CHAPTER 16

Qualitative and Limited Dependent Variable Models

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Give some examples of economic decisions in which the observed outcome is a binary variable.
2. Explain why probit, or logit, is usually preferred to least squares when estimating a model in which the dependent variable is binary.
3. Give some examples of economic decisions in which the observed outcome is a choice among several alternatives, both ordered and unordered.
4. Compare and contrast the multinomial logit model to the conditional logit model.
5. Give some examples of models in which the dependent variable is a count variable.
6. Discuss the implications of censored data for least squares estimation.
7. Describe what is meant by the phrase ‘‘sample selection.’’

KEYWORDS

alternative specific variables
binary choice models
censored data
conditional logit
count data models
feasible generalized least squares
Heckit
identification problem
independence of irrelevant alternatives (IIA)
index models
individual specific variables
latent variables
likelihood function
likelihood ratio
limited dependent variables
linear probability model
logistic random variable
logit
log-likelihood function
marginal effect
maximum likelihood estimation
multinomial choice models
multinomial logit
ordered probit
ordinal variables
Poisson random variable
Poisson regression model
probability ratio
probit
selection bias
Tobit model
truncated regression

In this book, we focus primarily on econometric models in which the dependent variable is continuous and fully observable; quantities, prices, and outputs are examples of such variables. However, microeconomics is a general theory of choice, and many of the choices that individuals and firms make cannot be measured by a continuous outcome variable. In this chapter, we examine some fascinating models that are used to describe choice behavior, and which do not have the usual continuous dependent variable. Our descriptions will be brief, since we will not go into all the theory, but we will reveal to you a rich area of economic applications.

We also introduce a class of models with dependent variables that are *limited*. By that we mean that they are continuous but that their range of values is constrained in some way, and their values not completely observable. Alternatives to least squares estimation must be considered for such cases, since the least squares estimator is both biased and inconsistent.

16.1 Introducing Models with Binary Dependent Variables

Many of the choices that individuals and firms make are "either-or" in nature. For example, a high-school graduate decides either to attend college or not. A worker decides either to drive to work or to get there using a different means of transportation. A household decides either to purchase a house or to rent. A firm decides either to advertise its product in a local newspaper or it decides not to. As economists we are interested in explaining why particular choices are made, and what factors enter into the decision process. We also want to know *how much* each factor affects the outcome and how to predict outcomes. Such questions lead us to the problem of constructing a statistical model of binary, either-or, choices. Such choices can be represented by a binary (indicator) variable that takes the value 1 if one outcome is chosen and the value 0 otherwise. The binary variable describing a choice is the dependent variable rather than an independent variable. This fact affects our choice of a statistical model.

The list of economic applications in which choice models may be useful is a long one. These models are useful in any economic setting in which an agent must choose one of two alternatives. Examples include the following:

- An economic model explaining why some individuals take a second or third job and engage in "moonlighting."
- An economic model of why some legislators in the U.S. House of Representatives vote for a particular bill and others do not.
- An economic model explaining why some loan applications are accepted and others are not at a large metropolitan bank.
- An economic model explaining why some individuals vote for increased spending in a school board election and others vote against.
- An economic model explaining why some female college students decide to study engineering and others do not.

This list illustrates the great variety of circumstances in which a model of binary choice may be used. In each case, an economic decision-maker chooses between two mutually exclusive outcomes.

The key feature of **binary choice models** is the nature of the outcome variable. It is an indicator variable representing the choice between two alternatives. We represent the ith individual's choice as

$$y_i = \begin{cases} 1 & \text{alternative one is chosen} \\ 0 & \text{alternative two is chosen} \end{cases} \tag{16.1}$$

Individuals make choices to maximize their utility, or well-being, and we economists would like to understand the process. What are the important factors leading to the choice and how much weight is given to each? Can we predict what the choice will be? These questions lead us to

consider how individuals make their decisions, how to build an econometric model of the choice process, and how to model the probability of choosing one alternative or the other.

It's always best to start at the beginning. Unlike the outcome of a game of chance, such as flipping a coin and observing a head or a tail, the probability that alternative one will be chosen varies from individual to individual, and the probability depends on many factors, describing the individual and the characteristics of the alternatives. As in a regression model, let these factors be denoted $\mathbf{x}_i = (x_{i1} = 1, x_{i2}, \ldots, x_{iK})$. Then the conditional probability that the ith individual chooses alternative one is $P(y_i = 1|\mathbf{x}_i) = p(\mathbf{x}_i)$, where $p(\mathbf{x}_i)$ is a function of the factors $\mathbf{x}_i$, and because it is a probability, $0 \le p(\mathbf{x}_i) \le 1$. The conditional probability of choosing alternative two is $P(y_i = 0|\mathbf{x}_i) = 1 - p(\mathbf{x}_i)$. We can represent the conditional probability function for the random variable y_i in equation (16.1) as

$$f(y_i|\mathbf{x}_i) = p(\mathbf{x}_i)^{y_i}\left[1 - p(\mathbf{x}_i)\right]^{1-y_i} \qquad y_i = 0, 1 \tag{16.2}$$

Then $P(y_i = 1|\mathbf{x}_i) = f(1|\mathbf{x}_i) = p(\mathbf{x}_i)$ and $P(y_i = 0|\mathbf{x}_i) = f(0|\mathbf{x}_i) = 1 - p(\mathbf{x}_i)$. The standard models of probabilistic choice are simply alternative ways of representing, or approximating, $P(y_i = 1|\mathbf{x}_i) = p(\mathbf{x}_i)$.

EXAMPLE 16.1 | A Transportation Problem

An important problem in transportation economics is explaining an individual's choice between driving (private transportation) and taking the bus (public transportation) when commuting to work, assuming, for simplicity, that these are the only two alternatives. We can imagine many factors that affect the choice, including an individual's characteristics, such as age, income, and sex; the characteristics of their automobile, such as its reliability, comfort, and fuel economy; the characteristics of the public transportation, such as reliability, cost, and safety. In our example, we will focus on a single factor, commuting time. Define the explanatory variable

x_i = (commuting time by bus
− commuting time by car, for the ith individual)

A priori we expect that as x_i increases, and commuting time by bus increases relative to commuting time by car, and holding all else constant, an individual would be more inclined to drive. Suppose that alternative one is driving to work, $y_i = 1$, and alternative two is taking public transportation, $y_i = 0$. Then the probability that the ith individual drives to work is $P(y_i = 1|x_i) = p(x_i)$. Our reasoning suggests that there is a positive relationship between the difference in commuting time and the probability that an individual will drive to work. Using data on individuals and their choices, we will obtain estimates of how much increases in commuting time by bus relative to driving will affect the probability that an individual will drive. Using the estimates, we can predict the choice of an individual when the commuting time by bus is, for example, 20 minutes longer than the commuting time by car. We will also develop methods for testing hypotheses about the nature of the relationship, such as testing whether the difference in commuting time is a statistically significant factor in the decision.

16.1.1 The Linear Probability Model

We discussed the **linear probability model** in Sections 7.4 and 8.7. It is a regression model that arises straightforwardly from the definition of expected value. Using the probability model in (16.2),

$$E(y_i|\mathbf{x}_i) = \sum_{y_i=0}^{1} y_i f(y_i|\mathbf{x}_i) = 0 \times f(0|\mathbf{x}_i) + 1 \times f(1|\mathbf{x}_i) = p(\mathbf{x}_i) \tag{16.3}$$

The population average outcome, the average choice, is the probability that the first alternative is chosen. It is natural to specify a linear regression model for the probability

$$p(\mathbf{x}_i) = E(y_i|\mathbf{x}_i) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} \tag{16.4}$$

Let the random error e_i account for the difference between the observed outcome y_i and the conditional mean $E(y_i|\mathbf{x}_i)$,

$$e_i = y_i - E(y_i|\mathbf{x}_i) \tag{16.5}$$

Then

$$y_i = E(y_i|\mathbf{x}_i) + e_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i \tag{16.6}$$

If $E(e_i|\mathbf{x}_i) = 0$, then the least squares estimator of the parameters is unbiased, or if random error e_i is uncorrelated with $\mathbf{x}_i = (x_{i1} = 1, x_{i2}, \ldots, x_{iK})$, then the least squares estimator is consistent. These are the usual OLS properties.

For a continuous variable x_{ik}, the **marginal effect** is

$$\partial E(y_i|\mathbf{x}_i)/\partial x_{ik} = \beta_k \tag{16.7}$$

Here is where some difficulty enters. Suppose that $\beta_k > 0$. Increasing x_{ik} by one unit increases $p(\mathbf{x}_i)$, the probability of alternative one being chosen, by a constant amount β_k. This puts us into the uncomfortable position of concluding that the probability can become one, or greater than one, if x_{ik} becomes large enough. Similarly, if $\beta_k < 0$ then the probability of alternative one being chosen can become negative if x_{ik} becomes large enough. These are the logical inconsistencies in the linear probability model. It is because of these difficulties that we develop alternatives to the linear probability model in Section 16.2. Nevertheless, the regression model approach is very familiar, and by now easy, and it is a useful approximation tool for the purpose of estimating marginal effects in nonextreme cases.

Apart from the logical problem noted above, which is important, there are two other more minor consequences of using the linear probability model. First, since y_i takes only two values, one and zero, it must be true that $\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$ takes the same two values. If $y_i = 1$, then it follows that $\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i = 1$, so that

$$e_i = 1 - (\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})$$

If $y_i = 0$, then $\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i = 0$ so that

$$e_i = -(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})$$

This seems very odd—the random error that accounts for all omitted factors and other specification errors takes only two values. This is the result of imposing a linear regression structure on a choice problem in which the outcome is binary, one or zero.

Secondly, the conditional variance in the random error is

$$\text{var}(e_i|\mathbf{x}_i) = p(\mathbf{x}_i)\left[1 - p(\mathbf{x}_i)\right] = \sigma_i^2 \tag{16.8}$$

and is necessarily heteroskedastic. When estimating the linear probability model, this feature must be recognized. When using the OLS estimator, we must at least use heteroskedasticity robust standard errors. Alternatively use the FGLS, **feasible generalized least squares**, estimation methodology discussed in Section 8.6.

EXAMPLE 16.2 | A Transportation Problem: The Linear Probability Model

Ben-Akiva and Lerman[1] have sample data on automobile and public transportation travel times and the alternative chosen for $N = 21$ individuals in the data file *transport*. The variable *AUTO* is an indicator variable taking the value one if automobile transportation is chosen and is zero if public transportation is chosen,

$$AUTO = \begin{cases} 1 & \text{auto is chosen} \\ 0 & \text{public transportation (bus) is chosen} \end{cases}$$

[1] (1985) *Discrete Choice Analysis*, MIT Press.

The variables *AUTOTIME* and *BUSTIME* are minutes of commuting time. The explanatory variable we consider is $DTIME = (BUSTIME - AUTOTIME) \div 10$, which is the commuting time differential in 10-minute increments. The linear probability model is $AUTO_i = \beta_1 + \beta_2 DTIME_i + e_i$. The OLS fitted model, with heteroskedasticity robust standard errors, is

$$\widehat{AUTO}_i = 0.4848 + 0.0703 DTIME_i \quad R^2 = 0.61$$
$$\text{(robse)} \quad (0.0712) \quad (0.0085)$$

We estimate that if travel times by public transportation and automobile are equal, so that $DTIME = 0$, then the probability of a person choosing automobile travel is 0.4848, close to 50–50, with a 95% interval estimate of [0.34, 0.63]. We estimate that, holding all else constant, an increase of 10 minutes in the difference in travel time, increasing public transportation travel time relative to automobile travel time, increases the probability of choosing automobile travel by 0.07, with a 95% interval estimate of [0.0525, 0.0881], which seems relatively precise. In truth, any judgment about precision depends on the use to which the results will be put. The fitted model can be used to estimate the probability of automobile travel for any commuting time differential. For example, if $DTIME = 1$, a 10-minute longer commute by public transportation, we estimate the probability of automobile travel to be $\widehat{AUTO}_i = 0.4848 + 0.0703(1) = 0.5551$.

How well does the model fit the data? The $R^2 = 0.61$ suggests that 61% of the variation in the outcome variable is explained by the model. With probability models, we can examine how well the model predicts the outcomes. Let's predict the choice using a probability threshold of 0.50. That is, if $\widehat{AUTO}_i \geq 0.50$ we predict that a person will drive to work, and otherwise, we predict that a person will use public transportation. In the sample of 21 individuals, 10 drove to work and 11 used public transportation. Using the classification rule, we successfully predict 9 of the 10 drivers, and 10 of the 11 bus riders. That is 19 successful predictions out of the 21 cases. Looking at individual estimated probabilities of driving, we find three negative values. If the commute is 69 minutes or less by public transportation, then the estimated probability of driving is zero or negative. If commuting time is 73 minutes or more by public transportation, then the estimated probability of driving is one or greater.

16.2 Modeling Binary Choices

It is the probability of choosing one alternative or the other that is the key concept when modeling binary choice. Probabilities must be between zero and one, and the flaw in the linear probability model in Section 16.1 is that it does not impose this constraint. We now turn to two nonlinear models for binary choices, the **probit model** and the **logit model**, which ensure that choice probabilities remain between zero and one. To keep the choice probability $p(\mathbf{x}_i)$ within the interval (0, 1), a nonlinear S-shaped "sigmoid" curve can be used. In Figure 16.1(a), one such curve is illustrated for the case of a single explanatory variable, x. If, for example, $\beta_2 > 0$, then, as x increases, and, $\beta_1 + \beta_2 x$ increases, the probability curve rises rapidly at first, and then begins to increase at a decreasing rate, keeping the probability less than one no matter how large x becomes. In the other direction, the probability approaches but never reaches zero. The *slope* of the probability curve, $dp(x_i)/dx$, is the change in probability given a unit change in x. It is the **marginal effect** and, unlike in the linear probability model, the slope is not constant.

The curve shown in Figure 16.1(a) is the cumulative distribution function (*cdf*) of the standard normal random variable. This choice of the S-curve leads to a model called **probit**. Any *cdf* function for a continuous random variable will work, and many have been tried over the years. These days the main competitor to the standard normal *cdf* is the *cdf* of a **logistic random variable**, leading to a model called **logit**. In binary choice cases, probit and logit provide very similar inferences. Economists tend to choose probit rather than logit in individual choice applications because it follows logically from utility maximizing behavior and random utility models (RUMs) under the assumption that the unobserved components of utility for the two alternatives are jointly normal. To obtain a logit model within this framework, the unobserved components of utility for the two alternatives must be statistically independent and have an unusual probability density function (*pdf*).[2] However, the logit model is widely used in many disciplines and leads to very convenient generalizations. We will discuss both the probit and logit models.

[2]For more on RUM and choice models, see Appendix 16B. Also Kenneth Train (2009) *Discrete Choice Methods with Simulation, Second Edition*, Cambridge University Press.

