilon) = \frac{\exp(\varepsilon)}{[1 + \exp(\varepsilon)]^2}$$

and an even simpler equation for the cdf:

$$\Lambda(\varepsilon) = \frac{\exp(\varepsilon)}{1 + \exp(\varepsilon)}$$


Panel A: pdf's for logistic and normal distributions

Panel B: cdf's for logistic and normal distributions

Figure 3.3. Normal and Logistic Distributions

These distributions are drawn with long dot-dashes in Figure 3.3. The standard logistic pdf is flatter than the normal distribution since it has a larger variance.

If we rescale the logistic distribution to have a unit variance, known as the *standardized* (not standard) logistic distribution, the logistic and normal cdf's are nearly identical, as shown in panel B of Figure 3.3. However, the pdf and cdf for the standardized logistic distribution with a unit variance are more complicated:

$$\lambda^S(\varepsilon) = \frac{\gamma \exp(\gamma\varepsilon)}{[1 + \exp(\gamma\varepsilon)]^2} \quad \text{and} \quad \Lambda^S(\varepsilon) = \frac{\exp(\gamma\varepsilon)}{1 + \exp(\gamma\varepsilon)} \qquad [3.2]$$

where $\gamma = \pi/\sqrt{3}$. Because of the simpler equations for the standard (not standard*ized*) logistic distribution, it is generally used for deriving

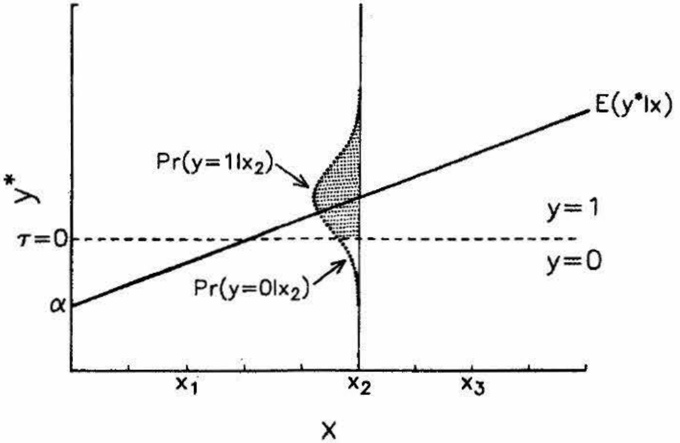


Figure 3.4. Probability of Observed Values in the Binary Response Model

the logit model. The consequences of assuming different variances for the probit and logit models are considered in Section 3.3.

By assuming a specific form for the distribution of ε, it is possible to compute the probability of $y = 1$ for a given $\mathbf{x}$. To see this, consider Figure 3.4, where ε is distributed either logistically or normally around $E(y^* \mid x) = \alpha + \beta x$. Values of $y = 1$ are observed for the shaded portion of the error distribution above τ. Even if $E(y^* \mid x)$ is in the shaded region where $y = 1$ (e.g., at x_2), it is possible to observe a 0 if ε is large and negative. The negative error moves y^* into the unshaded region of the curve.

Figure 3.5 illustrates the translation of these ideas into a formula for computing $\Pr(y = 1 \mid \mathbf{x})$. Panel A takes the error distribution from Figure 3.4 and places it on its side. Since $y = 1$ when $y^* > 0$,

$$\Pr(y = 1 \mid \mathbf{x}) = \Pr(y^* > 0 \mid \mathbf{x})$$

Substituting $y^* = \mathbf{x}\boldsymbol{\beta} + \varepsilon$, it follows that

$$\Pr(y = 1 \mid \mathbf{x}) = \Pr(\mathbf{x}\boldsymbol{\beta} + \varepsilon > 0 \mid \mathbf{x})$$

Subtracting $\mathbf{x}\boldsymbol{\beta}$ from each side of the inequality corresponds to shifting the x-axis as shown in panel B. Then

$$\Pr(y = 1 \mid \mathbf{x}) = \Pr(\varepsilon > -\mathbf{x}\boldsymbol{\beta} \mid \mathbf{x})$$

Since cdf's express the probability of a variable being less than some value, we must change the direction of the inequality. The normal and logistic distributions are symmetric, which means that the shaded area of the distribution greater than $-\mathbf{x}\boldsymbol{\beta}$ in panel B equals the shaded area less than $\mathbf{x}\boldsymbol{\beta}$ in panel C. Consequently,

$$\Pr(y = 1 \mid \mathbf{x}) = \Pr(\varepsilon \leq \mathbf{x}\boldsymbol{\beta} \mid \mathbf{x})$$

This is simply the cdf of the error distribution evaluated at $\mathbf{x}\boldsymbol{\beta}$. Accordingly,

$$\Pr(y = 1 \mid \mathbf{x}) = F(\mathbf{x}\boldsymbol{\beta}) \qquad [3.3]$$

where F is the normal cdf Φ for the probit model and the logistic cdf Λ for the logit model. The probability of observing an event given $\mathbf{x}$ is the cumulative density evaluated at $\mathbf{x}\boldsymbol{\beta}$.

To understand the functional form of the resulting model, consider the BRM for a single independent variable:

$$\Pr(y = 1 \mid x) = F(\alpha + \beta x) \qquad [3.4]$$

As x increases by one unit, the argument of F increases by β units. Plotting Equation 3.4 corresponds to plotting the cdf of either the normal or the logistic distribution as its argument increases. This is shown in Figure 3.6. Panel A illustrates the error distribution for nine values of x. The region of the distribution where $y^* > \tau$ corresponds to $\Pr(y = 1 \mid x)$ and has been shaded. Panel B plots $\Pr(y = 1 \mid x)$. At x_1, only a small portion of the tail of the curve crosses the threshold in panel A, resulting in a small value of $\Pr(y = 1 \mid x)$ in panel B. As we move to x_2, the error distribution shifts up slightly. (*This shift is exactly $\beta(x_2 - x_1)$. Why?*

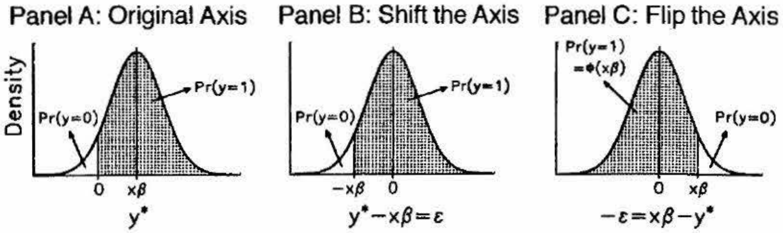


Figure 3.5. Computing $\Pr(y = 1 \mid x)$ in the Binary Response Model

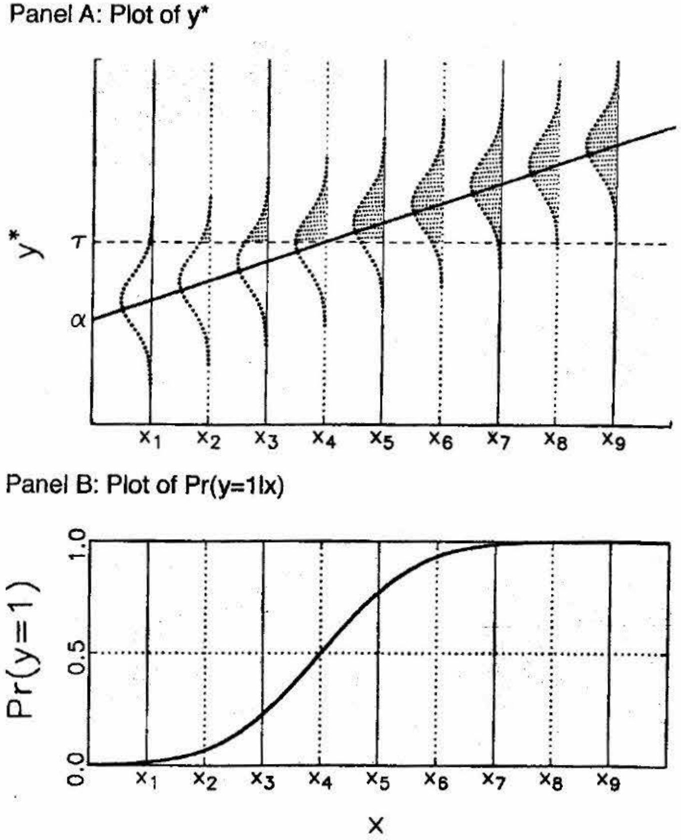


Figure 3.6. Plot of y^* and $\Pr(y = 1 \mid \mathbf{x})$ in the Binary Response Model

What is the amount of the change in the probability?) Since only a small portion of the thin tail moves over the threshold, $\Pr(y = 1 \mid x)$ increases only slightly as shown in panel B. As we continue to move to the right, from x_2 to x_3 to x_4, thicker regions of the error distribution slide over the threshold and the increase in $\Pr(y = 1 \mid x)$ becomes larger. After x_4, increasingly thinner sections of the distribution cross the threshold and the value of $\Pr(y = 1 \mid x)$ increases increasingly more slowly as it approaches 1. The resulting curve is the well-known S-curve associated with the BRM.

Before considering the interpretation of the parameters and how they are related to the predicted probability of an event, we must consider the issue of identification.

3.3. Identification

In specifying the BRM, we made three identifying assumptions: (1) the threshold is 0: $\tau = 0$; (2) the conditional mean of ε is 0: $E(\varepsilon \mid \mathbf{x}) = 0$; and (3) the conditional variance of ε is a constant: $\text{Var}(\varepsilon \mid \mathbf{x}) = 1$ in the probit model and $\text{Var}(\varepsilon \mid \mathbf{x}) = \pi^2/3$ in the logit model. These assumptions are *arbitrary* in the sense that they cannot be tested, but they are *necessary* to identify the model. Identification is an issue that is essential for understanding models with latent variables. Since a latent variable is unobserved, its mean and variance cannot be estimated. For example, in the covariance structure model, commonly referred to as the LISREL model, the variance of a latent variable is unidentified. Assumptions are required to fix the variance to a constant or to link the latent variable to an observed variable (Bollen, 1989, pp. 238–246; Long, 1983, pp. 49–52). In the BRM, the model is not identified until we impose assumptions that determine the mean and variance of y^*.

To see the relationship between the variance of the dependent variable and the identification of the β's in a regression model, consider the model $y = \mathbf{x}\boldsymbol{\beta}_y + \varepsilon_y$, where y is observed. Construct a new dependent variable $w = \delta y$, where δ is any nonzero constant. The variance of w equals:

$$\text{Var}(w) = \text{Var}(\delta y) = \delta^2\,\text{Var}(y)$$

For example, if $\delta = 1/\sqrt{\text{Var}(y)}$, then $\text{Var}(w) = 1$. Since $w = \delta y$ and $y = \mathbf{x}\boldsymbol{\beta}_y + \varepsilon_y$, it follows that

$$w = \delta(\mathbf{x}\boldsymbol{\beta}_y + \varepsilon_y) = \mathbf{x}(\delta\boldsymbol{\beta}_y) + \delta\varepsilon_y$$

Therefore, the β's in a regression of w on $\mathbf{x}$ are δ times the β's in the regression of y on $\mathbf{x}$. That is,

$$\boldsymbol{\beta}_w = \delta\boldsymbol{\beta}_y \qquad [3.5]$$

Since the magnitude of the slope depends on the scale of the dependent variable, if we do not know the variance of the dependent variable, then the slope coefficients are not identified.

To apply this result to the BRM and to understand the relationship between the magnitudes of the logit compared to the probit coefficients, we need to distinguish between the structural models for logit and probit. Let

$$y_L^* = \mathbf{x}\boldsymbol{\beta}_L + \varepsilon_L \quad \text{and} \quad y_P^* = \mathbf{x}\boldsymbol{\beta}_P + \varepsilon_P$$

where L indicates the logit model and P the probit model. Since y_L^* and y_P^* are latent, it is impossible to determine their variances from the observed data, and, consequently, $\boldsymbol{\beta}_L$ and $\boldsymbol{\beta}_P$ are unidentified. For both models, the variance of y^* is determined by assuming the variance of ε. Since $\text{Var}(\varepsilon_L \mid \mathbf{x}) = (\pi^2/3)\,\text{Var}(\varepsilon_P \mid \mathbf{x})$ *(Why?)*, it follows that $\varepsilon_L \approx (\pi/\sqrt{3})\varepsilon_P$. The errors are not identical since the logistic and normal distributions with unit variance are only approximately equal (see Figure 3.3). From Equation 3.5,

$$\boldsymbol{\beta}_L \approx \sqrt{\text{Var}(\varepsilon_L \mid \mathbf{x})}\,\boldsymbol{\beta}_P \approx \sqrt{\pi^2/3}\,\boldsymbol{\beta}_P \approx 1.81\boldsymbol{\beta}_P$$

where $\sqrt{\pi^2/3} \approx 1.81$. This transformation can be used to compare coefficients from a published logit analysis to comparable coefficients from a probit analysis and vice versa.

The approximation $\boldsymbol{\beta}_L \approx 1.8\,\boldsymbol{\beta}_P$ is based on equating the variances of the logistic and normal distributions. Amemiya (1981) suggested making the cdf's of the logistic and normal distributions as close as possible, not just making their variances equal. He proposed that the cdf's were most similar when $\varepsilon_L \approx 1.6\varepsilon_P$, which led to his approximation: $\boldsymbol{\beta}_L \approx 1.6\,\boldsymbol{\beta}_P$. My own calculations indicate that the cdf's are closest when $\varepsilon_L \approx 1.7\varepsilon_P$, which, conveniently, corresponds to the results in the example I now present.

Example of Logit and Probit: Labor Force Participation

Even though we have not considered estimation, it is useful to examine the logit and probit estimates from our model of labor force participation. The model is

$$\Pr(LFP = 1) = F(\beta_0 + \beta_1 K5 + \beta_2 K618 + \beta_3 AGE \\ + \beta_4 WC + \beta_5 HC + \beta_6 LWG + \beta_7 INC)$$

Estimates are given in Table 3.3. The first thing to notice is that the log likelihood and z-tests are nearly identical. This reflects the basic similarity, except for scaling, in the structure of the logit and probit models, and the fact that these statistics are unaffected by the assumed variance of the error. The effects of the identifying assumptions about $\text{Var}(\varepsilon)$ are seen by taking the ratio of the logit coefficients to the probit coefficients, contained in the column labeled "Ratio." The logit coefficients are about 1.7 times larger than the corresponding probit coefficients, with the exception of the coefficient for *HC* which is the least statistically significant parameter. Clearly, interpretation of the β's must take the effects of the identifying assumptions into account. This issue is now considered.

TABLE 3.3 Logit and Probit Analyses of Labor Force Participation

	Logit		*Probit*		*Ratio*	
Variable	β	z	β	z	β	z
Constant	3.182	4.94	1.918	5.04	1.66	0.98
K5	−1.463	−7.43	−0.875	−7.70	1.67	0.96
K618	−0.065	−0.95	−0.039	−0.95	1.67	1.00
AGE	−0.063	−4.92	−0.038	−4.97	1.66	0.99
WC	0.807	3.51	0.488	3.60	1.65	0.97
HC	0.112	0.54	0.057	0.46	1.95	1.18
LWG	0.605	4.01	0.366	4.17	1.65	0.96
INC	−0.034	−4.20	−0.021	−4.30	1.68	0.98
$-2\ln L$	905.27		905.39		1.00	

NOTE: $N = 753$. β is an unstandardized coefficient; z is the z-test for β. "Ratio" is the ratio of a logit to a probit coefficient.

3.3.1. The Identification of Probabilities

Since the β's are unidentified without assumptions about the mean and variance of ε, the β's are arbitrary in this sense: if we change the identifying assumption regarding $\text{Var}(\varepsilon \mid \mathbf{x})$, the β's also change. *Accordingly, the β's cannot be interpreted directly since they reflect both*: (1) *the relationship between the x's and y^*; and* (2) *the identifying assumptions*. While the identifying assumptions affect the β's, they do not affect $\Pr(y = 1 \mid \mathbf{x})$. More technically, $\Pr(y = 1 \mid \mathbf{x})$ is an *estimable function*. An estimable function is a function of the parameters that is invariant to the identifying assumptions (Searle, 1971, pp. 180–188).

Consider the logit model where

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i\boldsymbol{\beta})} = \frac{1}{1 + \exp(-\mathbf{x}_i\boldsymbol{\beta})}$$

(*Prove the last equality.*) The right-hand side is the cdf for the logistic distribution with variance $\sigma^2 = \pi^2/3$. We can standardize ε to have a unit variance by dividing the structural model by σ:

$$\frac{y_i^*}{\sigma} = \frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma} + \frac{\varepsilon_i}{\sigma}$$

ε/σ has a standardized logistic distribution with cdf (see Equation 3.2):

$$\Lambda^S\left(\frac{\varepsilon_i}{\sigma}\right) = \frac{\exp\left(\frac{\pi}{\sqrt{3}}\frac{\varepsilon_i}{\sigma}\right)}{1+\exp\left(\frac{\pi}{\sqrt{3}}\frac{\varepsilon_i}{\sigma}\right)}$$

Since $\sigma = \pi/\sqrt{3}$,

$$\Lambda^S\left(\frac{\varepsilon_i}{\sigma}\right) = \frac{\exp(\varepsilon_i)}{1+\exp(\varepsilon_i)} = \Lambda(\varepsilon_i)$$

Consequently, *the probability of an event is unaffected by the identifying assumption regarding* $\text{Var}(\varepsilon\,|\,\mathbf{x})$. While the specific value assumed for $\text{Var}(\varepsilon\,|\,\mathbf{x})$ is arbitrary and affects the β's, it does not affect the quantity that is of fundamental interest, namely, the probability that an event occurred. The same result holds for the probit model.

The critical point is that while the β's are affected by the arbitrary scale assumed for ε, the probabilities are not affected. Consequently, the probabilities can be interpreted without concern about the arbitrary assumption that is made to identify the model. That is to say, the probabilities are estimable functions. Further, any function of the probabilities is also estimable. Importantly, we can interpret changes in probabilities and odds, which are ratios of probabilities. This is done in Section 3.7, but first we consider an alternative method of deriving the logit and probit models.

3.4. A Nonlinear Probability Model

The BRM can also be derived without appealing to an underlying latent variable. This is done by specifying a nonlinear model relating the x's to the probability of an event. For example, Aldrich and Nelson (1984, pp. 31–32) derive the logit model by starting with the problem that the LPM can predict values of $\Pr(y=1\,|\,\mathbf{x})$ that are greater than 1 or less than 0. To eliminate this problem, they transform $\Pr(y=1\,|\,\mathbf{x})$ into a function that ranges from $-\infty$ to ∞. First, the probability is transformed into the *odds*:

$$\frac{\Pr(y=1\,|\,\mathbf{x})}{\Pr(y=0\,|\,\mathbf{x})} = \frac{\Pr(y=1\,|\,\mathbf{x})}{1-\Pr(y=1\,|\,\mathbf{x})}$$

The odds indicate how often something (e.g., $y = 1$) happens relative to how often it does not happen (e.g., $y = 0$), and range from 0 when $\Pr(y=1\,|\,\mathbf{x}) = 0$ to ∞ when $\Pr(y=1\,|\,\mathbf{x}) = 1$. The log of the odds, known as the *logit*, ranges from $-\infty$ to ∞. This suggests a model that is linear in the logit:

$$\ln\left[\frac{\Pr(y=1\,|\,\mathbf{x})}{1-\Pr(y=1\,|\,\mathbf{x})}\right] = \mathbf{x}\boldsymbol{\beta} \qquad [3.6]$$

This is equivalent to the logit model derived above (*Show this.*):

$$\Pr(y=1\,|\,\mathbf{x}) = \frac{\exp(\mathbf{x}\boldsymbol{\beta})}{1+\exp(\mathbf{x}\boldsymbol{\beta})} \qquad [3.7]$$

Other probability models can be constructed by choosing functions of $\mathbf{x}\boldsymbol{\beta}$ that range from 0 to 1. Cumulative distribution functions have this property and readily provide a number of examples. The cdf for the standard normal distribution results in the probit model:

$$\Pr(y=1\,|\,\mathbf{x}) = \int_{-\infty}^{\mathbf{x}\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)dt = \Phi(\mathbf{x}\boldsymbol{\beta})$$

Another example is the complementary log-log model (Agresti, 1990, pp. 104–107; McCullagh & Nelder, 1989, p. 108), defined by

$$\ln(-\ln[1-\Pr(y=1\,|\,\mathbf{x})]) = \mathbf{x}\boldsymbol{\beta}$$

or, equivalently,

$$\Pr(y=1\,|\,\mathbf{x}) = 1-\exp[-\exp(\mathbf{x}\boldsymbol{\beta})]$$

Unlike the logit and probit models, the complementary log-log model is asymmetric. In the logit and probit models, if you are at that point on the probability curve where $\Pr(y=1\,|\,x) = .5$, increasing x by a given amount δ changes the probability by the same amount as if x is decreased by δ. This is not the case for the complementary log-log model as shown in Figure 3.7. As x increases, the probability increases slowly at the left until it reaches about .2; the change from .8 toward 1 occurs much more rapidly. The log-log model, which is defined as

$$\Pr(y=1\,|\,\mathbf{x}) = \exp[-\exp(-\mathbf{x}\boldsymbol{\beta})]$$

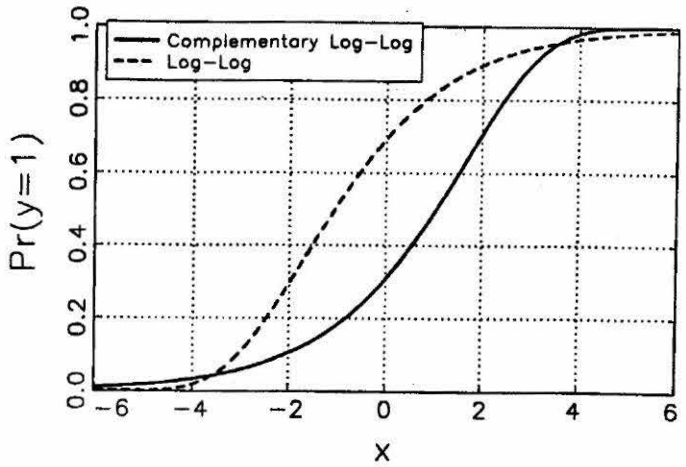


Figure 3.7. Complementary Log-Log and Log-Log Models

has the opposite pattern. These models can be estimated with GLIM, Stata, and SAS, and have links to the proportional hazards model (see Allison, 1995, pp. 216–217, or Petersen, 1995, p. 499, for details).