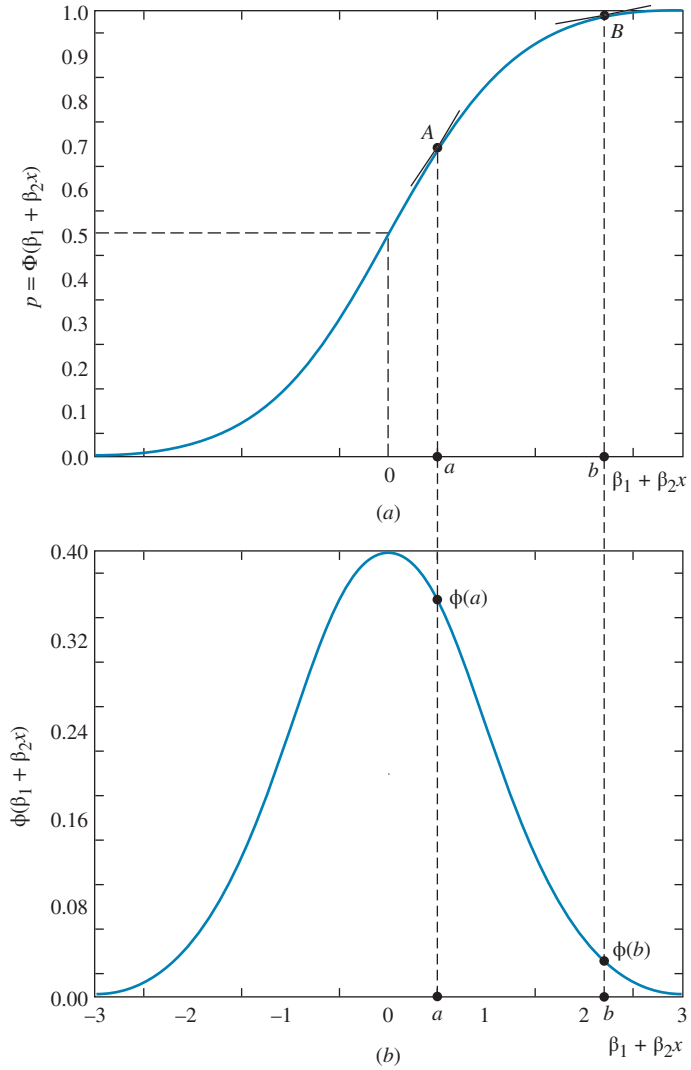

FIGURE 16.1 **(*a*) Standard normal *cdf*; (*b*) standard normal *pdf*.**

16.2.1 The Probit Model for Binary Choice

As noted above, the probit model is based on the standard normal *cdf*. If Z is a standard normal random variable, then its *pdf* is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-0.5z^2} \quad -\infty < z < \infty \tag{16.9a}$$

The *cdf* of the standard normal distribution is

$$\Phi(z) = P[Z \leq z] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-0.5u^2} du \tag{16.9b}$$

This integral expression is the probability that a standard normal random variable falls to the left of point z. In geometric terms, it is the area under the standard normal *pdf* to the left of z. The function $\Phi(z)$ is the *cdf* that we have worked with to compute normal probabilities.

The probit statistical model expresses the probability $p(\mathbf{x}_i)$ that alternative one is chosen, $y_i = 1$, to be

$$P(y_i = 1|\mathbf{x}_i) = p(\mathbf{x}_i) = P[Z \le \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}] = \Phi(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}) \quad (16.10)$$

where $\Phi(z)$ is the standard normal *cdf*. The probit model is said to be *nonlinear* because (16.10) is a nonlinear function of the parameters $\beta_1, \ldots, \beta_K$. If the parameters $\beta_1, \ldots, \beta_K$ were known, we could use (16.10) to find the probability that alternative one is chosen for any set of predictor values $\mathbf{x}_i = (x_{i1} = 1, x_{i2}, \ldots, x_{iK})$. Because these parameters are not known we will estimate them.

16.2.2 Interpreting the Probit Model

Interpreting the probit model requires a bit of work. How we proceed to measure the impact of any one variable x_{ik} depends on whether it is continuous or discrete, like an indicator variable. When an explanatory variable is continuous, we can examine the marginal effect of a change in its value on the probability $p(\mathbf{x}_i)$. When the explanatory variable is an indicator variable, we can calculate the difference in the probability $p(\mathbf{x}_i)$ associated with $x_{ik} = 0$ and $x_{ik} = 1$. In both of these cases, we must deal with the fact that the magnitudes of the effects depend not only on the parameter values, $\beta_1, \ldots, \beta_K$, but also on the values of the explanatory variables, $\mathbf{x}_i = (x_{i1} = 1, x_{i2}, \ldots, x_{iK})$. We will examine these cases separately.

Marginal Effect of a Continuous Explanatory Variable

If x_k is a continuous variable then we can calculate the marginal effect by finding the derivative of (16.10). The marginal effect is

$$\frac{\partial p(\mathbf{x}_i)}{\partial x_{ik}} = \frac{\partial \Phi(t_i)}{\partial t_i} \cdot \frac{\partial t_i}{\partial x_{ik}} = \phi(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})\beta_k \quad (16.11)$$

where $t_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$ and $\phi(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})$ is the standard normal *pdf* evaluated at $\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$. To obtain this result, we have used the chain rule of differentiation (see Derivative Rule 9 in Appendix A.3.1). Note that the marginal effect includes the *pdf* of the standard normal random variable, $\phi(\bullet)$.

To simplify the algebra, suppose that there is a single continuous explanatory variable, x. Then, the probit probability model is $p(x_i) = P[Z \le \beta_1 + \beta_2 x_i] = \Phi(\beta_1 + \beta_2 x_i)$. Assuming $\beta_2 > 0$, this is the equation of the sigmoid S-shaped curve in Figure 16.1(a). At point A in Figure 16.1(a), where $\beta_1 + \beta_2 x_i = a$, the marginal effect of a change in x on the probability is the slope of the tangent line. At point B in Figure 16.1(a), where $\beta_1 + \beta_2 x_i = b$ and the probability $\Phi(b)$ is larger, the marginal effect is smaller, which it must be to keep the probability function less than one as x increases.

The equation of the marginal effect $dp(x_i)/dx_i = \phi(\beta_1 + \beta_2 x_i)\beta_2$ is the slope of the probability function at the point $\beta_1 + \beta_2 x_i$. The *pdf* $\phi(\beta_1 + \beta_2 x_i)$, plotted in Figure 16.1(b), appears in the marginal effect because of its relationship to the cumulative distribution function $\Phi(\beta_1 + \beta_2 x_i)$. As noted in (16.9), the *cdf* is the integral of the *pdf*, and it follows that the *pdf* is the derivative of the *cdf* in (16.11). The marginal effect at point A is larger because $\phi(a) > \phi(b)$. The marginal effect equation, $dp(x_i)/dx_i = \phi(\beta_1 + \beta_2 x_i)\beta_2$, has the following implications.

1. Since $\phi(\beta_1 + \beta_2 x_i)$ is a *pdf* its value is always *positive*. Consequently, the sign of $dp(x_i)/dx_i$ is determined by the sign of β_2. If $\beta_2 > 0$ then $dp(x_i)/dx_i > 0$, and if $\beta_2 < 0$ then $dp(x_i)/dx_i < 0$.
2. As x_i changes the value of the function $\phi(\beta_1 + \beta_2 x_i)$ changes. The standard normal *pdf* reaches its maximum when $\beta_1 + \beta_2 x_i = 0$. In this case $p(x_i) = P[Z \le 0] = \Phi(0) = 0.5$; the alternatives one and two are equally likely to be chosen. It makes sense that in this case the effect of a change in x_i has its greatest effect, the marginal effect is largest, because the individual is "on the borderline."

3. On the other hand, if $\beta_1 + \beta_2 x_i$ is large, say near 3, then the probability that the individual chooses alternative one, $p(x_i)$, is very large and close to 1. In this case, a change in x_i will have relatively little effect since $\phi(\beta_1 + \beta_2 x_i)$ is nearly 0. The same is true if $\beta_1 + \beta_2 x_i$ is a large negative value, say near -3. These results are consistent with the notion that if an individual is "set" in their ways, with $p(x_i)$ near 0 or 1, the effect of a small change in x_i will be negligible.

Discrete Change Effect of an Indicator Explanatory Variable The marginal effect in (16.11) is valid only if the explanatory variable x_k is continuous. If x_k is a discrete variable, such as an indicator variable for an individual's sex, then the derivative in (16.11) cannot be used. Instead we can compute the discrete change in probability effect of x_k changing from zero to one,

$$\Delta p(\mathbf{x}_i) = p(\mathbf{x}_i | x_{ki} = 1) - p(\mathbf{x}_i | x_{ki} = 0) \tag{16.12a}$$

To simplify the notation, suppose $p(\mathbf{x}_i) = \Phi(\beta_1 + \beta_2 x_{i2} + \delta D_i)$ where D_i is an indicator variable. The difference in the probability of choosing alternative one given $D_i = 1$ as compared to when $D_i = 0$ is

$$\Delta p(\mathbf{x}_i) = p(\mathbf{x}_i | D_i = 1) - p(\mathbf{x}_i | D_i = 0) = \Phi(\beta_1 + \beta_2 x_{i2} + \delta) - \Phi(\beta_1 + \beta_2 x_{i2}) \tag{16.12b}$$

The change can be positive or negative, depending on the sign of the parameter δ. If $\delta > 0$, then there is an increase in the probability of choosing alternative one. If $\delta < 0$, then the probability of choosing alternative one decreases. Note that the magnitude of the effect depends on the sign and magnitude of the parameter δ but also on the values of the other explanatory variables and their parameters.

Discrete Change Effect of any Explanatory Variable The use of the discrete change approach is not limited to indicator variables. It can also be used for an explanatory variable that is a count, such as $x_3 = 0, 1, 2, \ldots$ Suppose that y_i is an individual's health outcome, such as whether their blood pressure reading is too high, or not, and x_3 is the person's number of periods of exercise per week. We might be interested in the change in the probability of high blood pressure of increasing from one workout per week to three workouts per week. The discrete change approach can also be used for a continuous variable. Suppose that x_3 is the number of minutes of exercise per week. We might be interested in the change in the probability of high blood pressure of increasing the number of minutes of exercise from 90 to 120 per week. In general, suppose that we are interested in the change $x_{i3} = c$ to $x_{i3} = c + \delta$. Then the discrete change in probability is

$$\begin{aligned} \Delta p(\mathbf{x}_i) &= p(\mathbf{x}_i | x_{i3} = c + \delta) - p(\mathbf{x}_i | x_{i3} = c) \\ &= \Phi(\beta_1 + \beta_2 x_{i2} + \beta_3 c + \beta_3 \delta) - \Phi(\beta_1 + \beta_2 x_{i2} + \beta_3 c) \end{aligned} \tag{16.12c}$$

Because the model is nonlinear, the values of c and δ will affect the change in probability.

Estimating Marginal and Discrete Change Effects In order to estimate the marginal effect in (16.11) or the discrete change effect (16.12), we must have parameter estimates, $\tilde{\beta}_1, \ldots, \tilde{\beta}_K$. The estimates are obtained by **maximum likelihood estimation**, which we will discuss in Section 16.2.3. For the moment, suppose that we have these estimates. In practice, they are obtained just like OLS estimates, with a simple computer command. Focus now on the possible values of the explanatory variables $\mathbf{x}_i = (x_{i1} = 1, x_{i2}, \ldots, x_{iK})$. There are several options for reporting marginal effects:

1. **Marginal effect at means (MEM)**[3] One choice is $\bar{\mathbf{x}} = (1, \bar{x}_2, \ldots, \bar{x}_K)$ where $\bar{x}_k$ is the sample mean of the values for the kth explanatory variable. There are two points of interest here. First, unlike the linear regression model, the fitted probit model does not pass through the "point of the means," so choosing the point $\bar{\mathbf{x}}$ has no special significance. Second, for an indicator variable, such as $x_{ik} = 1$ for females and $x_{ik} = 0$ for males, the average value $\bar{x}_k$ is the fraction of the sample that is female. Instead of a 1 or a 0, we might have $\bar{x}_k = 0.53$, indicating that 53% of the sample is female.
2. **Marginal effect at a representative value (MER)** Another possibility is to choose the values of $\mathbf{x}_i = (x_{i1} = 1, x_{i2}, \ldots, x_{iK})$ to reflect a particular scenario, a set of values that tell a "story" about the results. That is, suppose that x_{i2} is a person's years of schooling, x_{i3} is the person's sex (1 = female), and x_{i4} is their income ($1000s). We might specify $\mathbf{x}_i = (1, x_{i2} = 14, x_{i3} = 1, x_{i4} = 100)$, representing a female with 14 years of schooling and $100,000 income. This approach is more work because the representative values for the variables should have some meaning within the context of the research problem, but in some sense, it is also the most meaningful when describing the results. Of course, some of the variables' representative values might be variable means, medians, or quartiles.
3. **Average marginal effect (AME)** A third option is to calculate the sample average marginal effect. For a continuous variable, the AME is the sample average of (16.11) evaluated at each sample observation,

$$\text{AME}(x_k) = N^{-1}\sum_{i=1}^{N} \partial p(\mathbf{x}_i)/\partial x_{ik} = \beta_k \sum_{i=1}^{N} \phi(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})/N \tag{16.13a}$$

   For a discrete variable, we average the differences in (16.12a). In the simple model $p(\mathbf{x}_i) = \Phi(\beta_1 + \beta_2 x_{i2} + \delta D_i)$, this average is

$$\begin{aligned}\text{AME}(D) &= N^{-1}\sum_{i=1}^{N} \Delta p(\mathbf{x}_i) \\ &= \sum_{i=1}^{N} \Phi(\beta_1 + \beta_2 x_{i2} + \delta)/N - \sum_{i=1}^{N} \Phi(\beta_1 + \beta_2 x_{i2})/N\end{aligned} \tag{16.13b}$$

   If, for example, $D_i = 1$ if a person is female, then the first term $\sum_{i=1}^{N} \Phi(\beta_1 + \beta_2 x_{i2} + \delta)/N$ assigns the female sex to everyone in the sample, and the second term $\sum_{i=1}^{N} \Phi(\beta_1 + \beta_2 x_{i2})/N$ assigns the male sex to everyone in the sample. There are two advantages to computing the AME. First, it relieves us of having to make a choice about what to do. Second, relying on a "law of large numbers" argument, the sample average marginal, or discrete change, effect can be thought of as estimating the population average response to a change in a variable.
4. **A Histogram** A fourth option is to examine a histogram of the marginal effects computed for each $\mathbf{x}_i$ in the sample.

Predicting Choice with a Probit Model Last but not least, we can use the probit model to not only estimate the probability that an individual chooses one alternative or another but also predict the choice they will make. The probability model is $p(\mathbf{x}_i) = \Phi(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})$. Given values of the explanatory variables, and parameter estimates, $\tilde{\beta}_1, \ldots, \tilde{\beta}_K$, we can estimate the probability that an individual will choose alternative one as $\tilde{p}(\mathbf{x}_i) = \Phi(\tilde{\beta}_1 + \tilde{\beta}_2 x_{i2} + \cdots + \tilde{\beta}_K x_{iK})$. By comparing the estimated probability to a suitable threshold, τ, we can predict choice. The first threshold that comes to mind is 0.5. If we estimate the probability to be greater than or equal to 0.5 we predict $\tilde{y}_i = 1$, and if the estimated probability is less than 0.5, then we predict $\tilde{y}_i = 0$.

The threshold 0.5 is not necessarily the best threshold value to use. For example, suppose that we are the loan officer at a lending institution and must decide whether to give a loan to an

[3] We use the abbreviations MEM, MER, and AME, following Cameron and Trivedi (2010) *Microeconometrics Using Stata, Second Edition*, pp. 343–356.

applicant. Using data on previous borrowers, we can estimate a probit model for whether a loan was repaid on time, $y_i = 1$, or not, $y_i = 0$, as a function of borrower and loan characteristics. The fact is that most borrowers do repay their loans. If 90% of borrowers pay their loan back and if our applicant's estimated probability of repayment is 0.60, then that is a weak endorsement for giving a loan. For a lender, choosing the profit maximizing threshold τ^* is not an easy task. The correct decision is to give a loan to someone who will repay it and to not give a loan to someone who won't repay it. Lenders must weigh two types of incorrect decisions. If the lender gives a loan to someone who does not repay it, then there are costs (losses) associated with collecting the loan; further correspondence, legal action, and so on. If the lender does not give a loan to someone who would repay, there are foregone profits, opportunity costs. Lenders must compare the costs of these errors. If the threshold is raised, there are increased foregone profits; if the threshold is lowered, there are more collection costs. There is no one universal threshold that is suitable for every type of situation.

16.2.3 Maximum Likelihood Estimation of the Probit Model

The maximum likelihood estimation (MLE) methodology is discussed in Appendix C.8. Maximum likelihood estimation is based on a principle that is an alternative to the least squares principle or to other principles such as generalized least squares, or the method of moments, although it sometimes yields the same results. The MLE methodology is well suited to models we discuss in this chapter, including the probit binary choice model. Under some suitable conditions, maximum likelihood estimators have properties that are valid in large samples. If $\tilde{\beta}_k$ is the maximum likelihood estimator of the parameter β_k, then it is a consistent estimator, plim $\tilde{\beta}_k = \beta_k$, and it has an approximate normal distribution in large samples, $\tilde{\beta}_k \overset{a}{\sim} N\left[\beta_k, \text{var}(\tilde{\beta}_k)\right]$. The estimator variance is known (though complicated algebraically) and can be consistently estimated in several ways. If $\widehat{\text{var}}(\tilde{\beta}_k)$ is a consistent estimator of $\text{var}(\tilde{\beta}_k)$, then we can calculate a standard error, $\text{se}(\tilde{\beta}_k) = \sqrt{\widehat{\text{var}}(\tilde{\beta}_k)}$. Using the standard error, we can compute interval estimates, $\tilde{\beta}_k \pm z_{(1-\alpha/2)}\text{se}(\tilde{\beta}_k)$, carry out "$t$-tests," and so on in the usual way. All of these theoretical results are illustrated in Appendix C.8. In Example 16.3, we present the essence of the maximum likelihood estimation method.

EXAMPLE 16.3 | Probit Maximum Likelihood: A Small Example

We first illustrate the idea of maximum likelihood estimation in an abbreviated version of the transportation choice model from Examples 16.1 and 16.2. Suppose that we randomly select three individuals and observe that the first two drive to work and the third takes the bus; $y_1 = 1$, $y_2 = 1$, $y_3 = 0$. Furthermore, suppose that the differences in commuting times for these individuals, in 10-minute units, are $x_1 = 1.5$, $x_2 = 0.6$, $x_3 = 0.7$. What is the joint probability of observing $y_1 = 1$, $y_2 = 1$, $y_3 = 0$? The probability function for y_i is given by (16.2), which we now combine with the probit model (16.10) to obtain

$$f(y_i|x_i) = \left[\Phi(\beta_1 + \beta_2 x_i)\right]^{y_i}\left[1 - \Phi(\beta_1 + \beta_2 x_i)\right]^{1-y_i}, \quad y_i = 0, 1$$

If the three individuals are independently drawn, then the joint *pdf* for y_1, y_2, and y_3 is the product of the marginal probability functions:

$$f(y_1, y_2, y_3|x_1, x_2, x_3) = f(y_1|x_1)\, f(y_2|x_2)\, f(y_3|x_3)$$

Consequently, the probability of observing $y_1 = 1$, $y_2 = 1$, and $y_3 = 0$ is

$$P(y_1 = 1, y_2 = 1, y_3 = 0|x_1, x_2, x_3) = f(1, 1, 0|x_1, x_2, x_3) = f(1|x_1)\, f(1|x_2)\, f(0|x_3)$$

Substituting the y and x values, we have

$$\begin{aligned} P(y_1 = 1, y_2 = 1, y_3 = 0|x_1, x_2, x_3) &= \Phi\left[\beta_1 + \beta_2(1.5)\right] \times \Phi\left[\beta_1 + \beta_2(0.6)\right] \\ &\quad \times \left\{1 - \Phi\left[\beta_1 + \beta_2(0.7)\right]\right\} \\ &= L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x}) \end{aligned} \tag{16.14}$$

In statistics, the function (16.14), which gives us the probability of observing the sample data, is called the **likelihood function**. The notation $L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x})$ indicates that the likelihood function is a function of the unknown parameters once we are given the data. It is intuitively reasonable to use as estimates those values $\tilde{\beta}_1$ and $\tilde{\beta}_2$ that maximize the probability,

or likelihood, of the observed outcome. Unfortunately, for the probit model, there are no formulas that give us the values for $\tilde{\beta}_1$ and $\tilde{\beta}_2$ as there are in least squares estimation of the linear regression model. Consequently, we must use the computer and techniques from numerical analysis to find the values $\tilde{\beta}_1$ and $\tilde{\beta}_2$ that maximize $L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x})$. In practice, instead of maximizing (16.14), we maximize the logarithm of (16.14), which is called the **log-likelihood function**

$$\begin{aligned}\ln L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x}) &= \ln\left\{\Phi[\beta_1 + \beta_2(1.5)] \times \Phi[\beta_1 + \beta_2(0.6)]\right. \\ &\quad \left.\times \left\{1 - \Phi[\beta_1 + \beta_2(0.7)]\right\}\right\} \\ &= \ln \Phi[\beta_1 + \beta_2(1.5)] + \ln\Phi[\beta_1 + \beta_2(0.6)] \\ &\quad + \ln\left\{1 - \Phi[\beta_1 + \beta_2(0.7)]\right\} \qquad (16.15)\end{aligned}$$

On the surface, this appears to be a difficult task, because $\Phi(z)$ from (16.9) is such a complicated function. As it turns out, however, using a computer to maximize (16.15) is a relatively easy process.

The maximization of the log-likelihood function $\ln L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x})$ is easier than the maximization of (16.14), because it is a sum of terms and not a product of terms. The logarithm is a nondecreasing, or monotonic, function so that the maximum values of the two functions $L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x})$ and $\ln L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x})$ occur at the same values of β_1 and β_2, namely, $\tilde{\beta}_1$ and $\tilde{\beta}_2$. The value of the log-likelihood function (16.15) evaluated at the maximizing values $\tilde{\beta}_1$ and $\tilde{\beta}_2$ is very useful for hypothesis testing, which is discussed in Sections 16.2.4 and 16.2.5. Using econometric software, we find that the parameter values that maximize (16.15) are $\tilde{\beta}_1 = -1.1525$ and $\tilde{\beta}_2 = 0.1892$. These values maximize the log-likelihood function, $\ln L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x})$, and also maximize the likelihood function $L(\beta_1, \beta_2|\mathbf{y}, \mathbf{x})$. They are the *maximum likelihood estimates*. Any other values of the parameters that we might try will yield a lower value of the log-likelihood function. Plugging these values into (16.15), we obtain the value of the log-likelihood function evaluated at the maximum likelihood estimates, which is $L(\tilde{\beta}_1, \tilde{\beta}_2|\mathbf{y}, \mathbf{x}) = -1.5940$.

An interesting feature of the maximum likelihood estimation procedure is that while its properties in small samples are not known, we can show that in large samples the maximum likelihood estimator is normally distributed, consistent and *best*, in the sense that no competing estimator has smaller variance. The properties of maximum likelihood estimators are fully discussed in Appendix C.8.

We have used only three observations in the numerical illustration above for demonstration purposes only. In practice, such maximum likelihood estimation procedures should only be used when large samples are available. In the following section, we present another simple example that will demonstrate more aspects of the probit choice model.

EXAMPLE 16.4 | The Transportation Data: Probit

In Example 16.2, we estimated a linear probability model using the transportation data, *transport*. In this example, we carry out probit estimation. The probit model is $P(AUTO = 1) = \Phi(\beta_1 + \beta_2 DTIME)$. The maximum likelihood estimates of the parameters are

$$\begin{aligned}\tilde{\beta}_1 + \tilde{\beta}_2 DTIME &= -0.0644 + 0.3000 DTIME \\ \text{(se)} \qquad & \quad (0.3992) \quad (0.1029)\end{aligned}$$

The values in parentheses below the parameter estimates are estimated standard errors that are valid in large samples. These standard errors can be used to carry out hypothesis tests and construct interval estimates in the usual way, with the qualification that they are valid in large samples. The negative sign of $\tilde{\beta}_1$ implies that when commuting times via bus and auto are equal so $DTIME = 0$, individuals have a bias against driving to work, relative to public transportation, The estimated probability of a person choosing to drive to work when $DTIME = 0$ is $\hat{P}(AUTO = 1|DTIME = 0) = \Phi(-0.0644) = 0.4743$. The positive sign of $\tilde{\beta}_2$ indicates that an increase in public transportation travel time, relative to auto travel time, increases the probability that an individual will choose to drive to work, and this coefficient is statistically significant.

Suppose that we wish to estimate the marginal effect of increasing public transportation time, given that travel via public transportation currently takes 20 minutes longer than auto travel. Using (16.11),

$$\begin{aligned}\widehat{\frac{dp}{dDTIME}} &= \phi(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME)\tilde{\beta}_2 \\ &= \phi(-0.0644 + 0.3000 \times 2)(0.3000) \\ &= \phi(0.5355)(0.3000) = 0.3456 \times 0.3000 = 0.1037\end{aligned}$$

For the probit probability model, an incremental (10-minute) increase in the travel time via public transportation increases the probability of travel via auto by approximately 0.1037, given that taking the bus already requires 20 minutes more travel time than driving.

The estimated parameters of the probit model can also be used to "predict" the behavior of an individual who must choose between auto and public transportation to travel to work. If an individual is faced with the situation that it takes 30 minutes longer to take public transportation than to drive to work, then the estimated probability that auto transportation will be selected is calculated using (16.12):

$$\hat{p} = \Phi(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME) = \Phi(-0.0644 + 0.3000 \times 3) = 0.7983$$

Since the estimated probability that the individual will choose to drive to work is 0.7983, which is greater than 0.5, we "predict" that when public transportation takes 30 minutes longer than driving to work, the individual will choose to drive.

EXAMPLE 16.5 | The Transportation Data: More Postestimation Analysis

In Example 16.4, we estimated the probit model for transportation choice and illustrated basic calculations. In this example, we carry out further, more advanced, postestimation analysis.

Marginal Effect at a Representative Value (MER)

The marginal effect of a change in the travel time differential is

$$\widehat{\frac{dp}{dDTIME}} = \phi(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME)\tilde{\beta}_2 = g(\tilde{\beta}_1, \tilde{\beta}_2)$$

The marginal effect is an estimator, since, given *DTIME*, it is a function of the estimators $\tilde{\beta}_1$ and $\tilde{\beta}_2$. The discussions of the "delta method" in Section 5.7.4 and Appendix 5B are relevant because the marginal effect is a *nonlinear* function of $\tilde{\beta}_1$ and $\tilde{\beta}_2$. The marginal effect estimator is consistent and asymptotically normal with a variance given by equation (5B.4). Using this result, we can test marginal effects or compute interval estimates for them. For example, if the time differential is currently 20 minutes, so that the representative value is $DTIME = 2$, the estimated marginal effect (MER) is 0.1037 and the estimated standard error of the marginal effect is 0.0326 using the delta method. Therefore, a 95% interval estimate of the marginal effect, using the *t*-critical value $t_{(0.975,19)} = 2.093$, is [0.0354, 0.1720]. This interval is fairly wide. Recall, however, that the maximum likelihood estimates are based on only 21 observations, which is a very small sample. The details of the calculation of the standard error are given in Appendix 16A.1.

Marginal Effect at the Mean (MEM)

If particular values of interest are difficult to identify, many researchers evaluate the marginal effect "at the means," MEM. In these data, the average time travel differential is $\overline{DTIME} = -0.1224$ (1.2 minutes), and for this value, the marginal effect of a 10-minute increase in the time travel differential is 0.1191. The slightly larger effect, compared to $DTIME = 2$, is consistent with the second point in the Section 16.2.1 discussion. When the mean difference in travel time is near zero, the effect of a change in travel time difference is greater. We can compute a standard error for this marginal effect just as we did for MER, if we treat $\overline{DTIME}$ as given.

Average Marginal Effect (AME)

Rather than evaluate the marginal effect at a specific value, or the mean value, we can compute the average of the marginal effects evaluated at each sample data point. That is,

$$\widehat{\text{AME}} = \frac{1}{N}\sum_{i=1}^{N}\phi(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME_i)\tilde{\beta}_2 = \frac{1}{N}\tilde{\beta}_2\sum_{i=1}^{N}\phi(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME_i)$$

The average marginal effect has become a popular alternative to computing the marginal effect at the mean as it summarizes the response of individuals in the sample to a change in the value of an explanatory variable. For the current example, $\widehat{\text{AME}} = 0.0484$, which is the sample average estimated increase in probability given a 10-minute increase in bus travel time relative to auto travel time. Because the estimated marginal effect is different for each individual in the sample, we are interested in not only its average value but also its variation in the sample. The sample standard deviation of $\phi(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME_i)\tilde{\beta}_2$ is 0.0365, and its minimum and maximum values are 0.0025 and 0.1153.

We can evaluate the standard error of the average marginal effect using the delta method. Recall that $\widehat{\text{AME}} = 0.0484$. Its standard error estimated using the delta method is 0.0034. Details of this calculation are given in Appendix 16A.2. A 95% interval estimate of the population average marginal effect, using the *t*-critical value, is [0.0413, 0.0556]. This is much narrower than the MER interval estimate because we are estimating a different quantity, namely $\text{AME} = \frac{1}{N}\beta_2\sum_{i=1}^{N}\phi(\beta_1 + \beta_2 DTIME_i)$.

Estimated Probability of Driving

The estimated probability that $AUTO = 1$ given that the commuting time difference is 30 minute is calculated as $\hat{p} = \Phi\left(\tilde{\beta}_1 + \tilde{\beta}_2 DTIME\right) = \Phi(-0.0644 + 0.3000 \times 3) = 0.7983$. Note that the predicted probability is a nonlinear function of the parameter estimates. Using the delta method, we can compute a standard error for the prediction and thus an interval estimate. The details of the calculation of the standard error are given in Appendix 16A.3. The calculated standard error is 0.1425, so that a 95% prediction interval, again using the t-critical value $t_{(0.975,19)} = 2.093$, is [0.5000, 1.0966]. Note that the upper endpoint of the interval is greater than 1, which means that some of the values are infeasible.

This example has been used to illustrate in a simple problem how probit works. In reality, estimating complicated models like probit and logit with as few observations as we are using, $N = 21$, is not a good idea. In fact, microeconometric models can have many more parameters and sometimes are estimated using very large data sets.

16.2.4 The Logit Model for Binary Choices

A frequently used alternative to the probit model for binary choice situations is the **logit** model. These models differ only in the particular S-shaped curve used to constrain probabilities to the [0, 1] interval. If L is a **logistic random variable**, then its *pdf* is

$$\lambda(l) = \frac{e^{-l}}{\left(1 + e^{-l}\right)^2}, \quad -\infty < l < \infty \tag{16.16}$$

The corresponding cumulative distribution function, unlike the normal distribution, has a closed-form expression, which makes analysis somewhat easier. The cumulative distribution function for a logistic random variable is

$$\Lambda(l) = p[L \le l] = \frac{1}{1 + e^{-l}} \tag{16.17}$$

In the logit model, if there is a single explanatory variable x, the probability $p(x)$ that the observed value y takes the value 1 is

$$p(x) = P\left[L \le \gamma_1 + \gamma_2 x\right] = \Lambda\left(\gamma_1 + \gamma_2 x\right) = \frac{1}{1 + e^{-(\gamma_1 + \gamma_2 x)}} \tag{16.18}$$

A more generally useful form of $p(x)$ is

$$p(x) = \frac{1}{1 + e^{-(\gamma_1 + \gamma_2 x)}} = \frac{\exp\left(\gamma_1 + \gamma_2 x\right)}{1 + \exp\left(\gamma_1 + \gamma_2 x\right)}$$

Then the probability that $y = 0$ is

$$1 - p(x) = \frac{1}{1 + \exp\left(\gamma_1 + \gamma_2 x\right)}$$

Represented in this way, the logit model can be extended to cases in which the choice is between more than two alternatives, as we will see in Section 16.3.

In maximum likelihood estimation of the logit model, the probability given in (16.18) is used to form the likelihood function (16.14) by inserting "Λ" for "Φ." To interpret the logit estimates, the equations (16.11) and (16.12) are still valid, using (16.16) instead of the normal *pdf*.