3.5. ML Estimation[2]

To specify the likelihood equation, define p as the probability of observing whatever value of y was actually observed for a given observation:

$$p_i = \begin{cases} \Pr(y_i = 1 \mid \mathbf{x}_i) & \text{if } y_i = 1 \text{ is observed} \\ 1 - \Pr(y_i = 1 \mid \mathbf{x}_i) & \text{if } y_i = 0 \text{ is observed} \end{cases} \qquad [3.8]$$

$\Pr(y_i = 1 \mid \mathbf{x}_i)$ is defined by Equation 3.3. If the observations are independent, the likelihood equation is

$$L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} p_i \qquad [3.9]$$

[2] When there is more than one observation for each combination of values of independent variables, Berkson's minimum chi-square estimation can be used. Since the requirement of many observations per cell is rarely satisfied in social science research, I do not consider this method. See Hanushek and Jackson (1977, pp. 190–200) or Maddala (1983, pp. 28–34).

Combining Equations 3.8 and 3.9,

$$L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \prod_{y=1} \Pr(y_i = 1 \mid \mathbf{x}_i) \prod_{y=0} [1 - \Pr(y_i = 1 \mid \mathbf{x}_i)]$$

where the index for multiplication indicates that the product is taken over only those cases where $y = 1$ and $y = 0$, respectively.

The β's are incorporated into the likelihood equation by substituting the right-hand side of Equation 3.3:

$$L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \prod_{y=1} F(\mathbf{x}_i\boldsymbol{\beta}) \prod_{y=0} [1 - F(\mathbf{x}_i\boldsymbol{\beta})]$$

Taking logs, we obtain the log likelihood equation:

$$\ln L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \sum_{y=1} \ln F(\mathbf{x}_i\boldsymbol{\beta}) + \sum_{y=0} \ln[1 - F(\mathbf{x}_i\boldsymbol{\beta})]$$

Amemiya (1985, pp. 273–274) proves that under conditions that are likely to apply in practice, the likelihood function is globally concave which ensures the uniqueness of the ML estimates. These estimates are consistent, asymptotically normal, and asymptotically efficient.

3.5.1. Maximum Likelihood and Sample Size

For ML estimation, the desirable properties of consistency, normality, and efficiency are asymptotic. This means that these properties have been proven to hold as the sample size approaches ∞. While ML estimators are not necessarily bad estimators in small samples, indeed OLS for the linear regression model is an ML estimator that works quite well in small-samples, the small-sample behavior of ML estimators for the models in this book is largely unknown. Since alternative estimators with known small sample properties are generally not available for the models we consider, the practical question is: *When is the sample large enough to use the ML estimates and the resulting significance tests*? While I am reluctant to give advice without firm evidence to justify the advice, it seems necessary to add a cautionary note since it is easy to get the impression that ML estimation works well with any sample size. For example, the 32 observations from a study by Spector and Mazzeo (1980) are used frequently to illustrate the logit and probit models, yet 32 is too small of a sample to justify the use of ML. The following guidelines are not hard and fast. They are based on my experience of when the models seem to produce reasonable and robust results and my discussions with other researchers who use these methods.

It is risky to use ML with samples smaller than 100, while samples over 500 seem adequate. These values should be raised depending on characteristics of the model and the data. First, if there are a lot of parameters in the model, more observations are needed. In the literature on the covariance structure model, the rule of at least five observations per parameter is often given. A rule of at least 10 observations per parameter seems reasonable for the models in this book. This rule does not imply that a minimum of 100 is not needed if you have only two parameters. Second, if the data are ill conditioned (e.g., independent variables are highly collinear) or if there is little variation in the dependent variable (e.g., nearly all of the outcomes are 1), a larger sample is required. Third, some models seem to require more observations. The ordinal regression model of Chapter 5 is an example. In discussing the use of ML for small samples, Allison (1995, p. 80) makes a useful point. While the standard advice is that with small samples you should accept larger p-values as evidence against the null hypothesis, given that the degree to which ML estimates are normally distributed in small samples is unknown, it is more reasonable to require smaller p-values in small samples.

3.6. Numerical Methods for ML Estimation

For the LRM, ML estimates are obtained by setting the gradient of the log likelihood to 0 and solving for the parameters using algebra. Algebraic solutions are rarely possible with nonlinear models. Consequently, *numerical methods* are used to find the estimates that maximize the log likelihood function. Numerical methods start with a guess of the values of the parameters and iterate to improve on that guess. While you may be tempted to dismiss numerical methods as an esoteric topic of little practical concern, programs using numerical methods for estimation can produce incorrect estimates or fail to provide any estimates. To recognize and correct such problems, an elementary understanding of numerical methods is useful. I begin with an introduction to numerical methods, followed by practical advice on using these methods.

3.6.1. Iterative Solutions

Assume that we are trying to estimate the vector of parameters $\boldsymbol{\theta}$. We begin with an initial guess $\boldsymbol{\theta}_0$, called *start values*, and attempt to improve on this guess by adding a vector $\boldsymbol{\zeta}_0$ of adjustments:

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \boldsymbol{\zeta}_0$$

We proceed by updating the previous iteration according to the equation:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \boldsymbol{\zeta}_n$$

Iterations continue until there is *convergence*. Roughly, convergence occurs when the gradient of the log likelihood is close to 0 or the estimates do not change from one step to the next. Convergence must occur to obtain the ML estimator $\widehat{\boldsymbol{\theta}}$.

The problem is to find a $\boldsymbol{\zeta}_n$ that moves the process rapidly toward a solution. It is useful to think of $\boldsymbol{\zeta}_n$ as consisting of two parts: $\boldsymbol{\zeta}_n = \mathbf{D}_n\boldsymbol{\gamma}_n$. $\boldsymbol{\gamma}_n$ is the *gradient* vector defined as $\partial \ln L/\partial\boldsymbol{\theta}_n$, which indicates the direction of the change in the log likelihood for a change in the parameters. $\mathbf{D}_n$ is a *direction matrix* that reflects the curvature of the log likelihood function; that is, it indicates how rapidly the gradient is changing. A clearer understanding of these components is gained by examining the simplest methods of maximization.

The Method of Steepest Ascent. The method of steepest ascent lets $\mathbf{D}=\mathbf{I}$:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \frac{\partial \ln L}{\partial \boldsymbol{\theta}_n}$$

An estimate increases if the gradient is positive, and it decreases if the gradient is negative. Iterations stop when the derivative becomes nearly 0. The problem with this approach is that it considers the slope of $\ln L$, but not how quickly the slope is changing. To see why this is a problem, consider two log likelihood functions with the same gradient at a given point but with one function changing shape more quickly than the other. (*Sketch these functions.*) You should move more gradually for the function that is changing quickly, in order to avoid moving too far. Steepest descent tends to work poorly since it treats both functions in the same way.

The next three commonly used methods address this problem by adding a direction matrix that assesses how quickly the log likelihood function is changing. They differ in their choice of a direction matrix. In all cases, it takes longer to compute the direction matrix than the identity matrix used with the method of steepest ascent. Usually, the additional computational costs are made up for by the fewer iterations that are required to reach convergence.

No one method works best all of the time. An algorithm applied to one set of data may not converge, while another algorithm applied to the same data may converge rapidly. For a different set of data, the opposite may occur. In general, the algorithm used in commercial software depends on the preferences of the programmer and the ease with which an algorithm can be programmed for a given model.

The Newton-Raphson Method. The rate of change in the slope of $\ln L$ is indicated by the second derivatives, which are contained in the *Hessian matrix* $\partial^2 \ln L/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$. For example, with two parameters $\boldsymbol{\theta} = (\alpha\ \beta)'$, the Hessian is

$$\frac{\partial^2 \ln L}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \begin{pmatrix} \dfrac{\partial^2 \ln L}{\partial\alpha\partial\alpha} & \dfrac{\partial^2 \ln L}{\partial\alpha\partial\beta} \\ \dfrac{\partial^2 \ln L}{\partial\beta\partial\alpha} & \dfrac{\partial^2 \ln L}{\partial\beta\partial\beta} \end{pmatrix}$$

If $\partial^2 \ln L/\partial\alpha\partial\alpha$ is large relative to $\partial^2 \ln L/\partial\beta\partial\beta$, the gradient is changing more rapidly as α changes than as β changes. Thus, smaller a djustments to the estimate of α would be indicated. The Newton-Raphson algorithm proceeds according to the equation:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \left(\frac{\partial^2 \ln L}{\partial\boldsymbol{\theta}_n\partial\boldsymbol{\theta}_n'}\right)^{-1} \frac{\partial \ln L}{\partial\boldsymbol{\theta}_n}$$

(Why are we taking the inverse of the Hessian?)

The Method of Scoring. In some cases, the expectation of the Hessian, known as the *information matrix*, can be easier to compute than the Hessian. The method of scoring uses the information matrix as the direction matrix, which results in

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \left(E\left[\frac{\partial^2 \ln L}{\partial\boldsymbol{\theta}_n\partial\boldsymbol{\theta}_n'}\right]\right)^{-1} \frac{\partial \ln L}{\partial\boldsymbol{\theta}_n}$$

The BHHH Method. When the Hessian and the information matrix are difficult to compute, Berndt et al. (1974) propose using an outer product of the gradient approximation to the information matrix:

$$\sum_{i=1}^{N} \frac{\partial \ln L_i}{\partial\boldsymbol{\theta}_n} \frac{\partial \ln L_i}{\partial\boldsymbol{\theta}_n}'$$

where $\ln L_i$ is the value of the likelihood function evaluated for the ith observation. This approximation is often simpler to compute since only the gradient needs to be evaluated. Iterations proceed according to

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \left(\sum_{i=1}^{N} \frac{\partial \ln L_i}{\partial\boldsymbol{\theta}_n} \frac{\partial \ln L_i}{\partial\boldsymbol{\theta}_n}'\right)^{-1} \frac{\partial \ln L}{\partial\boldsymbol{\theta}_n}$$

which is known as the BHHH (pronounced "B-triple-H") algorithm or the modified method of scoring.

Numerical Derivatives. If you cannot obtain an algebraic solution for the gradient or the Hessian, numerical methods can be used to estimate them. For example, consider a log likelihood based on a single parameter θ. The gradient is approximated by computing the slope of the change in $\ln L$ when θ changes by a small amount. If Δ is a small number relative to θ,

$$\frac{\partial \ln L}{\partial\theta} \approx \frac{\ln L(\theta+\Delta) - \ln L(\theta)}{\Delta}$$

Using numerical estimates can greatly increase the time and number of iterations needed, and results can be sensitive to the choice of Δ. Further, different start values can result in different estimates of the Hessian at convergence, which translates into different estimates of the standard errors. Programs that use numerical methods for computing derivatives should only be used if no alternatives are available. When they must be used, you should experiment with different starting values to make sure that the estimates that you obtain are stable.

3.6.2. The Variance of the ML Estimator

In addition to estimating the parameters $\boldsymbol{\theta}$, numerical methods provide estimates of the asymptotic covariance matrix $\mathrm{Var}(\widehat{\boldsymbol{\theta}})$, which are used for the statistical tests in Chapter 4. The theory of maximum likelihood shows that if the assumptions justifying ML estimation hold, then the asymptotic covariance matrix equals

$$\mathrm{Var}(\widehat{\boldsymbol{\theta}}) = \left(-E\left[\frac{\partial^2 \ln L}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\right)^{-1} \qquad [3.10]$$

In words, the asymptotic covariance equals the inverse of the negative of the expected value of the Hessian, known as the *information matrix*.

The covariance matrix is often written in an equivalent form using the outer product of the gradient:

$$\mathrm{Var}(\widehat{\boldsymbol{\theta}}) = \left(E\left[\frac{\partial \ln L}{\partial \boldsymbol{\theta}} \frac{\partial \ln L'}{\partial \boldsymbol{\theta}}\right]\right)^{-1} \qquad [3.11]$$

In both cases, the expression is evaluated at $\boldsymbol{\theta}$. Since we only have an estimate of $\boldsymbol{\theta}$, the covariance matrix must be estimated. Three consistent estimators of $\mathrm{Var}(\widehat{\boldsymbol{\theta}})$ are commonly used.

The first estimator evaluates Equation 3.10 using the ML estimates $\widehat{\boldsymbol{\theta}}$:

$$\widehat{\mathrm{Var}}_1(\widehat{\boldsymbol{\theta}}) = -\left(E\left[\frac{\partial^2 \ln L}{\partial \widehat{\boldsymbol{\theta}} \partial \widehat{\boldsymbol{\theta}}'}\right]\right)^{-1}$$

This estimator is generally used with the method of scoring since that method requires evaluating the information matrix at each iteration.

A second estimator is obtained by evaluating the negative of the Hessian, sometimes referred to as the observed information matrix, rather than the information matrix itself:

$$\widehat{\mathrm{Var}}_2(\widehat{\boldsymbol{\theta}}) = -\left(\sum_{i=1}^{N} \frac{\partial^2 \ln L_i}{\partial \widehat{\boldsymbol{\theta}} \partial \widehat{\boldsymbol{\theta}}'}\right)^{-1} \qquad [3.12]$$

$\widehat{\mathrm{Var}}_2(\widehat{\boldsymbol{\theta}})$ is generally used with the Newton-Raphson algorithm. Equation 3.12 shows the relationship between the curvature of the likelihood function and the variance of the estimator. The size of the variance is inversely related to the second derivative: the smaller the second derivative, the larger the variance. When the second derivative is smaller, the likelihood function is flatter. If the likelihood equation is very flat, the variance will be large. This should match your intuition that the flatter the likelihood function, the harder it will be to find the maximum of the function, and the less confidence (i.e., the more variance) you should have in the solution you obtain.

A third estimator, which is related to the BHHH algorithm, is simple to compute since it does not require evaluation of the second derivatives:

$$\widehat{\mathrm{Var}}_3(\widehat{\boldsymbol{\theta}}) = \left(\sum_{i=1}^{N} \frac{\partial \ln L_i}{\partial \widehat{\boldsymbol{\theta}}} \frac{\partial \ln L_i}{\partial \widehat{\boldsymbol{\theta}}'}\right)^{-1}$$

While these estimators of the covariance matrix are asymptotically equivalent, in practice they sometimes provide very different estimates, especially when the sample is small or the data are ill conditioned. Consequently, if you estimate the same model with the same data using two programs that use different estimators, you can get different results.

3.6.3. Problems With Numerical Methods and Possible Solutions

While numerical methods generally work well, there can be problems. First, it may be difficult or impossible to reach convergence. You might get an error such as "Convergence not obtained after 250 iterations." Or, it might not be possible to invert the Hessian when $\ln L$ is nearly flat. This generates a message such as "Singularity encountered," "Hessian could not be inverted," or "Hessian was not of full rank." The message might refer to the covariance matrix or the information matrix. Second, sometimes convergence occurs, but the wrong solution is obtained. This occurs when $\ln L$ has more than one location where the gradient is 0. The iterative process might locate a saddle point or local maximum, where the gradient is also 0, rather than the global maximum. (Think of a two-humped Bactrian camel. The top of the smaller hump is a local maximum; the low spot between the two humps is a saddle point.) In such cases, the covariance matrix which should be positive definite is negative definite. When $\ln L$ is globally concave, there is only one solution, and that is a maximum. This is the case for most of the models considered in this book. However, even when the log likelihood is globally concave, it is possible to have false convergence. This can occur when the function is very flat and the precision of the estimates of the gradient is insufficient. This is common when numerical gradients are used and can also be caused by problems with scaling (discussed below). Finally, in some cases, ML estimates do not exist for a particular pattern of data. For example, with a binary outcome and a single binary independent variable, ML estimates are not possible if there is no variation in the independent variable for one of the outcomes. You can try estimating a probit model using: $\mathbf{y}' = (0\ 0\ 1\ 1\ 1)$ and $\mathbf{x}' = (1\ 0\ 1\ 1\ 0)$. This works fine, since there are x's equal to 0 and 1 for both $y = 1$ and $y = 0$. However, now try to estimate the model for: $\mathbf{y}' = (0\ 0\ 1\ 1)$ and $\mathbf{x}' = (1\ 0\ 1\ 1)$. Your program will "crash" since whenever $y = 1$, all x's are 1's.

When you cannot get a solution or appear to get the wrong solution, the first thing to check is that the software is estimating the model that you want to estimate. It is easy to make an error in specifying the commands to estimate your model. If the model and commands are correct, there may be problems with the data.

Incorrect variables. Most simply, you may have constructed a variable incorrectly. Be sure to check the descriptive statistics for all variables. My experience suggests that most problems with numerical methods are due to data that have not been "cleaned."

Number of observations. Convergence generally occurs more rapidly when there are more observations, and when the ratio of the number of observations to the number of variables is larger. While there is generally little you can do about sample size, it can explain why you are having problems getting your models to converge.

Scaling of variables. Scaling is a very common cause of problems with numerical methods. The larger the ratio between the largest standard deviation and the smallest standard deviation, the more problems you will have with numerical methods. For example, if you have income measured in dollars, it may have a very large standard deviation relative to other variables. Recoding income to thousands of dollars, may solve the problem. My experience suggests that problems are much more likely when the ratio between the largest and smallest standard deviation exceeds 10.

Distribution of the outcome. If a large proportion of cases are censored in the tobit model or if one of the categories of a categorical variable has very few cases, convergence may be difficult. There is little that can be done with such data limitations.

Numerical methods for ML estimation tend to work well when your model is appropriate for your data. In such cases, convergence generally occurs quite rapidly, often within five iterations. If you have too few cases, too many variables, or a poor model, convergence may be a problem. In such cases, rescaling your data can solve the problem. If that does not work, you can try using a program that uses a different numerical algorithm. A problem that may be very difficult for one algorithm may work quite well for another.

While numerical methods generally work well, I heartily endorse Cramer's (1986, p. 10) advice: "Check the data, check their transfer into the computer, check the actual computations (preferably by repeating at least a sample by a rival program), and always remain suspicious of the results, regardless of the appeal."

3.6.4. Software Issues

There are several issues related to software for logit and probit that should be considered.

The Method of Numerical Maximization. Different programs use different methods of numerical maximization. In most cases, estimates of the parameters from the different programs are identical to at least four decimal digits. Estimates of the standard errors and the z-values may differ at the first decimal digit as a result of the different methods used to estimate $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$.

Parameterizations of the Model. A more basic difference is found in the outcome being modeled. While most programs model the probability of a 1, some programs (e.g., SAS) model the probability of a 0. This is a trivial difference *if* you are aware of what the program is doing. For the BRM,

$$\Pr(y_i = 0 \mid \mathbf{x}_i) = 1 - \Pr(y_i = 1 \mid \mathbf{x}_i) = 1 - F(\mathbf{x}_i'\boldsymbol{\beta}) = F(-\mathbf{x}_i'\boldsymbol{\beta})$$

where the last equality follows from the symmetry of the pdf for the logit and probit models. Thus, all coefficients will have the opposite sign. Note that this will *not* be the case for the complementary log-log model since it is asymmetric.

With estimates in hand, we can consider the interpretation of the binary response model.

3.7. Interpretation

In this section, I present four methods of interpretation, each of which is generalized to other models in later chapters. First, I show how to present predicted probabilities using graphs and tables. Second, I examine the partial change in y^* and in the probability. Third, I use discrete change in the probability to summarize the effects of each variable. Finally, for the logit model, I derive a simple transformation of the parameters that indicates the effect of a variable on the odds that the event occurred.

Since the BRM is nonlinear, no single approach to interpretation can fully describe the relationship between a variable and the outcome probability. You should search for an elegant and concise way to summarize the results that does justice to the complexities of the nonlinear model. For any given application, you may need to try each method before a final approach is determined. For example, you might have to construct a plot of the predicted probabilities before realizing that a single measure of discrete change is sufficient to summarize the effect of a variable. I illustrate these methods with the data on the labor force participation of women. You should be able to replicate many of the results using Tables 3.1 and 3.3, although your answers may differ slightly due to rounding error.

I begin by showing how the intercept and the slope affect the curve relating an independent variable to the probability of an event. Understanding how the parameters affect the probability curves is fundamental to applying each method of interpretation.

3.7.1. The Effects of the Parameters

Consider the BRM with a single x:

$$\Pr(y = 1 \mid x) = F(\alpha + \beta x)$$

Panel A of Figure 3.8 shows the effect of the intercept on the probability curve. When $\alpha = 0$, shown by the short dashed line, the curve passes through the point (0, .5). As α gets larger, the curve shifts to the left; as α gets smaller, the curve shifts to the right. (*Why does the curve shift to the left when α increases?*) When the curve shifts, the slope at a given value of $\Pr(y = 1 \mid x)$ does not change. This idea of shifting, "parallel" curves is used to explain several of the methods presented below. It is also fundamental to understanding the ordinal regression model in Chapter 5.

Panel B of Figure 3.8 shows the effects of changing the slope. Since $\alpha = 0$, the curves go through point (0, .5). The smaller the β, the more stretched out the curve. At $\beta = .25$, shown by the solid line, the curve increases steadily as it moves from -20 to 20. When β increases to .5, shown by the long dashed line, the curve initially increases more slowly. As x approaches 0, the increase is more rapid. In general, as β increases, the curve increases more rapidly as x approaches 0. While I have not drawn the curves, when the slope is negative, the curve is rotated 180° around $x = 0$. For example, if $\beta = -.25$, the curve would be near 1 at $x = -20$, and would gradually decrease toward 0 at $x = 20$.

It is also important to understand how the probability curve generalizes to more than one variable. Figure 3.9 plots the probit model:

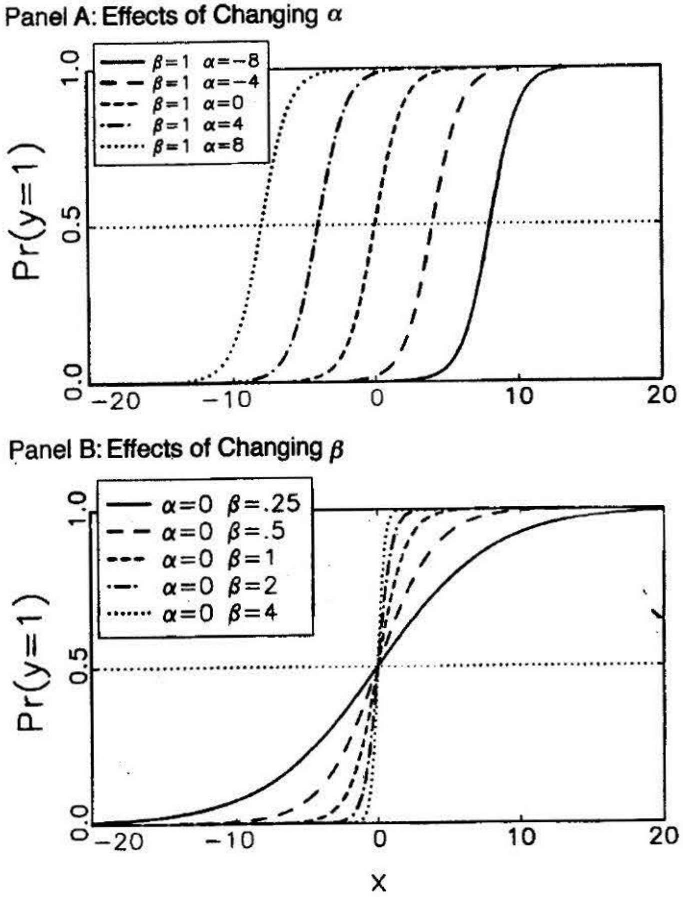
$$\Pr(y = 1 \mid x, z) = \Phi(1 + 1x + .75z)$$

Similar results apply for the logit model. The surface begins near zero when $x = -4$ and $z = -8$. If we fix $z = -8$, then

$$\Pr(y = 1 \mid x, z = -8) = \Phi(1 + 1x + [.75 \times -8]) = \Phi(-5.0 + 1x)$$

which is the first S-shaped curve along the x-axis. If we increase z by 1, which corresponds to the next curve back along the z-axis, then

$$\Pr(y = 1 \mid x, z = -7) = \Phi(1 + 1x + [.75 \times -7]) = \Phi(-4.25 + 1x)$$


Panel A: Effects of Changing α

Panel B: Effects of Changing β

Figure 3.8. Effects of Changing the Slope and Intercept on the Binary Response Model: $\Pr(y = 1 \mid x) = F(\alpha + \beta x)$

Only the intercept has changed, which causes the curve to shift to the left (see panel A of Figure 3.8). The level of z affects the intercept of the curve, but does not affect the slope. Conversely, controlling for x affects the intercept of the curve for z, but not the slope.