The shapes of the logistic and normal *pdfs* are somewhat different and maximum likelihood estimates of β_1 and β_2 will differ from γ_1 and γ_2. Roughly[4]

$$\tilde{\gamma}_{\text{Logit}} \cong 4\hat{\beta}_{\text{LPM}}$$
$$\tilde{\beta}_{\text{Probit}} \cong 2.5\hat{\beta}_{\text{LPM}}$$
$$\tilde{\gamma}_{\text{Logit}} \cong 1.6\tilde{\beta}_{\text{Probit}}$$

While the probit and logit parameter estimates differ, the marginal effects and predicted probabilities differ very little in most cases. In these expressions LPM denotes the linear probability model.

EXAMPLE 16.6 | An Empirical Example from Marketing

In Section 7.4.1, we introduced the example of a linear probability model for the choice between Coke and Pepsi. Here, we compare the linear probability model to the probit and logit models for this binary choice. The outcome variable is *COKE*

$$COKE = \begin{cases} 1 & \text{if Coke is chosen} \\ 0 & \text{if Pepsi is chosen} \end{cases}$$

The expected value of this variable is $E(COKE|\mathbf{x}) = p_{COKE}$ = probability that Coke is chosen. As explanatory variables, $\mathbf{x}$, we use the relative price of Coke to Pepsi (*PRATIO*), as well as *DISP_COKE* and *DISP_PEPSI*, which are indicator variables taking the value 1 if the respective store display is present and 0 if it is not present. We anticipate that the presence of a Coke display will increase the probability of a Coke purchase, and the presence of a Pepsi display will decrease the probability of a Coke purchase.

The data file *coke* contains "scanner" data on 1140 individuals who purchased Coke or Pepsi. The linear probability, probit, and logit models for the choice are

$$\begin{aligned} p_{COKE} &= E(COKE|\mathbf{x}) \\ &= \alpha_1 + \alpha_2 PRATIO + \alpha_3 DISP\_COKE \\ &\quad + \alpha_4 DISP\_PEPSI \\ p_{COKE} &= E(COKE|\mathbf{x}) \\ &= \Phi(\beta_1 + \beta_2 PRATIO + \beta_3 DISP\_COKE \\ &\quad + \beta_4 DISP\_PEPSI) \\ p_{COKE} &= E(COKE|\mathbf{x}) \\ &= \Lambda(\gamma_1 + \gamma_2 PRATIO + \gamma_3 DISP\_COKE \\ &\quad + \gamma_4 DISP\_PEPSI) \end{aligned}$$

We have given the choice model parameters different symbols to emphasize that the parameters have different meanings. The estimates are given in Table 16.1.

The parameters and their estimates vary across the models and no direct comparison is very useful. More relevant, however, is the comparison of the estimated probabilities and marginal effects implied by the alternative models.

TABLE 16.1 Coke-Pepsi Choice Models

	LPM	Probit	Logit
C	0.8902	1.1081	1.9230
	(0.0653)	(0.1900)	(0.3258)
PRATIO	−0.4009	−1.1460	−1.9957
	(0.0604)	(0.1809)	(0.3146)
DISP_COKE	0.0772	0.2172	0.3516
	(0.0339)	(0.0966)	(0.1585)
DISP_PEPSI	−0.1657	−0.4473	−0.7310
	(0.0344)	(0.1014)	(0.1678)

Standard errors in parentheses (White robust se for LPM)

Estimated probabilities at representative values Suppose that *PRATIO* = 1.1, indicating that the price of Coke is 10% higher than the price of Pepsi, and no store displays are present. Using the linear probability model, the estimated probability of Coke choice is 0.4493 with standard error 0.0202. Using probit, the estimated probability is 0.4394 with standard error 0.0218, and for logit, the estimated probability is 0.4323 with standard error 0.0224.

Average marginal effects (AME) In the linear probability model, the estimated marginal effect of *PRATIO* is −0.4009. This does not depend on the values of the variables. For the probit model, the average marginal effect of *PRATIO* is −0.4097 with standard error 0.0616, and for the logit model, the average marginal effect of *PRATIO* is −0.4333 with standard error 0.0639. In this example, the average marginal effect from the probit model is not too different from that implied by the linear probability model.

[4]T. Amemiya (1981) "Qualitative response models: A Survey," *Journal of Economic Literature*, 19, pp. 1483–1536, or A. Colin Cameron and Pravin K. Trivedi (2010) *Microeconometrics Using Stata: Revised Edition*, Stata Press, p. 465.

Marginal effect at a representative value (MER) If we examine specific scenarios then differences appear. For example, suppose *PRATIO* = 1.1, indicating that the price of Coke is 10% higher than the price of Pepsi, and no store displays are present. The estimated marginal effect of *PRATIO* from the probit model is −0.4519, with standard error 0.0703. Using the logit estimates, the marginal effect is −0.4898 with standard error 0.0753.

Prediction success Another basis for comparison is how well the alternative models predict choice outcomes. For the linear probability model, compute the predicted value $\widehat{COKE}$, then predict consumer choice by comparing this value to 0.5. If $\widehat{COKE}$ is greater than 0.5, we predict the consumer will choose Coke. For the probit model, we estimate the probability of choosing Coke using equation (16.10). Using the 0.5 threshold, we find that of the 510 consumers who chose *COKE*, 247 were correctly predicted. Of the 630 who chose *PEPSI*, 507 were correctly predicted. In this example, the number of correct predictions is identical for the linear probability model, probit and logit.

16.2.5 Wald Hypothesis Tests

Hypothesis tests concerning individual coefficients in probit and logit models are carried out in the usual way based on an "asymptotic-*t*" test. If the null hypothesis is $H_0: \beta_k = c$, then the test statistic using the probit model is

$$t = \frac{\tilde{\beta}_k - c}{\text{se}(\tilde{\beta}_k)} \overset{a}{\sim} N(0,1)$$

where $\tilde{\beta}_k$ is the probit parameter estimator. The test is asymptotically justified and we should use the test critical values from the standard normal distribution. For two-tail tests, these are the familiar 1.645 for 10%, 1.96 for 5%, and 2.58 for 1%. However, it is not uncommon to take a more conservative approach and, if the sample size is not very large, to use critical values from the $t_{(N-K)}$ distribution, where K is the number of parameters estimated. Your software may report "z" statistics instead of "t" and automatically compute p-values and calculate interval estimates with critical numbers from the standard normal distribution, rather than the t-distribution.

The t-test is based on the *Wald principle*, which uses the model coefficient estimates, estimated variances, covariances, and standard errors that are asymptotically valid. This testing principle is discussed in Appendix C.8.4. It is common for software packages to have "built in" Wald test statements (something like "TEST") that are convenient to use after a model is estimated. For linear hypotheses, such as $H_0: c_2\beta_2 + c_3\beta_3 = c_0$, the test statistic is of the familiar form,

$$t = \frac{\left(c_2\tilde{\beta}_2 + c_3\tilde{\beta}_3\right) - c_0}{\sqrt{c_2^2\widehat{\text{var}}\left(\tilde{\beta}_2\right) + c_3^2\widehat{\text{var}}\left(\tilde{\beta}_3\right) + 2c_2c_3\widehat{\text{cov}}\left(\tilde{\beta}_2, \tilde{\beta}_3\right)}}$$

If the null hypothesis is true, then this statistic has an asymptotic $N(0, 1)$ distribution but again $t_{(N-K)}$ might be used if the sample is not truly large. For **joint linear hypotheses**, such as

$$H_0: c_2\beta_2 + c_3\beta_3 = c_0, \quad a_4\beta_4 + a_5\beta_5 = a_0$$

a valid large sample Wald test is based on the chi-square distribution. If there are J joint hypotheses, the Wald statistic has an asymptotic $\chi^2_{(J)}$ distribution. The null hypothesis is rejected if the Wald test statistic, W, is greater than or equal to the $(1 - \alpha)$ percentile of the $\chi^2_{(J)}$ distribution, $\chi^2_{(1-\alpha,J)}$. In Section 6.1.5, we discuss large sample tests in the linear regression model. The chi-square test was labeled $\hat{V}_1$ in equation (6.14), and it was calculated as the difference between the sums of squared residuals from an unrestricted and a restricted model, divided by the estimated error variance. That is *not* the way the statistic is calculated in nonlinear models such as probit and logit, but the interpretation is the same. There is a "small-sample" conservative correction using the F-statistic, $F = W/J \overset{a}{\sim} F_{(J,N-K)}$, which is similar to using t-critical values instead of those from the $N(0, 1)$ distribution. Do not be surprised if your software reports a chi-square statistic instead of a t-statistic even when only one hypothesis is being tested.

EXAMPLE 16.7 | Coke Choice Model: Wald Hypothesis Tests

Here are some examples of various tests in the Coke choice model.

Test of significance Using the estimates in Table 16.1, we can test the significance of the coefficients in the usual way. The probit model for *COKE* is

$$p_{COKE} = \Phi\left(\beta_1 + \beta_2 PRATIO + \beta_3 DISP\_COKE + \beta_4 DISP\_PEPSI\right)$$

We might like to test the null hypothesis $H_0: \beta_3 \leq 0$ against $H_1: \beta_3 > 0$. The test statistic is $t = \tilde{\beta}_3 / \text{se}(\tilde{\beta}_3) \overset{a}{\sim} N(0,1)$ if the null hypothesis is true. Using a 5% one-tail test, the critical value is $z_{(0.95)} = 1.645$. The calculated value of the test statistic is $t = \tilde{\beta}_3 / \text{se}(\tilde{\beta}_3) = 2.2481$, and thus, we reject the null hypothesis at the 5% level and conclude that a display for Coke has a positive effect on the probability that a consumer will purchase Coke. Using a TEST statement might also produce the Wald statistic $W = 5.0540$. For a single hypothesis $W = t^2$. The Wald test statistic is designed for two-tail tests; in this case $H_0: \beta_3 = 0$ versus $H_1: \beta_3 \neq 0$, yields a two-tail p-value of $p = 0.0246$. If your software reports a t-statistic or an F-statistic, the p-value will be slightly larger, $p = 0.0248$. There is little difference here because the sample is large with $N = 1140$ observations. The Wald test critical value is $\chi^2_{(0.95,1)} = 3.841$ from Statistical Table 3.

Testing an economic hypothesis Another hypothesis of interest is $H_0: \beta_3 = -\beta_4$ versus $H_1: \beta_3 \neq -\beta_4$. This hypothesis is that the coefficients on the display variables are equal in magnitude but opposite in sign or that the effects of the Coke and Pepsi displays have an equal but opposite effect on the probability of choosing Coke. The t-test statistic is

$$t = \frac{\tilde{\beta}_3 + \tilde{\beta}_4}{\text{se}(\tilde{\beta}_3 + \tilde{\beta}_4)} \overset{a}{\sim} N(0,1)$$

Noting that it is a two-tail alternative hypothesis, we reject the null hypothesis at the $\alpha = 0.05$ level if $t \geq 1.96$ or $t \leq -1.96$. The calculated t-value is $t = -2.3247$, so we reject the null hypothesis and conclude that the effects of the Coke and Pepsi displays are not of equal magnitude with opposite sign. This test is asymptotically valid because $N - K = 1140 - 4 = 1136$ is a large sample. Automatic TEST statements usually generate the chi-square distribution version of the test, which in this case is the square of the t-statistic, $W = 5.4040$. The 5% critical value is $\chi^2_{(0.95,1)} = 3.841$ so we reject the null hypothesis. We reach the same conclusion as using the t-test. The link between the t- and chi-square test is fully explained in Appendix C.8.4.

Testing joint significance Another hypothesis of interest is

$$H_0: \beta_3 = 0,\ \beta_4 = 0 \qquad H_1: \beta_3 \neq 0 \text{ and/or } \beta_4 \neq 0$$

This joint null hypothesis is that neither the Coke nor Pepsi display affects the probability of choosing Coke. Here we are testing $J = 2$ hypotheses, so that the Wald statistic has an asymptotic $\chi^2_{(2)}$ distribution. Using Statistical Table 3, the 0.95 percentile value for this distribution is 5.991. In this case, the value of the Wald statistic is $W = 19.4594$, and thus, we reject the null hypothesis and conclude that the Coke or Pepsi display has an effect on the probability of choosing Coke. This test statistic value can be computed using the automatic TEST statement in your software.

Testing the overall model significance As in the linear regression model, we are interested in testing the overall significance of the probit model. In the Coke choice example, the null hypothesis for this test is $H_0: \beta_2 = 0,\ \beta_3 = 0,\ \beta_4 = 0$. The alternative hypothesis is that at least one of the parameters is not zero. The value of the Wald test statistic is 132.54. The test statistic has an asymptotic $\chi^2_{(3)}$ distribution if the null hypothesis is true. The 0.95 percentile value for this distribution is 7.815, so we reject the null hypothesis that none of the explanatory variables help explain the choice of Coke versus Pepsi.

16.2.6 Likelihood Ratio Hypothesis Tests

When using maximum likelihood estimators, such as probit and logit, tests based on the **likelihood ratio principle** are generally preferred. Appendix C.8.4 contains a discussion of this methodology. The idea is much like the F-test in the linear regression model. One test component is the log-likelihood function value in the unrestricted, full model (call it $\ln L_U$) evaluated at the maximum likelihood estimates. This calculation was illustrated in Example 16.3. Whenever a model is estimated by maximum likelihood, the maximized value of the log-likelihood function is automatically reported by econometric software. The second ingredient in a likelihood ratio test is the log-likelihood function value from the model that is "restricted" by imposing the condition that the null hypothesis is true (call it $\ln L_R$). Thus, the likelihood ratio test has the disadvantage of requiring two estimations of the model; once for the original model and once for the model that assumes the hypothesis is true. The likelihood ratio test statistic is $LR = 2\left(\ln L_U - \ln L_R\right)$.

The idea is that if the null hypothesis is true, then there should be little difference between the log-likelihood function with or without the hypothesis being assumed true. In that case, the *LR* statistic will be small but always greater than zero. If the null hypothesis is not true, then when we estimate the model assuming that it is true, the model should not fit as well, and the maximum value of the restricted log-likelihood function will be lower, making *LR* larger. Large values of the *LR* test statistic are evidence against the null hypothesis. If the null hypothesis is true, the statistic has an asymptotic chi-square distribution with degrees of freedom equal to the number of hypotheses, J, being tested. The null hypothesis is rejected if the value *LR* is larger than the chi-square distribution critical value, $\chi^2_{(1-\alpha,J)}$.

EXAMPLE 16.8 | Coke Choice Model: Likelihood Ratio Hypothesis Tests

We can use likelihood ratio tests for the same hypotheses considered in Example 16.7.

Test of significance The probit model for *COKE* is

$$p_{COKE} = \Phi(\beta_1 + \beta_2 PRATIO + \beta_3 DISP\_COKE + \beta_4 DISP\_PEPSI)$$

To test the null hypothesis $H_0: \beta_3 = 0$ against $H_1: \beta_3 \neq 0$ using the likelihood ratio principle, we first note that the maximized value of the log-likelihood function is $\ln L_U = -710.9486$. If the null hypothesis is true, then the restricted model is $p_{COKE} = \Phi(\beta_1 + \beta_2 PRATIO + \beta_4 DISP\_PEPSI)$. Estimating this model by maximum likelihood, we find $\ln L_R = -713.4803$, which is smaller than in the original model, as it must be. Imposing constraints on a probit model will reduce the maximized value of the log-likelihood function. Then

$$LR = 2(\ln L_U - \ln L_R) = 2[-710.9486 - (-713.4803)] = 5.0634$$

The 5% critical value is $\chi^2_{(0.95,1)} = 3.841$. We reject the null hypothesis that a display for Coke has no effect.