With these ideas in mind, we can consider several methods for interpreting the binary response model.

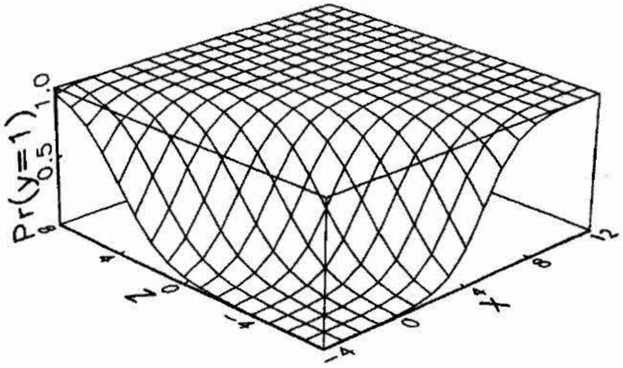


Figure 3.9. Plot of Probit Model: $\Pr(y=1 \mid x, z) = \Phi(1.0 + 1.0x + 0.75z)$

3.7.2. Interpretation Using Predicted Probabilities

The most direct approach for interpretation is to examine the predicted probabilities of an event for different values of the independent variables. When there are more than two variables, it is no longer possible to plot the entire probability surface and a decision must be made regarding which probabilities to compute and how to present them. A useful first step is to examine the range of predicted probabilities within the sample, and the degree to which each variable affects the probabilities. If the range of probabilities is between .2 and .8 (or, more conservatively, between .3 and .7), the relationship between the x's and the predicted probability is nearly linear, and simple measures can be used to summarize the results. Or, if the range of the probability is small, the relationship between the x's and the probability will also be approximately linear. For example, the segment of the probability curve between .05 and .10 is nearly linear. These points are illustrated below.

Determining the Range of Probabilities

The predicted probability of an event given $\mathbf{x}$ for the ith individual is

$$\widehat{\Pr}(y_i = 1 \mid \mathbf{x}_i) = F(\mathbf{x}_i\widehat{\boldsymbol{\beta}})$$

The minimum and maximum probabilities in the sample are defined as

$$\min \widehat{\Pr}(y = 1 \mid \mathbf{x}) = \min_i F(\mathbf{x}_i\widehat{\boldsymbol{\beta}})$$
$$\max \widehat{\Pr}(y = 1 \mid \mathbf{x}) = \max_i F(\mathbf{x}_i\widehat{\boldsymbol{\beta}})$$

where $\min_i$ indicates taking the minimum value over all observations, and similarly for $\max_i$. In our example, the predicted probabilities from the probit model range from .01 to .97, which indicates that the nonlinearities that occur below .2 and above .8 need to be taken into account. If the coefficients from the logit model are used, the predicted probabilities range from .01 to .96. This illustrates the great similarity between the predictions of the logit and probit models, even for observations that fall in the tail of the distribution. Consequently, in the remainder of this section, only the results from the probit analysis are shown.

Computing the minimum and maximum predicted probabilities requires your software to save each observation's predicted probability for further analysis. If this is not possible, or if you are doing a meta-analysis, the minimum and maximum can be approximated by using the estimated β's and the descriptive statistics. The *lower extreme* of the variables is defined by setting each variable associated with a positive β to its minimum and each variable associated with a negative β to its maximum. In our example, this involves taking the maximum number of young children (since *K6* has a negative effect), the minimum anticipated wage (since *LWG* has a positive effect), and so on. Formally, let

$$\overleftarrow{x}_k = \begin{cases} \min\limits_i x_{ik} & \text{if } \beta_k \geq 0 \\ \max\limits_i x_{ik} & \text{if } \beta_k < 0 \end{cases}$$

and let $\overleftarrow{\mathbf{x}}$ be the vector whose kth element is $\overleftarrow{x}_k$. The *upper extreme* can be defined in a corresponding way, with the values contained in $\overrightarrow{\mathbf{x}}$. The minimum and maximum probabilities are computed as

$$\widehat{\Pr}(y = 1 \mid \overleftarrow{\mathbf{x}}) = F(\overleftarrow{\mathbf{x}}\widehat{\boldsymbol{\beta}}) \quad \text{and} \quad \widehat{\Pr}(y = 1 \mid \overrightarrow{\mathbf{x}}) = F(\overrightarrow{\mathbf{x}}\widehat{\boldsymbol{\beta}})$$

In our example, the computed probability at the lower extreme is less than .01 and at the upper extreme is .99. While these values are quite close to the minimum and maximum predicted probabilities for the sample, $\overleftarrow{\mathbf{x}}$ and $\overrightarrow{\mathbf{x}}$ are constructs that do not necessarily approximate any member of the sample. If they differ substantially from any $\mathbf{x}_i$ in the sample, then $\widehat{\Pr}(y = 1 \mid \overleftarrow{\mathbf{x}})$ and $\widehat{\Pr}(y = 1 \mid \overrightarrow{\mathbf{x}})$ will be poor approximations of the probabilities $\min \widehat{\Pr}(y = 1 \mid \mathbf{x})$ and $\max \widehat{\Pr}(y = 1 \mid \mathbf{x})$.

Warning on the Use of Minimums and Maximums. The use of the minimum or maximum value of a variable can be misleading if there are extreme values in the sample. For example, if our sample includes an extremely wealthy person, the change in the probability when we move from the minimum to the maximum income would be unrealistically large. Before using the minimum and maximum, you should examine the frequency distribution of each variable. If extreme values are present, you should consider using the 5th percentile and the 95th percentile, for example, rather than the minimum and maximum.

The Effect of Each Variable on the Predicted Probability

The next step is to determine the extent to which change in a variable affects the predicted probability. One way to do this is to allow one variable to vary from its minimum to its maximum, while all other variables are fixed at their means. Let $\Pr(y = 1 \mid \bar{\mathbf{x}}, x_k)$ be the probability computed when all variables except x_k are set equal to their means, and x_k equals some specified value. For example, $\Pr(y = 1 \mid \bar{\mathbf{x}}, \min x_k)$ is the probability when x_k equals its minimum. The predicted change in the probability as x_k changes from its minimum to its maximum equals

$$\Pr(y = 1 \mid \bar{\mathbf{x}}, \max x_k) - \Pr(y = 1 \mid \bar{\mathbf{x}}, \min x_k)$$

For our example, the results are given in Table 3.4. The range of predicted probabilities can be used to guide further analysis. For example, there is little to be learned by analyzing variables whose range of probabilities is small, such as *HC*. For variables that have a larger range, the end points of the range affect how interpretation should proceed. For example, the predicted probabilities for *AGE* range from .75 when age is 30 to .32 when age is 60, which is a region where the probability curve is nearly linear. The range for *INC*, however, is from .09 to .73, where nonlinearities are present. The implications of these differences are shown in the next section.

TABLE 3.4 Probabilities of Labor Force Participation Over the Range of Each Independent Variable for the Probit Model

Variable	*At Maximum*	*At Minimum*	*Range of* $\widehat{\Pr}$
K5	0.01	0.66	0.64
K618	0.48	0.60	0.12
AGE	0.32	0.75	0.43
WC	0.71	0.52	0.18
HC	0.59	0.57	0.02
LWG	0.83	0.17	0.66
INC	0.09	0.73	0.64

Plotting Probabilities Over the Range of a Variable

When there are more than two independent variables, we must examine the effects of one or two variables while the remaining variables are held constant. For example, consider the effects of age and the wife attending college on labor force participation. The effects of both variables can be plotted by holding all other variables at their means and allowing age and college status to vary. To do this, let $\mathbf{x}_0$ contain the mean of all variables, except let $WC = 0$ and allow AGE to vary. $\mathbf{x}_1$ is defined similarly for $WC = 1$. Then

$$\widehat{\Pr}(LFP = 1 \mid AGE, WC = 0) = \Phi(\mathbf{x}_0\widehat{\boldsymbol{\beta}})$$

is the predicted probability of being in the labor force for women of a given age who did not attend college and who are average on all other characteristics. $\widehat{\Pr}(LFP = 1 \mid AGE,\ WC = 1)$ can be computed similarly. These probabilities are plotted in Figure 3.10. As suggested by Table 3.4, the relationship between age and the probability of being employed is approximately linear. This allows a very simple interpretation:

- Attending college increases the probability of being employed by about .18 for women of all ages, holding all other variables at their means.
- For each additional 10 years of age, the probability of being employed decreases by about .13, holding all other variables at their means.

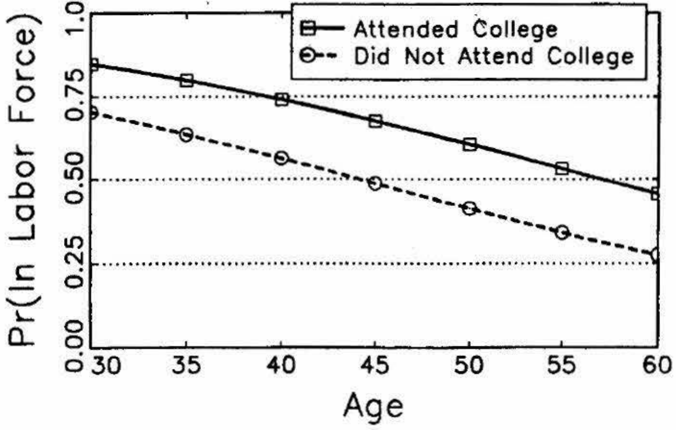


Figure 3.10. Probability of Labor Force Participation by Age and Wife's Education

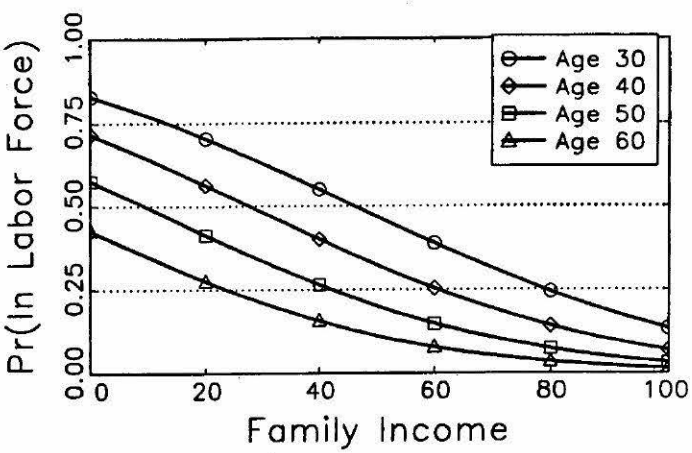


Figure 3.11. Probability of Labor Force Participation by Age and Family Income for Women Without Some College Education

The effect of age was computed by subtracting the predicted probability at age 30 (= .85) from that at age 60 (= .46) and dividing by 3 (for three periods of ten years). It would also be appropriate to use the marginal effect computed at the mean, which is discussed in Section 3.7.4.

The relationship between age and the probability of working was nearly linear and the plot was superfluous. In other cases, plotting is very useful. Consider the effects of income and age. While we could hold all other variables at their means and draw a three-dimensional plot, it is often more informative to divide one of the variables into groups and plot the results in two dimensions. Figure 3.11 shows the probability of employment as income changes for women aged 30, 40, 50, and 60. The nonlinearities are apparent, with the effect of income decreasing with age. When relationships are nonlinear, plots are often useful for uncovering relationships, even if they are not used to present the findings.

Tables of Predicted Probabilities at Selected Values

You can also use tables to present predicted probabilities. For example, the effects of young children and the wife's education on the probability of employment are shown in Table 3.5. The strong, nonlinear effect of having young children is clearly evident. It also shows that the effect of attending college decreases as the number of children increases. (*The difference in the probability for those attending and not attending college increases and then decreases. Draw the probability curves that produce this result.*)

TABLE 3.5 Probability of Employment by College Attendence and the Number of Young Children for the Probit Model

Number of Young Children	*Predicted Probability*		
	Did Not Attend	*Attended College*	*Difference*
0	0.61	0.78	0.17
1	0.27	0.45	0.18
2	0.07	0.16	0.09
3	0.01	0.03	0.02

Another strategy for presenting probabilities is to define combinations of characteristics that correspond to ideal types in the population. For example, in his study of factors that affected the retention of workers by their employer after training programs, Gunderson (1974) defined five "hypothetical trainees" based on combinations of the independent variables: typical, disadvantaged, advantaged, housewife, and teenage entrant. Predicted probabilities of being retained were computed for each hypothetical person. In some situations, this can quickly and convincingly summarize the effects of key variables.

3.7.3. The Partial Change in y^*

Measures of partial change can also be used to summarize the effects of each independent variable on the probability of an event occurring. Recall that the logit and probit models are linear in the latent variable:

$$y^* = \mathbf{x}\boldsymbol{\beta} + \varepsilon$$

Taking the partial derivative with respect to x_k,

$$\frac{\partial y^*}{\partial x_k} = \beta_k$$

Since the model is linear in y^*, the partial derivative can be interpreted as:

- For a unit change in x_k, y^* is expected to change by β_k units, holding all other variables constant.

The problem with this interpretation is that the variance of y^* is unknown, so the meaning of a change of β_k in y^* is unclear. This issue was discussed by Winship and Mare (1984, p. 517) and McKelvey and Zavoina (1975, pp. 114–116) regarding the ordinal regression model, but their concerns apply equally to the BRM. Since the variance of y^* changes when new variables are added to the model, the magnitudes of all β's will change even if the added variable is uncorrelated with the original variables. This makes it misleading to compare coefficients from different specifications of the independent variables. (*Why is this not a problem with the LRM?*) To compare coefficients across equations, McKelvey and Zavoina proposed fully standardized coefficients, while Winship and Mare suggested y^*-standardized coefficients.

If σ_{y^*} is the unconditional standard deviation of y^*, then the *y^*-standardized coefficient* for x_k is

$$\beta_k^{Sy^*} = \frac{\beta_k}{\sigma_{y^*}}$$

which can be interpreted as:

- For a unit increase in x_k, y^* is expected to increase by $\beta_k^{Sy^*}$ standard deviations, holding all other variables constant.

y^*-standardized coefficients indicate the effect of an independent variable in its original unit of measurement. This is sometimes preferable for substantive reasons and is necessary for binary independent variables.

Fully standardized coefficients also standardize the independent variable. If σ_k is the standard deviation of x_k, then the *fully standardized coefficient* for x_k is

$$\beta_k^S = \frac{\sigma_k \beta_k}{\sigma_{y^*}} = \sigma_k \beta_k^{Sy^*}$$

which can be interpreted as:

- For a standard deviation increase in x_k, y^* is expected to increase by β_k^S standard deviations, holding all other variables constant.

To compute $\widehat{\beta}_k^{Sy}$ and $\widehat{\beta}_k^S$, we need estimates of β_k, σ_k, and σ_{y^*}. The standard deviations of the x's can be computed directly from the observed data. Since $y^* = \mathbf{x}\boldsymbol{\beta} + \varepsilon$, and $\mathbf{x}$ and ε are uncorrelated, $\sigma_{y^*}^2$ can be estimated by the quadratic form:

$$\widehat{\text{Var}}(y^*) = \widehat{\boldsymbol{\beta}}' \widehat{\text{Var}}(\mathbf{x}) \widehat{\boldsymbol{\beta}} + \text{Var}(\varepsilon)$$

TABLE 3.6 Standardized and Unstandardized Probit Coefficients for Labor Force Participation

Variable	β	β^{Sy^*}	β^S	z
K5	−0.875	−0.759	−0.398	−7.70
K618	−0.039	−0.033	−0.044	−0.95
AGE	−0.038	−0.033	−0.265	−4.97
WC	0.488	0.424	0.191	3.60
HC	0.057	0.050	0.024	0.46
LWG	0.366	0.317	0.186	4.17
INC	−0.021	−0.018	−0.207	−4.30
$\widehat{\text{Var}}(y^*)$	1.328			

NOTE: $N = 753$. β is an unstandardized coefficient β^{Sy^*} is a y^*-standardized coefficient; β^S is a fully standardized coefficient. z is the z-test.

$\widehat{\text{Var}}(\mathbf{x})$ is the covariance matrix for the x's computed from the observed data; $\widehat{\boldsymbol{\beta}}$ contains ML estimates; and $\text{Var}(\varepsilon) = 1$ in the probit model and $\text{Var}(\varepsilon) = \pi^2/3$ in the logit model.

If you accept the notion that it is meaningful to discuss the latent propensity to work, the fully standardized and y^*-standardized coefficients in Table 3.6 can be interpreted just as their counterparts for the LRM.[3] For example,

- Each additional young child decreases the mother's propensity to enter the labor market by .76 standard deviations, holding all other variables constant.
- A standard deviation increase in age decreases a woman's propensity to enter the labor market by .27 standard deviations, holding all other variables constant.

3.7.4. The Partial Change in $\Pr(y = 1 \mid \mathbf{x})$

The β's can also be used to compute the partial change in the probability of an event. Let

$$\Pr(y = 1 \mid \mathbf{x}) = F(\mathbf{x}\boldsymbol{\beta}) \qquad [3.13]$$

where F is either the cdf Φ for the normal distribution or the cdf Λ for the logistic distribution. The corresponding pdf is indicated as f. The *partial change in the probability*, also called the *marginal effect*, is

[3] If you try to reproduce the standardized coefficients in Table 3.6 using the descriptive statistics from Table 3.1, your answers will only match to the first decimal digit due to rounding.

computed by taking the partial derivative of Equation 3.13 with respect to x_k:[4]

$$\frac{\partial \Pr(y=1\mid \mathbf{x})}{\partial x_k} = \frac{\partial F(\mathbf{x}\boldsymbol{\beta})}{\partial x_k} = \frac{dF(\mathbf{x}\boldsymbol{\beta})}{d\,\mathbf{x}\boldsymbol{\beta}}\frac{\partial \mathbf{x}\boldsymbol{\beta}}{\partial x_k} = f(\mathbf{x}\boldsymbol{\beta})\beta_k \qquad [3.14]$$

For the probit model,

$$\frac{\partial \Pr(y=1\mid \mathbf{x})}{\partial x_k} = \phi(\mathbf{x}\boldsymbol{\beta})\beta_k$$

and for the logit model,

$$\begin{aligned}\frac{\partial \Pr(y=1\mid \mathbf{x})}{\partial x_k} &= \lambda(\mathbf{x}\boldsymbol{\beta})\beta_k = \frac{\exp(\mathbf{x}\boldsymbol{\beta})}{[1+\exp(\mathbf{x}\boldsymbol{\beta})]^2}\,\beta_k \\ &= \Pr(y=1\mid \mathbf{x})[1-\Pr(y=1\mid \mathbf{x})]\beta_k\end{aligned}$$

(*Prove the last equality.*)

The marginal effect is the slope of the probability curve relating x_k to $\Pr(y=1\mid \mathbf{x})$, holding all other variables constant. The *sign* of the marginal effect is determined by β_k, since $f(\mathbf{x}\boldsymbol{\beta})$ is always positive. The *magnitude* of the change depends on the magnitude of β_k and the value of $\mathbf{x}\boldsymbol{\beta}$. This is shown in Figure 3.12, where the solid line graphs $\Pr(y=1\mid x)$ and the dashed line graphs the marginal effect. The marginal is largest at $x=x_2$, which corresponds to $\Pr(y=1\mid x)=.5$. The marginal is symmetric around x_2, reflecting the symmetry of f. Therefore,

$$\frac{\partial \Pr(y=1\mid x=x_1)}{\partial x} = \frac{\Pr(y=1\mid x=x_3)}{\partial x}.$$

The magnitude of the marginal effect depends on the values of the other variables and their coefficients, since f is computed at $\mathbf{x}\boldsymbol{\beta}$. Consequently, the marginal depends on the β's for all variables and the levels of all x's. To understand how the value of the marginal effect of x_k depends on the level of other variables, consider Figure 3.9 which plots the probability surface for variables x and z. Pick a point (x, z), which corresponds to the intersection of lines within the figure. The partial $\partial \Pr(y=1\mid x,z)/\partial x$ is the slope of the line parallel to the x-axis at the point (x, z); $\partial \Pr(y=1\mid x,z)/\partial z$ is the slope of the line parallel to the z axis at the point (x, z). For example, at $(-4, -8)$, the slope with respect to x is nearly 0. As z increases, the slope with respect to x increases steadily. At $(-4, 0)$, where $\Pr(y=1\mid x,z)$ is about .5, the slope is near its maximum. As z continues to increase, the slope gradually decreases. Hanushek and Jackson (1977, p. 189) show this relationship by taking the second derivative:

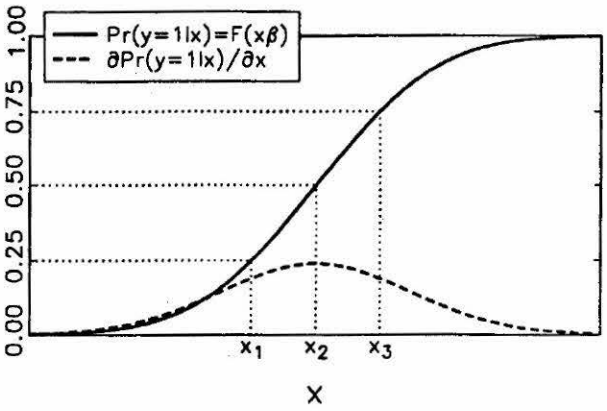


Figure 3.12. Marginal Effect in the Binary Response Model

$$\frac{\partial^2 \Pr(y=1\mid \mathbf{x})}{\partial x_k \partial x_\ell} = \beta_k\beta_\ell \Pr(y=1\mid \mathbf{x})[1-\Pr(y=1\mid \mathbf{x})][1-2\Pr(y=1\mid \mathbf{x})]$$

The β's can also be used to assess the relative magnitudes of the marginal effect for two variables. From Equation 3.14, the ratio of marginal effects for x_k and x_ℓ is

$$\frac{\dfrac{\partial \Pr(y=1\mid \mathbf{x})}{\partial x_k}}{\dfrac{\partial \Pr(y=1\mid \mathbf{x})}{\partial x_\ell}} = \frac{f(\mathbf{x}\boldsymbol{\beta})\beta_k}{f(\mathbf{x}\boldsymbol{\beta})\beta_\ell} = \frac{\beta_k}{\beta_\ell}$$

Thus, while the β's are only identified up to a scale factor, their ratio is identified and can be used to compare the effects of independent variables.