Test of an economic hypothesis To test $H_0: \beta_3 = -\beta_4$, we first obtain the unrestricted probit model log-likelihood value, $\ln L_U = -710.9486$. The restricted probit model is obtained by imposing the condition $\beta_3 = -\beta_4$ on the model, leading to

$$\begin{aligned} p_{COKE} &= \Phi(\beta_1 + \beta_2 PRATIO + \beta_3 DISP\_COKE + \beta_4 DISP\_PEPSI) \\ &= \Phi(\beta_1 + \beta_2 PRATIO - \beta_4 DISP\_COKE + \beta_4 DISP\_PEPSI) \\ &= \Phi(\beta_1 + \beta_2 PRATIO + \beta_4 (DISP\_PEPSI - DISP\_COKE)) \end{aligned}$$

Estimating this model by maximum likelihood probit, we obtain $\ln L_R = -713.6595$. The likelihood ratio test statistic value is then

$$LR = 2(\ln L_U - \ln L_R) = 2[-710.9486 - (-713.6595)] = 5.4218$$

This value is larger than the 0.95 percentile from the $\chi^2_{(1)}$ distribution, $\chi^2_{(0.95,1)} = 3.841$. Note that the values of the *LR* and Wald statistics (from Example 16.7) are not the same but are close in this case. The Wald test statistic value is easier to compute, since it requires only the maximum likelihood estimates for the original, unrestricted model. However, the likelihood ratio test has been found to be more reliable in a wide variety of more complex testing situations, and it is the preferred test.[5]

Test of joint significance To test the joint null hypothesis $H_0: \beta_3 = 0,\ \beta_4 = 0$, use the restricted model $E(COKE|\mathbf{x}) = \Phi(\beta_1 + \beta_2 PRATIO)$. The value of the likelihood ratio test statistic is 19.55, which is larger than the $\chi^2_{(2)}$ 0.95 percentile value 5.991. We reject the null hypothesis that neither the Coke nor Pepsi display has an effect on the choice of Coke.

Testing the overall model significance As in the linear regression model, we are interested in testing the overall significance of the probit model. In the Coke choice example, the null hypothesis for this test is $H_0: \beta_2 = 0,\ \beta_3 = 0,\ \beta_4 = 0$. The alternative hypothesis is that at least one of the parameters is not zero. If the null hypothesis is true, the restricted model is $E(COKE) = \Phi(\beta_1)$. The log-likelihood value for this restricted model is $\ln L_R = -783.8603$ and the value of the likelihood ratio test statistic is $LR = 145.8234$. The test statistic has an asymptotic $\chi^2_{(3)}$ distribution if the null hypothesis is true. The 0.95 percentile value for this distribution is 7.815, so we reject the null hypothesis that none of the explanatory variables help explain the choice of Coke versus Pepsi. In addition, like in the linear regression model, this "overall" test is reported in standard probit computer output.

[5] Griffiths, W. E., Hill, R. C., & Pope, P. (1987). Small Sample Properties of Probit Model Estimators. *Journal of the American Statistical Association*, 82, 929–937.

16.2.7 Robust Inference in Probit and Logit Models

You may be wondering if there are "robust" standard errors for use with probit and logit that correct for heteroskedasticity and/or serial correlation. Unfortunately, the answer is no. As noted in Chapter 8, equation (8.32), the 0-1 random variable y_i has conditional variance $\text{var}(y_i|\mathbf{x}_i) = p(\mathbf{x}_i)[1 - p(\mathbf{x}_i)]$. In the probit model, for example, this means that

$$\text{var}(y_i|\mathbf{x}_i) = \Phi(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})\left[1 - \Phi(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})\right]$$

There is no other possible variance if the probit model is correct. Maximum likelihood estimation of the probit model does not require any adjustment for this built in heteroskedasticity. Some software packages do have probit with a robust option, but it does not provide the type of robust results we have seen in Chapters 8 and 9. If you happen to use one of these options, and if the "robust" standard errors are much different from the usual probit standard errors, then, if anything, it is a symptom of some specification problem, such as incorrect functional form.

An exception is when there are data clusters. In Section 15.2.1, we introduced cluster-robust standard errors. There we discussed clusters in the context of panel data. However, clusters of observations, in which there are intracluster correlations, can occur in many contexts. We may observe individuals within different villages, and there may be a common unobserved heterogeneity within villages representing a "village effect." The unobserved heterogeneity causes a correlation among individuals in the same village, while there is no correlation among individuals across villages. In these situations, conventional standard errors may greatly overstate the precision of estimation. Therefore, using cluster-robust standard errors with probit and logit is recommended when the problem is suitable. In general, this means that there are many clusters with not too many observations in each.[6] Be careful when implementing cluster-robust standard errors as the computer command may be quite different from the usual "robust" standard error command.

16.2.8 Binary Choice Models with a Continuous Endogenous Variable

There are several ways that probit concepts can be combined with endogenous variables. The first is when the outcome variable is binary, as in the linear probability or probit models, and an explanatory variable is endogenous. As in our discussions of instrumental variables and two-stage least squares estimation in Chapters 10 and 11, the estimation methods here require instrumental variables.

The first, and easiest, option is to estimate a linear probability model for the binary outcome variable using IV/2SLS. To be specific, suppose that the equation of interest is

$$y_{i1} = \alpha_2 y_{i2} + \beta_1 + \beta_2 x_{i2} + e_i$$

where $y_{i1} = 1$ or 0, y_{i2} is a continuous endogenous variable, and x_{i2} is an exogenous variable, that is uncorrelated with the random error e_i. Suppose that we have an instrumental variable z_i so that the first-stage equation, or reduced form, is

$$y_{i2} = \pi_1 + \pi_2 x_{i2} + \pi_3 z_i + v_i$$

Using the IV/2SLS estimation approach, we first estimate this equation by OLS, obtain the fitted values $\hat{y}_{i2} = \hat{\pi}_1 + \hat{\pi}_2 x_{i2} + \hat{\pi}_3 z_i$. Substituting these fitted values into the equation of interest we have $y_{i1} = \alpha_2 \hat{y}_{i2} + \beta_1 + \beta_2 x_{i2} + e_i^*$. Estimating this model by OLS produces IV/2SLS estimates. However, as always, to obtain correct standard errors use IV/2SLS software, and in this case use heteroskedasticity robust standard errors.

[6] A complete but advanced resource is A. Colin Cameron and Douglas L. Miller (2015). A Practitioner's Guide to Cluster-Robust Inference, *The Journal of Human Resources*, 50(2), 317–372.

This approach is familiar and easy to implement. As always we must be concerned about the strength of the instrumental variable. The coefficient π_3 must not be zero, and when the first-stage model is estimated, it must be statistically very significant. As previously noted, using the linear probability model is not ideal when the outcome variable is binary. The procedure we have outlined ignores the binary character of the outcome variable, but it may reasonably estimate the population average marginal effect. There is another, more theoretically complicated, maximum likelihood estimator that is called *instrumental variables probit*, or simply *IV probit*.[7] This estimator is available in some software packages.

EXAMPLE 16.9 | Estimating the Effect of Education on Labor Force Participation

When studying the wages of married women, Examples 10.1–10.7 using data file *mroz*, we were very concerned with the endogeneity of education. In those examples, we only considered women who were in the labor force and had an observable market wage. Now we ask about the effect of education on the decision to join the labor force or not. Let

$$LFP = \begin{cases} 1 & \text{in labor force} \\ 0 & \text{not in labor force} \end{cases}$$

Consider the linear probability model

$$LFP = \alpha_1 EDUC + \beta_1 + \beta_2 EXPER + \beta_3 EXPER^2 + \beta_4 KIDSL6 + \beta_5 AGE + e$$

Suppose the instrumental variable for *EDUC* is *MOTHEREDUC*. The first-stage equation is

$$EDUC = \pi_1 + \pi_2 EXPER + \pi_3 EXPER^2 + \pi_4 KIDSL6 + \pi_5 AGE + \pi_6 MOTHEREDUC + v$$

In the first-stage estimation, the t-value for the coefficient of *MOTHEREDUC* is 12.85, which using conventional standards indicates that this instrument is not weak. The two-stage least squares estimates of the labor force participation equation, with robust standard errors, are

$$\widehat{LFP} = 0.0388 EDUC + 0.5919 + 0.0394 EXPER$$
$$\text{(se)} \quad (0.0165) \qquad (0.2382) \quad (0.0060)$$
$$- 0.0006 EXPER^2 - 0.2712 KIDSL6 - 0.0177 AGE$$
$$(0.0002) \qquad (0.03212) \qquad (0.0023)$$

We estimate that each additional year of education increases the probability of a married woman being in the labor force by 0.0388, holding all else constant. The regression-based Hausman test for the endogeneity of education, using robust standard errors, has a p-value of 0.646. Thus, we cannot reject the exogeneity of education in this model, using the instrument *MOTHEREDUC*.

16.2.9 Binary Choice Models with a Binary Endogenous Variable

Modify the model in Section 16.2.8 so that the endogenous variable y_{i2} is binary. The first, and easiest, option is again to estimate a linear probability model for the binary outcome variable using IV/2SLS. To be specific, suppose that the equation of interest is

$$y_{i1} = \alpha_2 y_{i2} + \beta_1 + \beta_2 x_{i2} + e_i$$

where $y_{i1} = 1$ or 0, $y_{i2} = 1$ or 0, and x_{i2} is an exogenous variable, that is uncorrelated with the random error e_i. Suppose that we have an instrumental variable z_i so that the first-stage equation, or reduced form, is

$$y_{i2} = \pi_1 + \pi_2 x_{i2} + \pi_3 z_i + v_i$$

Using the IV/2SLS estimation approach, we first estimate this equation by OLS, obtain the fitted values $\hat{y}_{i2} = \hat{\pi}_1 + \hat{\pi}_2 x_{i2} + \hat{\pi}_3 z_i$. Substituting these fitted values into the equation of interest we

[7] See William Greene (2018) *Econometric Analysis, Eighth Edition*, Prentice-Hall, page 773, or Jeffery M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data, Second Edition*, MIT Press, p. 585–594. These references are very advanced.

have $y_{i1} = \alpha_2 \hat{y}_{i2} + \beta_1 + \beta_2 x_{i2} + e_i^*$. Estimate this model by OLS to obtain IV/2SLS estimates. Of course, as always, use proper IV/2SLS software and, because the dependent variable is binary, use heteroskedasticity robust standard errors.

It is tempting but *incorrect* to think that the first-stage equation can be estimated by probit, followed by substituting $\tilde{p}_i = \tilde{P}(y_{i2} = 1) = \Phi(\tilde{\pi}_1 + \tilde{\pi}_2 x_{i2} + \tilde{\pi}_3 z_i)$ into the equation of interest, and then applying either probit or the linear probability model. The second estimation is called a forbidden regression.[8] Two-stage least squares works only when it consists of two OLS regressions, substituting OLS fitted values from a first-stage regression in for the endogenous variable in the first equation. 2SLS works because OLS has the property that the residuals are uncorrelated with the explanatory variables.

Once again the linear probability model approach "works" but does not use the fact that $y_{i1} = 1$ or 0 and $y_{i2} = 1$ or 0 are binary variables. A maximum likelihood estimation approach called bivariate probit[9] does take this into account.

EXAMPLE 16.10 | Women's Labor Force Participation and Having More Than Two Children

The Angrist and Evans (1998)[10] model of labor force participation, $LFP = 1$ or 0, includes as an explanatory variable the indicator variable $MOREKIDS = 1$ if the woman has three or more children, and $MOREKIDS = 0$ otherwise. Intuitively, we think having three or more children will have a negative effect on the probability of labor force participation. The very clever instrumental variable used is the indicator variable where the value $SAMESEX = 1$ if the woman's first two children are of the same sex, and $SAMESEX = 0$ otherwise. The idea behind this instrumental variable is that while it should have no direct effect on labor force participation it is correlated with a woman having three or more children. If a woman's first two children are both boys (girls), then she may be inclined to have another child in the hope of getting a girl (boy).

16.2.10 Binary Endogenous Explanatory Variables

Modify the model in Section 16.2.9 so that the outcome variable y_{i1} is continuous and the endogenous variable y_{i2} is binary. This model has long been studied and was first called a *dummy endogenous variable model* by Nobel prize winner James Heckman. The first, and easiest, option is to use IV/2SLS. To be specific, suppose that the equation of interest is

$$y_{i1} = \alpha_2 y_{i2} + \beta_1 + \beta_2 x_{i2} + e_i$$

where y_{i1} is continuous, the endogenous variable $y_{i2} = 1$ or 0, and x_{i2} is an exogenous variable, that is uncorrelated with the random error e_i. Suppose that we have an instrumental variable z_i so that the first-stage equation, or reduced form, is

$$y_{i2} = \pi_1 + \pi_2 x_{i2} + \pi_3 z_i + v_i$$

[8] Jeffery M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data, Second Edition*, MIT Press, p. 267–268 and 596–597.

[9] See William Greene (2018) *Econometric Analysis, Eighth Edition*, Prentice-Hall, Chapter 17.9, or Jeffery M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data*, Second Edition, MIT Press, pages. 594–599. These references are very advanced.

[10] Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size, *The American Economic Review*, Vol. 88, No. 3 (Jun., 1998), pp. 450–477. See also Jeffery M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data, Second Edition*, MIT Press, pp. 597–598.

Using the IV/2SLS estimation approach, we first estimate this linear probability model by OLS, obtain the fitted values $\hat{y}_{i2} = \hat{\pi}_1 + \hat{\pi}_2 x_{i2} + \hat{\pi}_3 z_i$. Substituting these fitted values into the equation of interest we have $y_{i1} = \alpha_2 \hat{y}_{i2} + \beta_1 + \beta_2 x_{i2} + e_i^*$. Estimating this model by OLS yields the 2SLS estimates. As always proper IV/2SLS software should be used.

The presence of an endogenous binary variable is an important feature in some **treatment effect** models.[11]

EXAMPLE 16.11 | Effect of War Veteran Status on Wages

A widely cited work by Joshua Angrist examines the effect of serving in the Vietnam war on the wages of male American workers. December 1, 1969, there was a lottery to determine eligibility for being drafted into service. Imagine 366 slips of paper each written with a birth date. The slips are placed in a jar, mixed up, and a slip drawn. The first date drawn was September 14. All men of eligible age with that birthday were given draft lottery number 1. The second date drawn was April 24 and was given lottery number 2, and so on. In the first lottery, all those with lottery numbers 195 or less were called to report for possible induction into the military. Some of those chosen did not serve for medical or other reasons, and some chose to volunteer. Thus, those who ultimately served, and became war veterans, did not correspond exactly to those with lottery numbers less than or equal to 195.

Consider a model of worker earnings, 10 years after the draft. Let *VETERAN* = 1 if a person was a veteran and = 0 otherwise. Because some chose to volunteer, the binary variable *VETERAN* is endogenous in the model

$$EARNINGS = \alpha_2 VETERAN + \beta_1 + \beta_2 OTHER\_FACTORS + e_i$$

What is a possible instrument? A person's lottery number is correlated with veteran status. More specifically, let *LOTTERY* = 1 if a person's draft lottery number was 195 or less, and *LOTTERY* = 0 otherwise. We anticipate that *LOTTERY* will be positively correlated with *VETERAN* and is a potential instrument. This type of binary IV leads to the Wald estimator, introduced in Exercises 10.5 and 10.6. The results of the IV estimation show that serving in the military has a negative and significant effect on wages.

16.2.11 Binary Choice Models and Panel Data

In Chapter 15, we used panel data to control for unobservable heterogeneity across individuals. The *fixed effects estimator* includes an indicator, or dummy, variable for each individual. Equivalently, the *within* estimator uses deviations about individual means to estimate coefficients of the regression function. We use the fixed effects estimator when the unobservable heterogeneity is correlated with the explanatory variables. The *random effects estimator* is a generalized least squares estimator that accounts for intra-individual error correlations caused by unobserved heterogeneity. It is more efficient than the fixed effects estimator but is inconsistent if the unobservable heterogeneity is correlated with any of the included explanatory variables.