[4] We use the chain rule:

$$\frac{\partial f(g(x))}{\partial x} = \frac{\partial f(g(x))}{\partial g(x)}\frac{\partial g(x)}{\partial x}$$

and the derivative:

$$\frac{\partial F(x)}{\partial x} = f(x)$$

Since the value of the marginal effect depends on the levels of all variables, we must decide on which values of the variables to use when computing the effect. One method is to compute the average over all observations:

$$\text{mean}\ \frac{\partial \Pr(y=1\mid \mathbf{x})}{\partial x_k} = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i\boldsymbol{\beta})\beta_k$$

Another method is to compute the marginal effect at the mean of the independent variables:

$$\frac{\partial \Pr(y=1\mid \bar{\mathbf{x}})}{\partial x_k} = f(\bar{\mathbf{x}}\boldsymbol{\beta})\beta_k$$

Stata's
mfx
default

The *marginal effect at the mean* is a popular summary measure for models with categorical dependent variables. It is frequently included in tables presenting results, and is automatically computed by programs such as LIMDEP. However, the measure is limited. First, given the nonlinearity of the model, it is difficult to translate the marginal effect into the change in the predicted probability that will occur if there is a discrete change in x_k. Second, since $\bar{\mathbf{x}}$ might not correspond to any observed values in the population, averaging over observations might be preferred. Finally, the measure is inappropriate for binary independent variables. For these reasons, I much prefer the measures of discrete change that are discussed in Section 3.7.5.

Table 3.7 contains marginal effects for our example of labor force participation. Several things should be noted. First, the marginal effects averaged over all observations are close to the marginals computed when all variables are held at their means. They are close since the predicted probability overall is approximately .5 in the sample. In general, these two measures of change can be quite different. Second, the marginal effect at the mean for *AGE* approximates the slope of the lines in Figure 3.10. If an independent variable varies over a region of the probability curve that is nearly linear, the marginal effect can be used to summarize the effect of a unit change in the variable on the probability of an event. However, if the range of an independent variable corresponds to a region of the probability curve that is nonlinear, the marginal cannot be used to assess the overall effect of the variable.

TABLE 3.7 Marginal Effects on the Probability of Being Employed for the Probit Model

Variable	*Average*	*At Mean*
K5	−0.300	−0.342
K618	−0.013	−0.015
AGE	−0.013	−0.015
WC	0.167	0.191
HC	0.020	0.022
LWG	0.125	0.143
INC	−0.007	−0.008

3.7.5. Discrete Change in $\Pr(y=1\mid \mathbf{x})$

The change in the predicted probabilities for a discrete change in an independent variable is an alternative to the marginal effect that I find more effective for interpreting the BRM (as well as other models for categorical outcomes). Let $\Pr(y=1\mid \mathbf{x}, x_k)$ be the probability of an event given $\mathbf{x}$, noting, in particular, the value of x_k. Thus, $\Pr(y=1\mid \mathbf{x}, x_k+\delta)$ is the probability with x_k increased by δ, while the other variables are unchanged. The *discrete change* in the probability for a change of δ in x_k equals

$$\frac{\Delta \Pr(y=1\mid \mathbf{x})}{\Delta x_k} = \Pr(y=1\mid \mathbf{x}, x_k+\delta) - \Pr(y=1\mid \mathbf{x}, x_k)$$

The discrete change can be interpreted as:

- For a change in the variable x_k from x_k to $x_k+\delta$, the predicted probability of an event changes by $\Delta \Pr(y=1\mid \mathbf{x})/\Delta x_k$, holding all other variables constant.

When interpreting the results of the BRM, it is essential to understand that the partial change does not equal the discrete change:

$$\frac{\partial \Pr(y=1\mid \mathbf{x})}{\partial x_k} \neq \frac{\Delta \Pr(y=1\mid \mathbf{x})}{\Delta x_k}$$

except in the limit as δ becomes infinitely small (which is, by definition, the partial change). The difference between these two measures is shown in Figure 3.13 which plots a segment of the probability curve. The partial change is the tangent at x_1, and its value corresponds to the solid triangle. For simplicity, assume that $\delta = 1$. The discrete change measures the change in the probability computed at x_1 and $x_1 + 1$. This is represented by a triangle formed of dashed lines. The discrete and partial changes are not equal since the rate of change in the curve changes

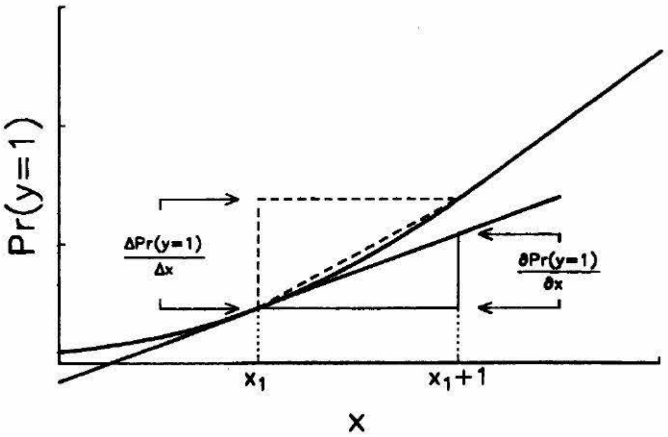


Figure 3.13. Partial Change Versus Discrete Change in Nonlinear Models

as x_k changes. While the measures are not equal, if the change in x_k occurs over a region of the probability curve that is roughly linear, the two measures will be close. This is the case for the example in Figure 3.10.

The amount of discrete change in the probability for a change in x_k depends on: (1) the amount of change in x_k; (2) the starting value of x_k; and (3) the values of all other variables. For example, if we have independent variables x_1 and x_2, the change in $\Pr(y = 1 \mid \mathbf{x})$ when x_1 changes from 1 to 2 does not necessarily equal the change when x goes from 2 to 3. (*Why would they be equal if* $\Pr(y = 1 \mid \mathbf{x}) = .5$?) Moreover, the change in $\Pr(y = 1 \mid \mathbf{x})$ when x_1 changes from 1 to 2 with $x_2 = 1$ does not necessarily equal the change when $x_2 = 2$. Thus, the practical problem is choosing which values of the variables to consider and how much to let them change.

Choosing Values of the Independent Variables

Since the change in the probability for a given change in x_k depends on the levels of all independent variables, we must decide at which values of the x's to compute the discrete change. A common approach is to assess the probability for an "average" member of the sample. For example, we could hold all values at their means. If the independent variables are highly skewed, assessing change relative to the mean may be misleading and changes relative to the median would be more useful.

Dummy variables require special consideration. If x_d is a dummy variable, $\bar{x}_d$ is the proportion of the sample with $x_d = 1$. The predicted probability at $\bar{x}_d$ is between the predicted probability at $x_d = 1$ and $x_d = 0$. Alternatively, you could compute the predicted probability for each combination of the dummy variables, with the other variables held at their means. In our labor force example, this would require four base probabilities: husband and wife attending college; only the husband attending; only the wife attending; and neither attending. Alternatively, dummy variables could be held at the modal value for each variable.

If there is a combination of the independent variables that is of particular substantive interest, those values could be used as a baseline. For example, if you were interested in the effects of education on labor force participation for young women without children, you could hold *AGE* at 30, *K5* at 0, *K618* at 0, and all other variables at their means. In the following examples, I hold all variables at their means.

Amounts of Change in the Independent Variables

Discrete change can be computed for any amount of change in an independent variable, holding all other variables at some fixed value. The amount of change that you allow for an independent variable depends on the type of variable and your purpose. Here are some useful options.

A Unit Change in x_k. If x_k increases from $\bar{x}_k$ to $\bar{x}_k + 1$,

$$\frac{\Delta \Pr(y = 1 \mid \bar{\mathbf{x}})}{\Delta x_k} = \Pr(y = 1 \mid \bar{\mathbf{x}}, \bar{x}_k + 1) - \Pr(y = 1 \mid \bar{\mathbf{x}}, \bar{x}_k)$$

By examining the probability curves (see Figure 3.8), it is clear that a unit *increase* in x_k from its mean will only have the same effect as a unit *decrease* in x_k from its mean when $\Pr(y = 1 \mid \bar{\mathbf{x}}) = .5$. This implies that if you have two variables such that $\beta_k = -\beta_\ell$, the effect of a unit increase in x_k will not equal the effect of a unit decrease in x_ℓ. For these reasons, Kaufman (1996) suggested examining a unit increase that is centered around $\bar{x}_k$. That is,

$$\frac{\Delta \Pr(y = 1 \mid \bar{\mathbf{x}})}{\Delta x_k} = \Pr\left(y = 1 \mid \bar{\mathbf{x}}, \bar{x}_k + \frac{1}{2}\right) - \Pr\left(y = 1 \mid \bar{\mathbf{x}}, \bar{x}_k - \frac{1}{2}\right)$$

The *centered discrete change* can be interpreted as:

- A unit change in x_k that is centered around $\bar{x}_k$ results in a change of $\Delta \Pr(y = 1 \mid \bar{\mathbf{x}})/\Delta x_k$ in the predicted probability, holding all other variables at their means.

A Standard Deviation Change in x_k. This idea can be extended to examine the effect of a standard deviation change:

$$\frac{\Delta \Pr(y=1\mid\bar{\mathbf{x}})}{\Delta x_k} = \Pr\left(y=1\mid\bar{\mathbf{x}},\bar{x}_k+\frac{s_k}{2}\right) - \Pr\left(y=1\mid\bar{\mathbf{x}},\bar{x}_k-\frac{s_k}{2}\right)$$

where s_k is the standard deviation of x_k.

A Change From 0 to 1 for Dummy Variables. When computing a discrete change in probability, you must make certain that the change in the variable does not result in values that exceed the variable's range. For example, if x_k is a dummy variable, either $\bar{x}_k + 1/2$ will exceed 1 or $\bar{x}_k - 1/2$ will be negative (unless $\bar{x}_k = 1/2$). Consequently, a preferred measure of discrete change for dummy variables is

$$\frac{\Delta \Pr(y=1\mid\bar{\mathbf{x}})}{\Delta x_k} = \Pr(y=1\mid\bar{\mathbf{x}},\ x_k=1) - \Pr(y=1\mid\bar{\mathbf{x}},\ x_k=0)$$

This is the change as x_k goes from 0 to 1, holding all other variables at their means.

Other Choices. The idea of discrete change can be extended in many ways depending on the application. If a change of a specific amount is substantively important, such as the addition of four years of schooling, changes other than 1 or s_k can be used.