If the outcome variable is binary, then using the panel data methods with the linear probability model is exactly the same as with the linear regression model. If there is unobserved heterogeneity that is correlated with one or more explanatory variables, then using the fixed effects estimator or the first difference estimator is appropriate. If the unobserved heterogeneity is not correlated with any explanatory variables, then using the random effects estimator is an option, as is the less efficient but consistent OLS estimator with robust cluster-corrected standard errors.

Using probit or logit with panel data is a different story. The probit model is a nonlinear model, that is, a nonlinear function of the parameters. If the unobserved heterogeneity is

[11] A discussion of the results and similar estimators can be found in Joshua D. Angrist and Jörn-Steffen Pischke (2009) *Mostly Harmless Econometrics: An Empiricist's Guide*, Princeton Press, pages 128–138. This reference is advanced. Other examples and estimation approaches for treatment effects are in Jeffery M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data, Second Edition*, MIT Press, Chapter 21. This reference is very advanced. For an advanced and exhaustive survey see G. W. Imbens and J. M. Wooldridge (2009) "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86.

correlated with the explanatory variables, we have a problem. The usual fixed effect approach to dealing with individual heterogeneity fails. If there are N individuals and $N \to \infty$ (gets large) while T remains fixed, then adding an indicator variable for each individual leads to a model in which the number of parameters we must estimate $N + K$ also approaches ∞. The probit estimator is no longer consistent because there are too many parameters. In statistics, this is called the *incidental parameters problem*. In the linear regression model, we avoid this problem by using the within-transformation, based on the Frisch–Waugh–Lovell theorem, so that we can estimate the regression function parameters without having to estimate all the fixed effects coefficients. This does not work in probit because of the nonlinear nature of the problem. In probit, we cannot apply the Frisch–Waugh–Lovell theorem and simply using variables in deviations about the mean form does not work. There is no fixed effects probit estimator, although researchers are considering methods for reducing the bias of the estimator so that it might be used. On the other hand, there is a type of panel logit fixed effects model called **conditional logit**, or sometimes *Chamberlain's conditional logit*,[12] recognizing the innovative econometrician Gary Chamberlain. It is not the same as introducing indicator variables for each individual into the logit model.

The probit model can however be combined with random effects to obtain a *random effects probit* model. The actual method of maximizing the likelihood function requires some tricky integrals, which can be solved using numerical approximations or simulations. As with the linear regression model, the random effects estimator is inconsistent if the random effects are correlated with the explanatory variables. It has been suggested that controls for time invariant factors, such as the time averages of the independent variables, $\bar{x}_{i\cdot}$, be introduced, similar to the Mundlak method for carrying out the Hausman test discussed in Chapter 15. The resulting model is called the *Mundlak–Chamberlain-correlated random effects probit model*.[13] The added variables $\bar{x}_{i\cdot}$ act like control variables, possibly reducing the random effects probit estimator bias.

A dynamic binary choice model, which includes the lagged value of the choice variable on the right-hand side as an explanatory variable, is an obvious way to handle habit persistence. Coke drinkers buying soda today are more likely to purchase Coke if they purchased Coke when shopping on the previous occasion. However, in such models, the lagged endogenous variable will be correlated with the random effect, as noted in Chapter 15. In this case, the previous estimators are inconsistent and new methods[14] must be considered.

16.3 Multinomial Logit

In probit and logit models, the decision-maker chooses between two alternatives. Clearly, we are often faced with choices involving more than two alternatives. These are called **multinomial choice** situations. Examples include the following:

- If you are shopping for a laundry detergent, which one do you choose? Tide, Cheer, Arm & Hammer, Wisk, and so on. The consumer is faced with a wide array of alternatives. Marketing researchers relate these choices to prices of the alternatives, advertising, and product characteristics.
- If you enroll in the business school, will you major in economics, marketing, management, finance, or accounting?
- If you are going to a mall on a shopping spree, which mall will you go to, and why?
- When you graduated from high school, you had to choose between not going to college and going to a private 4-year college, a public 4-year college, or a 2-year college. What factors led to your decision among these alternatives?

[12] See Wooldridge (2010, 620–622), Greene (2018, 787–789), or Baltagi (2013, 240–243). The material is advanced.

[13] See Greene (2018, 792–793) and Wooldridge (2010, 616–619).

[14] See Greene (2018, 794–796), Baltagi (2013, 248–253), and Wooldridge (2010, 625–630).

It would not take you long to come up with other illustrations. In each of these cases, we wish to relate the observed choice to a set of explanatory variables. More specifically, as in probit and logit models, we wish to explain and estimate the probability that an individual with a certain set of characteristics chooses one of the alternatives. The estimation and interpretation of such models is, in principle, similar to that in logit and probit models. The models themselves go under the names **multinomial logit**, **conditional logit**, and **multinomial probit**. We will discuss the most commonly used logit models.

16.3.1 Multinomial Logit Choice Probabilities

Suppose that a decision-maker must choose between several distinct alternatives. Let us focus on a problem with $J = 3$ alternatives. An example might be the choice facing a high-school graduate. Shall I attend a 2-year college, a 4-year college, or not go to college? The factors affecting this choice might include household income, the student's high-school grades, family size, race, and sex, and the parents' education. As in the logit and probit models, we will try to explain the probability that the *i*th person will choose alternative *j*,

$$p_{ij} = P[\text{individual } i \text{ chooses alternative } j]$$

In our example, there are $J = 3$ alternatives, denoted by $j = 1, 2$, or 3. These numerical values have no meaning because the alternatives in general have no particular ordering and are assigned arbitrarily. You can think of them as categories A, B, and C.

If we assume a single explanatory factor, x_i, then, in the multinomial logit specification, the probabilities of individual i choosing alternatives $j = 1, 2, 3$ are

$$p_{i1} = \frac{1}{1 + \exp(\beta_{12} + \beta_{22}x_i) + \exp(\beta_{13} + \beta_{23}x_i)}, \quad j = 1 \tag{16.19a}$$

$$p_{i2} = \frac{\exp(\beta_{12} + \beta_{22}x_i)}{1 + \exp(\beta_{12} + \beta_{22}x_i) + \exp(\beta_{13} + \beta_{23}x_i)}, \quad j = 2 \tag{16.19b}$$

$$p_{i3} = \frac{\exp(\beta_{13} + \beta_{23}x_i)}{1 + \exp(\beta_{12} + \beta_{22}x_i) + \exp(\beta_{13} + \beta_{23}x_i)}, \quad j = 3 \tag{16.19c}$$

The parameters β_{12} and β_{22} are specific to the second alternative and β_{13} and β_{23} are specific to the third alternative. The parameters specific to the first alternative are set to zero to solve an identification problem and to make the probabilities sum to one.[15] Setting $\beta_{11} = \beta_{21} = 0$ leads to the 1 in the numerator of p_{i1} and the 1 in the denominator of each part of (16.19). Specifically, the term that would be there is $\exp(\beta_{11} + \beta_{21}x_i) = \exp(0 + 0x_i) = 1$.

A distinguishing feature of the multinomial logit model in (16.19) is that there is a single explanatory variable that describes the individual, *not* the alternatives facing the individual. Such variables are called individual specific. To distinguish the alternatives, we give them different parameter values. This situation is common in the social sciences, where surveys record many characteristics of the individuals, and choices they made.

16.3.2 Maximum Likelihood Estimation

Let y_{i1}, y_{i2}, and y_{i3} be indicator variables representing the choice made by individual i. If alternative 1 is selected, then $y_{i1} = 1$, $y_{i2} = 0$, and $y_{i3} = 0$. If alternative 2 is selected, then $y_{i1} = 0$, $y_{i2} = 1$, and $y_{i3} = 0$. In this model, each individual must choose one, and only one, of the available alternatives.

[15] Some software may choose the parameters of the last (Jth) alternative to set to zero, or perhaps the most frequently chosen group. Check your software documentation.

Estimation of this model is by maximum likelihood. Suppose that we observe three individuals, who choose alternatives 1, 2, and 3, respectively. Assuming that their choices are independent, then the probability of observing this outcome is

$$\begin{aligned} P(y_{11}=1, y_{22}=1, y_{33}=1 \mid x_1, x_2, x_3) &= p_{11} \times p_{22} \times p_{33} \\ &= \frac{1}{1+\exp(\beta_{12}+\beta_{22}x_1)+\exp(\beta_{13}+\beta_{23}x_1)} \\ &\quad \times \frac{\exp(\beta_{12}+\beta_{22}x_2)}{1+\exp(\beta_{12}+\beta_{22}x_2)+\exp(\beta_{13}+\beta_{23}x_2)} \\ &\quad \times \frac{\exp(\beta_{13}+\beta_{23}x_3)}{1+\exp(\beta_{12}+\beta_{22}x_3)+\exp(\beta_{13}+\beta_{23}x_3)} \\ &= L(\beta_{12}, \beta_{22}, \beta_{13}, \beta_{23}) \end{aligned}$$

In the last line, we recognize that this joint probability depends on the unknown parameters and is in fact the likelihood function. Maximum likelihood estimation seeks those values of the parameters that maximize the likelihood or, more specifically, the **log-likelihood function**, which is easier to work with mathematically. In a real application, the number of individuals will be greater than three, and computer software will be used to maximize the log-likelihood function numerically. While the task might look daunting, finding the maximum likelihood estimates in this type of model is fairly simple.

16.3.3 Multinomial Logit Postestimation Analysis

Given that we can obtain maximum likelihood estimates of the parameters, which we denote as $\tilde{\beta}_{12}$, $\tilde{\beta}_{22}$, $\tilde{\beta}_{13}$, and $\tilde{\beta}_{23}$, what can we do then? The first thing we might do is estimate the probability that an individual will choose alternative 1, 2, or 3. For the value of the explanatory variable x_0, we can calculate the predicted probabilities of each outcome being selected using (16.19). For example, the probability that such an individual will choose alternative 1 is

$$\tilde{p}_{01} = \frac{1}{1+\exp(\tilde{\beta}_{12}+\tilde{\beta}_{22}x_0)+\exp(\tilde{\beta}_{13}+\tilde{\beta}_{23}x_0)}$$

The estimated probabilities for alternatives 2 and 3, $\tilde{p}_{02}$ and $\tilde{p}_{03}$, can similarly be obtained. If we wanted to predict which alternative would be chosen, we might choose to predict that alternative j will be chosen if $\tilde{p}_{0j}$ is the maximum of the estimated probabilities.

Because the model is such a complicated nonlinear function of the parameters, it will not surprise you to learn that the βs are not "slopes." In these models, the **marginal effect** is the effect of a change in x, everything else held constant, on the probability that an individual chooses alternative $m = 1, 2$, or 3. It can be shown[16] that

$$\left.\frac{\Delta p_{im}}{\Delta x_i}\right|_{\text{all else constant}} = \frac{\partial p_{im}}{\partial x_i} = p_{im}\left[\beta_{2m} - \sum_{j=1}^{3}\beta_{2j}p_{ij}\right] \tag{16.20}$$

Recall that the model we are discussing has a single explanatory variable, x_i, and that $\beta_{21} = 0$.

[16] One can quickly become overwhelmed by the mathematics when seeking references on this topic. Two relatively friendly sources with good examples are *Regression Models for Categorical and Limited Dependent Variables* by J. Scott Long (Thousand Oaks, CA: Sage Publications, 1997) [see Chapter 5] and *Quantitative Models in Marketing Research* by Philip Hans Franses and Richard Paap (Cambridge University Press, 2001) [see Chapter 5]. At a much more advanced level, see *Econometric Analysis, Eighth edition* by William Greene (Upper Saddle River, NJ: Pearson Prentice Hall, 2018) [see Section 18.2.3].

Alternatively, and somewhat more simply, the difference in probabilities can be calculated for two specific values of x_i. If x_a and x_b are two values of x_i, then the estimated change in probability of choosing alternative 1 [$m = 1$] when changing from x_a to x_b is

$$\begin{aligned}\widetilde{\Delta p_1} &= \tilde{p}_{b1} - \tilde{p}_{a1} \\ &= \frac{1}{1 + \exp\left(\tilde{\beta}_{12} + \tilde{\beta}_{22} x_b\right) + \exp\left(\tilde{\beta}_{13} + \tilde{\beta}_{23} x_b\right)} \\ &\quad - \frac{1}{1 + \exp\left(\tilde{\beta}_{12} + \tilde{\beta}_{22} x_a\right) + \exp\left(\tilde{\beta}_{13} + \tilde{\beta}_{23} x_a\right)}\end{aligned}$$

This approach is good if there are certain scenarios that you as a researcher have in mind as typical or important cases or if x is an indicator variable with only two values, $x_a = 0$ and $x_b = 1$.

Another useful interpretive device is the **probability ratio**. It shows how many times more likely category j is to be chosen relative to the first category and is given by

$$\frac{P(y_i = j)}{P(y_i = 1)} = \frac{p_{ij}}{p_{i1}} = \exp\left(\beta_{1j} + \beta_{2j} x_i\right), \quad j = 2, 3 \tag{16.21}$$

The effect on the probability ratio of changing the value of x_i is given by the derivative

$$\frac{\partial\left(p_{ij}/p_{i1}\right)}{\partial x_i} = \beta_{2j} \exp\left(\beta_{1j} + \beta_{2j} x_i\right), \quad j = 2, 3 \tag{16.22}$$

The value of the exponential function $\exp\left(\beta_{1j} + \beta_{2j} x_i\right)$ is always positive. Thus, the sign of β_{2j} tells us whether a change in x_i will make the jth category more or less likely relative to the first category.

An interesting feature of the probability ratio (16.21) is that it does not depend on how many alternatives there are in total. There is the implicit assumption in logit models that the probability ratio between any pair of alternatives is **independent of irrelevant alternatives (IIA)**. This is a strong assumption, and if it is violated, multinomial logit may not be a good modeling choice. It is especially likely to fail if several alternatives are similar. Tests for the IIA assumption work by dropping one or more of the available options from the choice set and then reestimating the multinomial model. If the IIA assumption holds, then the estimates should not change very much. A statistical comparison of the two sets of estimates, one set from the model with a full set of alternatives, and the other from the model using a reduced set of alternatives, is carried out using a Hausman contrast test proposed by Hausman and McFadden (1984).[17]

EXAMPLE 16.12 | Postsecondary Education Multinomial Choice

The National Education Longitudinal Study of 1988 (NELS:88) was the first nationally representative longitudinal study of eighth-grade students in public and private schools in the United States. It was sponsored by the National Center for Education Statistics. In 1988, some 25,000 eighth-graders and their parents, teachers, and principals were surveyed. In 1990, these same students (who were then mostly 10th graders, and some dropouts) and their teachers and principals were surveyed again. In 1992, the second follow-up survey was conducted of students, mostly in the 12th grade, but dropouts, parents, teachers, school administrators, and high school transcripts were also surveyed. The third follow-up was in 1994, after most students had graduated.[18]

[17] "Specification Tests for the Multinomial Logit Model," *Econometrica*, 49, pp. 1219–1240. A brief explanation of the test may be found in Greene (2018, Chapter 18.2.4), op. cit., p. 767.

[18] The study and data are summarized in *National Education Longitudinal Study: 1988–1994, Descriptive Summary Report with an Essay on Access and Choice in Post-Secondary Education*, by Allen Sanderson, Bernard Dugoni, Kenneth Rasinski, and John Taylor, C. Dennis Carroll project officer, NCES 96-175, National Center for Education Statistics, March 1996.