Example of Discrete Change: Labor Force Participation

Table 3.8 contains measures of discrete change for the probit model of women's labor force participation. Some of the effects can be interpreted as:

TABLE 3.8 Discrete Change in the Probability of Employment for the Probit Model

Variable	*Centered Unit* Change	*Centered Standard Deviation Change*	*Change From 0 to 1*
K5	−0.33	−0.18	—
K618	−0.02	−0.02	—
AGE	−0.01	−0.12	—
WC	—	—	0.18
HC	—	—	0.02
LWG	0.14	0.08	—
INC	−0.01	−0.09	—

NOTE: Changes are computed with other variables held at their means.

- For a woman who is average on all characteristics, an additional young child decreases the probability of employment by .33.
- A standard deviation change in age centered around the mean will decrease the probability of working by .12, holding all other variables constant.
- If a woman attends college, her probability of being in the labor force is .18 greater than a woman who does not attend college, holding all other variables at their means.

Notice that the discrete change from 0 to 1 for *WC* and *HC* is nearly identical to the effect of a unit change. This is a consequence of the near linearity of the probability curve over the range of these variables, and will not necessarily be true in other examples.

3.8. Interpretation Using Odds Ratios

Our final method of interpretation takes advantage of the tractable form of the logit model. A simple transformation of the β's in the logit model indicates the factor change in the odds of an event occurring. There is no corresponding transformation of the parameters of the probit model.

From Equation 3.6, the logit model can be written as the log-linear model:

$$\ln \Omega(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} \qquad [3.15]$$

where

$$\Omega(\mathbf{x}) = \frac{\Pr(y=1\mid\mathbf{x})}{\Pr(y=0\mid\mathbf{x})} = \frac{\Pr(y=1\mid\mathbf{x})}{1-\Pr(y=1\mid\mathbf{x})} \qquad [3.16]$$

is the odds of the event given **x**. $\ln \Omega(\mathbf{x})$ is the log of the odds, known as the *logit*. Equation 3.15 shows that the logit model is linear in the logit. Consequently,

$$\frac{\partial \ln \Omega(\mathbf{x})}{\partial x_k} = \beta_k$$

Since the model is linear, β_k can be interpreted as:

- For a unit change in x_k, we expect the logit to change by β_k, holding all other variables constant.

This interpretation is simple since the effect of a unit change in x_k on the logit does not depend on the level of x_k or on the level of any other variable. Unfortunately, most of us do not have an intuitive understanding of what a change in the logit means. This requires another transformation.

Taking the exponential of Equation 3.15,

$$\begin{aligned}\Omega(\mathbf{x}) &= \exp(\mathbf{x}\boldsymbol{\beta})\\ &= \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \cdots + \beta_K x_K)\\ &= \exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_k x_k)\cdots\exp(\beta_K x_K) = \Omega(\mathbf{x}, x_k)\end{aligned}$$

The last equality introduces notation that makes explicit the value of x_k. To assess the effect of x_k, we want to see how Ω changes when x_k changes by some quantity δ. Most often, we consider $\delta = 1$ or $\delta = s_k$. If we change x_k by δ, the odds become

$$\begin{aligned}&\Omega(\mathbf{x}, x_k + \delta)\\ &\quad = \exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_k(x_k + \delta))\cdots\exp(\beta_K x_K)\\ &\quad = \exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_k x_k)\exp(\beta_k\delta)\cdots\exp(\beta_K x_K)\end{aligned}$$

To compare the odds before and after adding δ to x_k, we take the *odds ratio*:

$$\begin{aligned}&\frac{\Omega(\mathbf{x}, x_k + \delta)}{\Omega(\mathbf{x}, x_k)}\\ &\quad = \frac{\exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_k x_k)\exp(\beta_k\delta)\cdots\exp(\beta_K x_K)}{\exp(\beta_0)\exp(\beta_1 x_1)\cdots\exp(\beta_k x_k)\cdots\exp(\beta_K x_K)}\\ &\quad = \exp(\beta_k\delta)\end{aligned}$$

Therefore, the parameters can be interpreted in terms of odds ratios:

- For a change of δ in x_k, the odds are expected to change by a factor of $\exp(\beta_k \times \delta)$, holding all other variables constant.

For $\delta = 1$, we have:

- *Factor change*. For a unit change in x_k, the odds are expected to change by a factor of $\exp(\beta_k)$, holding all other variables constant.

If $\exp(\beta_k)$ is greater than 1, you could say that the odds are "$\exp(\beta_k)$ times larger." If $\exp(\beta_k)$ is less than 1, you could say that the odds are "$\exp(\beta_k)$ times smaller." For $\delta = s_k$, we have:

- *Standardized factor change*. For a standard deviation change in x_k, the odds are expected to change by a factor of $\exp(\beta_k \times s_k)$, holding all other variables constant.

Notice that the effect of a change in x_k does not depend on the level of x_k or on the level of any other variable.

We can also compute the percentage change in the odds:

$$100\,\frac{\Omega(\mathbf{x}, x_k + \delta) - \Omega(\mathbf{x}, x_k)}{\Omega(\mathbf{x}, x_k)} = 100[\exp(\beta_k \times \delta) - 1]$$

This quantity can be interpreted as the percentage change in the odds for a δ unit change in x_k, holding all other variables constant.

The factor change and standardized factor change coefficients for the logit model analyzing labor force participation are presented in Table 3.9. Here is how some of the coefficients can be interpreted using factor and percentage changes:

- For each additional young child, the odds of being employed are decreased by a factor of .23, holding all other variables constant. Or, equivalently, for each additional young child, the odds of working are decreased 77%, holding all other variables constant.
- For a standard deviation increase in anticipated wages, the odds of being employed are 1.43 times greater, holding all other variables constant. Or, for a standard deviation increase in anticipated wages, the odds of working are 43% greater, holding all other variables constant.
- Being 10 years older decreases the odds by a factor of .52 ($= e^{-.063\times 10}$), holding all other variables constant.

TABLE 3.9 Factor Change Coefficients for Labor Force Participation for the Logit Model

Variable	*Logit Coefficient*	*Factor Change*	*Standard Factor Change*	*z-value*
Constant	3.182	—	—	4.94
K5	−1.463	0.232	0.465	−7.43
K618	−0.065	0.937	0.918	−0.95
AGE	−0.063	0.939	0.602	−4.92
WCOL	0.807	2.242	—	3.51
HCOL	0.112	1.118	—	0.54
WAGE	0.605	1.831	1.427	4.01
INC	−0.034	0.966	0.670	−4.20

The odds ratio is a multiplicative coefficient, which means that "positive" effects are greater than 1, while "negative" effects are between 0 and 1. *Magnitudes of positive and negative effects should be compared by taking the inverse of the negative effect* (*or vice versa*). For example, a positive factor change of 2 has the same magnitude as a negative factor change of $.5 = 1/2$. Thus, a coefficient of $.1 = 1/10$ indicates a stronger effect than a coefficient of 2. Another consequence of the multiplicative scale is that to determine the effect on the odds of the event not occurring, you simply take the inverse of the effect on the odds of the event occurring. For example,

- Being 10 years older makes the odds of not being in the labor force 1.9 $(= 1/.52)$ times greater, holding all other variables constant.

When interpreting the odds ratio, it is essential to keep the following in mind: *A constant factor change in the odds does not correspond to a constant change or constant factor change in the probability.* This can be seen in Table 3.10. While the odds are being changed by a constant factor of 2, the probabilities do not change by a constant factor or a constant amount. When the odds are very small, the factor change in the probability is approximately equal to the factor change in the odds. When the odds are large, the probability remains essentially unchanged. Consequently, when interpreting a factor change in the odds, it is *essential* to know what the current level of the odds is. This can be done using the methods in Section 3.7.2 to compute the predicted probability, and then computing the odds according to Equation 3.16.

TABLE 3.10 Factor Change of Two in the Odds With the Corresponding Factor Change and Change in the Probability

Original		*Changed*		*Factor Change*		*Change*
Odds	*Probability*	*Odds*	*Probability*	*Odds*	*Probability*	*in Probability*
1/1000	0.001	2/1000	0.002	2.000	1.998	0.001
1/100	0.010	2/100	0.020	2.000	1.980	0.010
1/10	0.091	2/10	0.167	2.000	1.833	0.076
1/2	0.333	2/2	0.500	2.000	1.500	0.167
1/1	0.500	2/1	0.667	2.000	1.333	0.167
2/1	0.667	4/1	0.800	2.000	1.200	0.133
10/1	0.909	20/1	0.952	2.000	1.048	0.043
100/1	0.990	200/1	0.995	2.000	1.005	0.005
1000/1	0.999	2000/1	0.999	2.000	1.000	0.000

3.9. Conclusions

The choice between the logit and probit models is largely one of convenience and convention, since the substantive results are generally indistinguishable. Chambers and Cox (1967) show that extremely large samples are necessary to distinguish whether observations were generated from the logit or the probit model. The availability of software is no longer an issue in choosing which model to use. Often the choice is a matter of convention. Some research areas tend to use logit, while others favor probit. For some users, the simple interpretation of logit coefficients as odds ratios is the deciding factor. In other cases, the need to generalize a model may be an issue. For example, multiple-equation systems involving qualitative dependent variables are based on the probit model, as discussed in Chapter 9. Or, if an analysis also includes equations with a nominal dependent variable, the logit model may be preferred since the probit model for nominal dependent variables is computationally too demanding. Or, in case-control studies where sampling is stratified by the binary outcome, the logit model is required (see Hosmer & Lemeshow, 1989, Chapter 6, for details).

Many of the ideas presented in this chapter are used to develop and interpret models for ordinal and nominal variables in Chapters 5 and 6. First, however, Chapter 4 considers hypothesis testing, methods for detecting outliers and influential observations, and measures of fit.

3.10. Bibliographic Notes

The very early history of these models begins in the 1860s and is discussed by Finney (1971, pp. 38–41). The more recent history of the probit model involves attempts to model the effects of toxins on insects. Work by Gaddum (1933) and Bliss (1934) was codified in Finney's influential *Probit Analysis* (1971), whose first edition appeared in 1947. The logit model was championed by Berkson (1944, 1951) in the 1940s as an alternative to the probit model. Cox's (1970) *The Analysis of Binary Data* was highly influential in the acceptance of the logit model. Applications of the logit and probit models appeared in economics in the 1950s (Cramer, 1991, p. 41). Goldberger's (1964, pp. 248–251) *Econometric Theory* was important in establishing these models as standard tools in economics, while Hanushek and Jackson's (1977) *Statistical Methods for Social Scientists* was important in disseminating these models to areas outside of economics.

McCullagh and Nelder (1989, Chapter 4) develop the logit and probit models, along with several alternatives, within the framework of the generalized linear model. Pudney (1989, Chapter 3) derives these models from behavioral assumptions associated with utility maximization. Agresti (1990, Chapter 4) presents both models with special attention to the links between logit analysis and log-linear models for categorical data. While the interpretation of the results of these models has often been neglected, each of the methods of interpretation considered in this chapter can be found in one form or another in earlier work. Recent treatments that focus on interpretation include Hanushek and Jackson (1977, pp. 187–207), King (1989a, pp. 97–117), Liao (1994), Long (1987), and Petersen (1985).

For a more advanced discussion of numerical methods, see Judge et al. (1985, pp. 951–979) and Greene (1993, pp. 343–357). For details on estimates of the covariance matrix, see Cramer (1986, pp. 27–29), Greene (1993, pp. 115–116), and Davidson and MacKinnon (1993, pp. 263–267).