We have taken a subset of the total data, namely those who stayed in the panel of data through the third follow-up. On this group, we have complete data on the individuals and their households, high-school grades, and test scores, as well as their postsecondary education choices. In the data file *nels_small*, we have 1000 observations on students who chose, upon graduating from high school, either no college (*PSECHOICE* = 1), a 2-year college (*PSECHOICE* = 2), or a 4-year college (*PSECHOICE* = 3). For illustration purposes, we focus on the explanatory variable *GRADES*, which is an index ranging from 1.0 (highest level, A+ grade) to 13.0 (lowest level, F grade) and represents combined performance in English, maths, and social studies.

Of the 1000 students, 22.2% selected not to attend a college upon graduation, 25.1% selected to attend a 2-year college, and 52.7% attended a 4-year college. The average value of *GRADES* is 6.53, with highest grade 1.74 and lowest grade 12.33. The estimated values of the parameters and their standard errors are given in Table 16.2. We selected the group who did not attend a college to be our base group, so that the parameters $\beta_{11} = \beta_{21} = 0$.

Based on these estimates, what can we say? Recall that a larger numerical value of *GRADES* represents a poorer academic performance. The parameter estimates for the coefficients of *GRADES* are negative and statistically significant. Using expression (16.22) on the effect of a change in an explanatory variable on the probability ratio, this means that if the value of *GRADES* increases, the probability that high-school graduates will choose a 2-year or a 4-year college goes down, relative to the probability of not attending college. This is the anticipated effect, as we expect that a poorer academic performance will increase the odds of not attending college.

We can also compute the estimated probability of each type of college choice using (16.19) for given values of *GRADES*. In our sample, the median value of *GRADES* is 6.64, and the top 5th percentile value is 2.635.[19] What are the choice probabilities of students with these grades? In Table 16.3, we show that the probability of choosing no college is 0.1810 for the student with median grades, but this probability is reduced to 0.0178 for students with top grades. Similarly, the probability of choosing a 2-year school is 0.2856 for the average student but is 0.0966 for the better student. Finally, the average student has a 0.5334 chance of selecting a 4-year college, but the better student has a 0.8857 chance of selecting a 4-year college.

The marginal effect of a change in *GRADES* on the choice probabilities can be calculated using (16.20). The marginal effect again depends on particular values for *GRADES*, and we report these in Table 16.3 for the median and 5th percentile students. An increase in *GRADES* of one point (worse performance) increases the probabilities of choosing either no college or a 2-year college and reduces the probability of attending a 4-year college. The probability of attending a 4-year college declines more for the average student than for the top student, given the one-point increase in *GRADES*. Note that for each value of *GRADES* the sum of the predicted probabilities is one, and the sum of the marginal effects is zero, except for rounding error. This is a feature of the multinomial logit specification.

TABLE 16.2 **Maximum Likelihood Estimates of PSE Choice**

Parameters	Estimates	Standard Errors	*t*-Statistics
β_{12}	2.5064	0.4183	5.99
β_{22}	−0.3088	0.0523	−5.91
β_{13}	5.7699	0.4043	14.27
β_{23}	−0.7062	0.0529	−13.34

TABLE 16.3 **Effects of Grades on Probability of PSE Choice**

PSE Choice	*GRADES*	$\hat{p}$	$\text{se}(\hat{p})$	Marginal Effect	se(ME)
No college	6.64	0.1810	0.0149	0.0841	0.0063
	2.635	0.0178	0.0047	0.0116	0.0022
Two-year college	6.64	0.2856	0.0161	0.0446	0.0076
	2.635	0.0966	0.0160	0.0335	0.0024
Four-year college	6.64	0.5334	0.0182	−0.1287	0.0095
	2.635	0.8857	0.0174	−0.0451	0.0030

[19] The 5th percentile value of *GRADES* is given as 2.635 which is halfway between observations 50 and 51 in this 1,000 observation data set. While this is a common way to calculate the 5th percentile, it is not the only way. Since $0.05 \times 1000 = 50$, some software will report the 50th value, after sorting according to increasing value, 2.63. Others may take a weighted average of the 50th and 51st values, such as $0.95 \times 2.63 + 0.05 \times 2.64 = 2.6305$. Thanks to Tom Doan (Estima) for noting this. Standard errors in Table 16.3 are computed via "the delta method," in a fashion similar to that described in Appendix 16A.

16.4 Conditional Logit

Suppose that a decision-maker must choose between several distinct alternatives, just as in the multinomial logit model. In a marketing context, suppose that our decision is between three types ($J = 3$) of soft drinks, say Pepsi, 7-Up, and Coke Classic, in 2-liter bottles. Shoppers will visit their supermarkets and make a choice, based on prices of the products and other factors. With the advent of supermarket scanners at checkout, data on purchases (what brand, how many units, and the price paid) are recorded. Of course, we also know the prices of the products that the consumer did not buy on a particular shopping occasion. The key point is that if we collect data on soda purchases from a variety of supermarkets, over a period of time, we observe consumer choices from the set of alternatives and we know the prices facing the shopper on each trip to the supermarket.

Let y_{i1}, y_{i2}, and y_{i3} be indicator variables that indicate the choice made by individual i. If alternative one (Pepsi) is selected, then $y_{i1} = 1$, $y_{i2} = 0$, and $y_{i3} = 0$. If alternative two (7-Up) is selected, then $y_{i1} = 0$, $y_{i2} = 1$, and $y_{i3} = 0$. If alternative 3 (Coke) is selected, then $y_{i1} = 0$, $y_{i2} = 0$, and $y_{i3} = 1$. The price facing individual i for brand j is $PRICE_{ij}$. That is, the price of Pepsi, 7-Up, and Coke is potentially different for each customer who purchases soda. Remember, different customers can shop at different supermarkets and at different times. Variables like *PRICE* are *individual- and alternative-specific* because they vary from individual to individual and are different for each choice the consumer might make. This type of information is very different from what we assumed was available in the multinomial logit model, where the explanatory variable x_i was *individual-specific*; it did not change across alternatives.

16.4.1 Conditional Logit Choice Probabilities

Our objective is to understand the factors that lead a consumer to choose one alternative over another. We construct a model for the probability that individual i chooses alternative j

$$p_{ij} = P[\text{individual } i \text{ chooses alternative } j]$$

The conditional logit model specifies these probabilities as

$$p_{ij} = \frac{\exp(\beta_{1j} + \beta_2 PRICE_{ij})}{\exp(\beta_{11} + \beta_2 PRICE_{i1}) + \exp(\beta_{12} + \beta_2 PRICE_{i2}) + \exp(\beta_{13} + \beta_2 PRICE_{i3})} \tag{16.23}$$

Note that unlike the probabilities for the multinomial logit model in (16.19), there is only one parameter β_2 relating the effect of each price to the choice probability p_{ij}. We have also included alternative specific constants (intercept terms). These cannot all be estimated, and one must be set to zero. We will set $\beta_{13} = 0$.

Estimation of the unknown parameters is by maximum likelihood. Suppose that we observe three individuals, who choose alternatives one, two, and three, respectively. Assuming that their choices are independent, then the probability of observing this outcome is

$$\begin{aligned} P(y_{11} = 1, y_{22} = 1, y_{33} = 1) &= p_{11} \times p_{22} \times p_{33} \\ &= \frac{\exp(\beta_{11} + \beta_2 PRICE_{11})}{\exp(\beta_{11} + \beta_2 PRICE_{11}) + \exp(\beta_{12} + \beta_2 PRICE_{12}) + \exp(\beta_2 PRICE_{13})} \\ &\quad \times \frac{\exp(\beta_{12} + \beta_2 PRICE_{22})}{\exp(\beta_{11} + \beta_2 PRICE_{21}) + \exp(\beta_{12} + \beta_2 PRICE_{22}) + \exp(\beta_2 PRICE_{23})} \\ &\quad \times \frac{\exp(\beta_2 PRICE_{33})}{\exp(\beta_{11} + \beta_2 PRICE_{31}) + \exp(\beta_{12} + \beta_2 PRICE_{32}) + \exp(\beta_2 PRICE_{33})} \\ &= L(\beta_{11}, \beta_{12}, \beta_2) \end{aligned}$$

16.4.2 Conditional Logit Postestimation Analysis

How a change in price affects the choice probability is different for "own price" changes and "cross-price" changes. Specifically, it can be shown that the own price effect is

$$\frac{\partial p_{ij}}{\partial PRICE_{ij}} = p_{ij}(1 - p_{ij})\beta_2 \tag{16.24}$$

The sign of β_2 indicates the direction of the own price effect.

The change in probability of alternative j being selected if the price of alternative k changes ($k \neq j$) is

$$\frac{\partial p_{ij}}{\partial PRICE_{ik}} = -p_{ij}p_{ik}\beta_2 \tag{16.25}$$

The cross-price effect is in the opposite direction of the own price effect.

An important feature of the conditional logit model is that the probability ratio between alternatives j and k is

$$\frac{p_{ij}}{p_{ik}} = \frac{\exp(\beta_{1j} + \beta_2 PRICE_{ij})}{\exp(\beta_{1k} + \beta_2 PRICE_{ik})} = \exp\left[(\beta_{1j} - \beta_{1k}) + \beta_2(PRICE_{ij} - PRICE_{ik})\right]$$

The probability ratio depends on the difference in prices but not on the prices themselves. As in the multinomial logit model, this ratio does not depend on the total number of alternatives, and there is the implicit assumption of the **independence of irrelevant alternatives (IIA)**. See the discussion at the end of Section 16.3.3. Models that do not require the IIA assumption have been developed, but they are difficult. These include the *multinomial probit* model, which is based on the normal distribution, and the *nested logit* and *mixed logit* models.[20]

EXAMPLE 16.13 | Conditional Logit Soft Drink Choice

We observe 1822 purchases, covering 104 weeks and 5 stores, in which a consumer purchased 2-liter bottles of either Pepsi (34.6%), 7-Up (37.4%), or Coke Classic (28%). These data are in the file *cola*. In the sample, the average price of Pepsi was \$1.23, 7-Up \$1.12, and Coke \$1.21. We estimate the conditional logit model shown in (16.22), and the estimates are shown in Table 16.4a.

TABLE 16.4a **Conditional Logit Parameter Estimates**

Variable	Estimate	Standard Error	*t*-Statistic	*p*-Value
$PRICE(\beta_2)$	−2.2964	0.1377	−16.68	0.000
$PEPSI(\beta_{11})$	0.2832	0.0624	4.54	0.000
$7\text{-}UP(\beta_{12})$	0.1038	0.0625	1.66	0.096

We see that all the parameter estimates are significantly different from zero at a 10% level of significance, and the sign of the coefficient of *PRICE* is negative. This means that a rise in the price of an individual brand will reduce the probability of its purchase, and the rise in the price of a competitive brand will increase the probability of its purchase. Table 16.4b contains the marginal effects of price changes on the probability of choosing Pepsi. The marginal effects are calculated using (16.24) and (16.25) with prices of Pepsi, 7-Up, and Coke set to \$1.00, \$1.25, and \$1.10, respectively. The standard errors are calculated using the delta method. Note two things about these estimates. First, they have the signs we anticipate. An increase in the price of Pepsi is estimated to have a negative effect on the probability of Pepsi purchase, while an increase in the price of either Coke or 7-Up increases the probability that Pepsi will be selected. Second, these values are very large for changes in probabilities because a "one-unit change" is \$1, which then represents almost a 100% change in price. For a 10-cent increase in

[20] For a brief description of these models at an advanced level, see William Greene, Econometric Analysis, Eighth Edition by (Upper Saddle River, NJ: Pearson Prentice Hall, 2018), Chapter 18.2.5. Mixed and nested logit models are important in applied research. David A. Hensher, John M. Rose, William H. Greene (2015) *Applied Choice Analysis, 2nd Edition*, Cambridge University Press, provide a comprehensive overview and integration of choice models, along with software instructions using the *NLOGIT* software package. Survey methodology is also discussed.

the prices the marginal effects, standard errors and interval estimate bounds should be multiplied by 0.10.

TABLE 16.4b **Marginal Effect of Price on Probability of Pepsi Choice**

PRICE	Marginal Effect	Standard Error	95% Interval Estimate
COKE	0.3211	0.0254	[0.2712, 0.3709]
PEPSI	−0.5734	0.0350	[−0.6421, −0.5048]
7-UP	0.2524	0.0142	[0.2246, 0.2802]

As an alternative to computing marginal effects, we can compute specific probabilities at given values of the explanatory variables. For example, at the prices used for Table 16.4b, the estimated probability of selecting Pepsi is then

$$\hat{p}_{i1} = \frac{\exp(\tilde{\beta}_{11} + \tilde{\beta}_2 \times 1.00)}{\left[\exp(\tilde{\beta}_{11} + \tilde{\beta}_2 \times 1.00) + \exp(\tilde{\beta}_{12} + \tilde{\beta}_2 \times 1.25) + \exp(\tilde{\beta}_2 \times 1.10)\right]} = 0.4832$$

The standard error for this predicted probability is 0.0154, which is computed via the delta method. If we raise the price of Pepsi to \$1.10, we estimate that the probability of its purchase falls to 0.4263 (se = 0.0135). If the price of Pepsi stays at \$1.00 but we increase the price of Coke by 15 cents, then we estimate that the probability of a consumer selecting Pepsi rises by 0.0445 (se = 0.0033). These numbers indicate to us the responsiveness of brand choice to changes in prices, much like elasticities.

16.5 Ordered Choice Models

The choice options in multinomial and conditional logit models have no natural ordering or arrangement. However, in some cases, choices are ordered in a specific way. Examples include the following:

1. Results of opinion surveys in which responses can be strongly in disagreement, in disagreement, neutral, in agreement, or strongly in agreement.
2. Assignment of grades or work performance ratings. Students receive grades A, B, C, D, and F, which are ordered on the basis of a teacher's evaluation of their performance. Employees are often given evaluations on scales such as outstanding, very good, good, fair, and poor, which are similar in spirit.
3. Standard and Poor's rates bonds as AAA, AA, A, BBB, and so on, as a judgment about the credit worthiness of the company or country issuing a bond, and how risky the investment might be.
4. Levels of employment as unemployed, part time, or full time.

When modeling these types of outcomes, numerical values are assigned to the outcomes, but the numerical values are **ordinal** and reflect only the ranking of the outcomes. In the first example, we might assign a dependent variable y the values

$$y = \begin{cases} 1 & \text{strongly disagree} \\ 2 & \text{disagree} \\ 3 & \text{neutral} \\ 4 & \text{agree} \\ 5 & \text{strongly agree} \end{cases}$$

In Section 16.3, we considered the problem of choosing what type of college to attend after graduating from high school as an illustration of a choice among unordered alternatives.

However, in this particular case, there may in fact be natural ordering. We might rank the possibilities as

$$y = \begin{cases} 3 & \text{four-year college (the full college experience)} \\ 2 & \text{two-year college (a partial college experience)} \\ 1 & \text{no college} \end{cases} \tag{16.26}$$

The usual linear regression model is not appropriate for such data, because in regression we would treat the y-values as having some numerical meaning when they do not. In the following section, we discuss how probabilities of each choice might be modeled.

16.5.1 Ordinal Probit Choice Probabilities

When faced with a ranking problem, we develop a "sentiment" about how we feel concerning the alternative choices, and the higher the sentiment, the more likely a higher ranked alternative will be chosen. This sentiment is, of course, unobservable to the econometrician. Unobservable variables that enter decisions are called **latent variables**, and we will denote our sentiment toward the ranked alternatives by y_i^*, with the "star" reminding us that this variable is unobserved. See Appendix 16B for more on latent variables.

Microeconomics is well described as the "science of choice." Economic theory will suggest that certain factors (observable variables) may affect how we feel about the alternatives facing us. As a concrete example, let us think about what factors might lead a high-school graduate to choose among the alternatives "no college," "2-year college," and "4-year college" as described by the ordered choices in (16.26). Some factors that affect this choice are household income, the student's high-school grades, how close a 2- or 4-year college is to the home, whether parents had attended a 4-year college, and so on. For simplicity, let us focus on the single explanatory variable *GRADES*. The model is then

$$y_i^* = \beta GRADES_i + e_i$$

This model is not a regression model because the dependent variable is unobservable. Consequently, it is sometimes called an **index model**. The error term is present for the usual reasons. The choices we observe are based on a comparison of "sentiment" toward higher education y_i^* relative to certain thresholds, as shown in Figure 16.2.

Because there are $M = 3$ alternatives, there are $M - 1 = 2$ thresholds μ_1 and μ_2, with $\mu_1 < \mu_2$. The index model does not contain an intercept because it would be exactly collinear with the threshold variables. If sentiment toward higher education is in the lowest category, then $y_i^* \le \mu_1$ and the alternative "no college" is chosen, if $\mu_1 < y_i^* \le \mu_2$ then the alternative "2-year college"

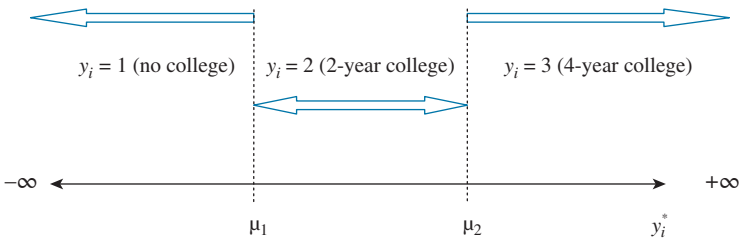

FIGURE 16.2 **Ordinal choices relative to thresholds.**

is chosen, and if sentiment toward higher education is in the highest category, then $y_i^* > \mu_2$ and "4-year college" is chosen. That is,

$$y_i = \begin{cases} 3 \text{ (four-year college)} & \text{if } y_i^* > \mu_2 \\ 2 \text{ (two-year college)} & \text{if } \mu_1 < y_i^* \le \mu_2 \\ 1 \text{ (no college)} & \text{if } y_i^* \le \mu_1 \end{cases}$$

We are able to represent the probabilities of these outcomes if we assume a particular probability distribution for y_i^*, or equivalently for the random error e_i. If we assume that the errors have the standard normal distribution, $N(0, 1)$, an assumption that defines the **ordered probit model**, then we can calculate the following:

$$\begin{aligned} P(y_i = 1) &= P(y_i^* \le \mu_1) = P(\beta GRADES_i + e_i \le \mu_1) \\ &= P(e_i \le \mu_1 - \beta GRADES_i) \\ &= \Phi(\mu_1 - \beta GRADES_i) \end{aligned}$$

$$\begin{aligned} P(y_i = 2) &= P(\mu_1 < y_i^* \le \mu_2) = P(\mu_1 < \beta GRADES_i + e_i \le \mu_2) \\ &= P(\mu_1 - \beta GRADES_i < e_i \le \mu_2 - \beta GRADES_i) \\ &= \Phi(\mu_2 - \beta GRADES_i) - \Phi(\mu_1 - \beta GRADES_i) \end{aligned}$$

and the probability that $y = 3$ is

$$\begin{aligned} P(y_i = 3) &= P(y_i^* > \mu_2) = P(\beta GRADES_i + e_i > \mu_2) \\ &= P(e_i > \mu_2 - \beta GRADES_i) \\ &= 1 - \Phi(\mu_2 - \beta GRADES_i) \end{aligned}$$

16.5.2 Ordered Probit Estimation and Interpretation

Estimation, as with previous choice models, is by maximum likelihood. If we observe a random sample of $N = 3$ individuals, with the first not going to college ($y_1 = 1$), the second attending a 2-year college ($y_2 = 2$), and the third attending a 4-year college ($y_3 = 3$), then the likelihood function is

$$L(\beta, \mu_1, \mu_2) = P(y_1 = 1) \times P(y_2 = 2) \times P(y_3 = 3)$$

Note that the probabilities depend on the unknown parameters μ_1 and μ_2 as well as the index function parameter β. These parameters are obtained by maximizing the log-likelihood function using numerical methods. Econometric software includes options for both **ordered probit**, which depends on the errors being standard normal, and **ordered logit**, which depends on the assumption that the random errors follow a logistic distribution. Most economists will use the normality assumption, but many other social scientists use the logistic. In the end, there is little difference between the results.

The types of questions we can answer with this model are the following:

1. What is the probability that a high-school graduate with $GRADES = 2.5$ (on a 13-point scale, with one being the highest) will attend a 2-year college? The answer is obtained by plugging in the specific value of *GRADES* into the estimated probability using the maximum likelihood estimates of the parameters,

$$\hat{P}(y = 2 | GRADES = 2.5) = \Phi(\tilde{\mu}_2 - \tilde{\beta} \times 2.5) - \Phi(\tilde{\mu}_1 - \tilde{\beta} \times 2.5)$$

2. What is the difference in probability of attending a 4-year college for two students, one with $GRADES = 2.5$ and another with $GRADES = 4.5$? The difference in the probabilities is calculated directly as

$$\hat{P}(y = 3|GRADES = 4.5) - \hat{P}(y = 3|GRADES = 2.5)$$

3. If we treat *GRADES* as a continuous variable, what is the marginal effect on the probability of each outcome, given a one-unit change in *GRADES*? These derivatives are

$$\frac{\partial P(y = 1)}{\partial GRADES} = -\phi(\mu_1 - \beta GRADES) \times \beta$$

$$\frac{\partial P(y = 2)}{\partial GRADES} = \left[\phi(\mu_1 - \beta GRADES) - \phi(\mu_2 - \beta GRADES)\right] \times \beta$$

$$\frac{\partial P(y = 3)}{\partial GRADES} = \phi(\mu_2 - \beta GRADES) \times \beta$$

In these expressions, "$\phi(\cdot)$" denotes the *pdf* of a standard normal distribution, and its values are always positive. Consequently, the sign of the parameter β is opposite the direction of the marginal effect for the lowest category, but it indicates the direction of the marginal effect for the highest category. The direction of the marginal effect for the middle category goes one way or the other, depending on the sign of the difference in brackets.

There are a variety of other devices that can be used to analyze the outcomes, including some useful graphics. For more on these, see (from a social science perspective) *Regression Models for Categorical and Limited Dependent Variables* by J. Scott Long (Sage Publications, 1997, Chapter 5) or (from a marketing perspective) *Quantitative Models in Marketing Research* by Philip Hans Franses and Richard Paap (Cambridge University Press, 2001, Chapter 6). A comprehensive reference is by William H. Greene and David A. Hensher (2010) *Modeling Ordered Choices: A Primer*, Cambridge University Press.

EXAMPLE 16.14 | Postsecondary Education Ordered Choice Model

To illustrate, we use the college choice data introduced in Section 16.3 and contained in the data file *nels_small*. We treat *PSECHOICE* as an ordered variable with 1 representing the least favored alternative (no college) and 3 denoting the most favored alternative (4-year college). The estimation results are in Table 16.5.

TABLE 16.5 **Ordered Probit Parameter Estimates for PSE Choice**

Parameters	Estimates	Standard Errors
β	−0.3066	0.0191
μ_1	−2.9456	0.1468
μ_2	−2.0900	0.1358

The estimated coefficient of *GRADES* is negative, indicating that the probability of attending a 4-year college goes down when *GRADES* increase (indicating a worse performance), and the probability of the lowest ranked choice, attending no college, increases. Let us examine the marginal effects of an increase in *GRADES* on attending a 4-year college. For a student with median grades (6.64), the marginal effect is −0.1221, and for a student in the 5th percentile (2.635), the marginal effect is −0.0538. These are similar in magnitude to the marginal effects shown in Table 16.3.

16.6 Models for Count Data

When the dependent variable in a regression model is a count of the number of occurrences of an event, the outcome variable is $y = 0, 1, 2, 3, \ldots$ These numbers are actual counts and thus different from the ordinal numbers of the previous section. Examples include the following:

- The number of trips to a physician a person makes during a year.
- The number of fishing trips taken by a person during the previous year.
- The number of children in a household.
- The number of automobile accidents at a particular intersection during a month.
- The number of televisions in a household.
- The number of alcoholic drinks a college student takes in a week.

While we are again interested in explaining and estimating probabilities, such as the probability that an individual will take two or more trips to the doctor during a year, the probability distribution we use as a foundation is the Poisson, not the normal or the logistic. If Y is a **Poisson random variable**, then its probability function is

$$f(y) = P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y = 0, 1, 2, \ldots \tag{16.27}$$

The factorial (!) term $y! = y \times (y-1) \times (y-2) \times \cdots \times 1$. This probability function has one parameter, λ, which is the mean (and variance) of Y. That is, $E(Y) = \text{var}(Y) = \lambda$. In a regression model, we try to explain the behavior of $E(Y)$ as a function of some explanatory variables. We do the same here, keeping the value of $E(Y) \geq 0$ by defining

$$E(Y|x) = \lambda = \exp(\beta_1 + \beta_2 x) \tag{16.28}$$

This choice defines the **Poisson regression model** for count data.

16.6.1 Maximum Likelihood Estimation of the Poisson Regression Model

The parameters β_1 and β_2 in (16.28) can be estimated by maximum likelihood. Suppose that we randomly select $N = 3$ individuals from a population and observe that their counts are $y_1 = 0$, $y_2 = 2$, and $y_3 = 2$, indicating 0, 2, and 2 occurrences of the event for these three individuals. Recall that the likelihood function is the joint probability function of the observed data, interpreted as a function of the unknown parameters. That is,

$$L(\beta_1, \beta_2) = P(Y = 0) \times P(Y = 2) \times P(Y = 2)$$

This product of functions like (16.27) will be very complicated and difficult to maximize. However, in practice, maximum likelihood estimation is carried out by maximizing the logarithm of the likelihood function, or

$$\ln L(\beta_1, \beta_2) = \ln P(Y = 0) + \ln P(Y = 2) + \ln P(Y = 2)$$

Using (16.28) for λ, the log of the probability function is

$$\begin{aligned} \ln\left[P(Y = y|x)\right] &= \ln\left[\frac{e^{-\lambda}\lambda^{y}}{y!}\right] = -\lambda + y\ln(\lambda) - \ln(y!) \\ &= -\exp(\beta_1 + \beta_2 x) + \left[y \times (\beta_1 + \beta_2 x)\right] - \ln(y!) \end{aligned}$$

Then the log-likelihood function, given a sample of N observations, becomes

$$\ln L(\beta_1, \beta_2) = \sum_{i=1}^{N} \left\{ -\exp(\beta_1 + \beta_2 x_i) + y_i \times (\beta_1 + \beta_2 x_i) - \ln(y_i!) \right\}$$

This log-likelihood function is a function of only β_1 and β_2 once we substitute in the data values (y_i, x_i). The log-likelihood function itself is still a nonlinear function of the unknown parameters, and the maximum likelihood estimates must be obtained by numerical methods. Econometric software has options that allow for the maximum likelihood estimation of count models with the click of a button.

16.6.2 Interpreting the Poisson Regression Model

As in other modeling situations, we would like to use the estimated model to predict outcomes, determine the marginal effect of a change in an explanatory variable on the mean of the dependent variable, and test the significance of coefficients.

Estimation of the conditional mean of y is straightforward. Given the maximum likelihood estimates $\tilde{\beta}_1$ and $\tilde{\beta}_2$, and given a value of the explanatory variable x_0,

$$\widehat{E(y_0)} = \tilde{\lambda}_0 = \exp(\tilde{\beta}_1 + \tilde{\beta}_2 x_0)$$

This value is an estimate of the expected number of occurrences observed if x takes the value x_0. The probability of a particular number of occurrences can be estimated by inserting the estimated conditional mean into the probability function, as

$$\widehat{P(Y = y)} = \frac{\exp(-\tilde{\lambda}_0)\,\tilde{\lambda}_0^{y}}{y!}, \quad y = 0, 1, 2, \ldots$$

The marginal effect of a change in a continuous variable x in the Poisson regression model is not simply given by the parameter because the conditional mean model is a nonlinear function of the parameters. Using our specification that the conditional mean is given by $E(y_i|x_i) = \lambda_i = \exp(\beta_1 + \beta_2 x_i)$, and using rules for derivatives of exponential functions, we obtain the marginal effect

$$\frac{\partial E(y_i|x_i)}{\partial x_i} = \lambda_i \beta_2 \tag{16.29}$$

To estimate this marginal effect, replace the parameters by their maximum likelihood estimates and select a value for x. The marginal effect is different depending on the value of x chosen. A useful fact about the Poisson model is that the conditional mean $E(y_i|x_i) = \lambda_i = \exp(\beta_1 + \beta_2 x_i)$ is always positive because the exponential function is always positive. Thus, the direction of the marginal effect can be determined from the sign of the coefficient β_2.

Equation (16.29) can be expressed as a percentage, which can be useful:

$$\frac{\%\Delta E(y_i|\mathbf{x})}{\Delta x_i} = 100 \frac{\partial E(y_i|\mathbf{x})/E(y_i|\mathbf{x})}{\partial x_i} = 100\beta_2\%$$

If x is not transformed, then a one-unit change in x leads to $100\beta_2\%$ change in the conditional mean.

Suppose that the conditional mean function contains a indicator variable, how do we calculate its effect? If $E(y_i|\mathbf{x}) = \lambda_i = \exp(\beta_1 + \beta_2 x_i + \delta D_i)$, we can examine the conditional expectation when $D = 0$ and when $D = 1$.

$$E(y_i|x_i, D_i = 0) = \exp(\beta_1 + \beta_2 x_i)$$

$$E(y_i|x_i, D_i = 1) = \exp(\beta_1 + \beta_2 x_i + \delta)$$

Then, the percentage change in the conditional mean is

$$100\left[\frac{\exp(\beta_1+\beta_2 x_i+\delta)-\exp(\beta_1+\beta_2 x_i)}{\exp(\beta_1+\beta_2 x_i)}\right]\% = 100\left[e^{\delta}-1\right]\%$$

This is identical to the expression we obtained for the effect of an indicator variable in a log-linear model. See Section 7.3.

Finally, hypothesis testing can be carried out using standard methods. The maximum likelihood estimators are asymptotically normal with a variance of a known form. The actual expression for the variance is complicated and involves matrix expressions, so we will not report the formula here.[21] Econometric software has the variance expressions encoded, and along with parameter estimates, it will provide standard errors, t-statistics, and p-values, which are used as always.

EXAMPLE 16.15 | A Count Model for the Number of Doctor Visits

The economic analysis of the health care system is a vital area of research and public interest. In this example, we consider data used by Riphahn, Wambach, and Million (2003).[22] The data file *rwm88_small* contains data on 1,200 individuals' number of doctor visits in the past three months (*DOCVIS*), their age in years (*AGE*), their sex (*FEMALE*), and whether or not they had public insurance (*PUBLIC*). The frequencies of doctor visits are illustrated in Table 16.6, with 90.5% of the sample having eight or fewer visits.

TABLE 16.6 **Number of Doctor Visits (*DOCVIS*)**

DOCVIS	Number
0	443
1	200
2	163
3	111
4	51
5	49
6	37
7	7
8	25

These are numerical **count data** (number of times an event occurs) so that the Poisson model is a feasible choice. Applying maximum likelihood estimation, we obtain the fitted model

$$\underset{(\text{se})}{\widehat{E(DOCVIS)}} = \exp\underset{(0.0918)}{(-0.0030}+\underset{(0.0015)}{0.0116}AGE+\underset{(0.0335)}{0.1283}FEMALE+\underset{(0.0680)}{0.5726}PUBLIC)$$

What can we say about these results? First, the coefficient estimates are all positive, implying that older individuals, females and those with public health insurance will have more doctor visits. Second, the coefficients of *AGE*, *FEMALE* and *PUBLIC* are significantly different from zero, with p-values less than 0.01. Using the fitted model, we can estimate the expected number of doctor visits. For example, the first person in the sample is a 29-year-old female who has public insurance. Substituting these values we estimate her expected number of doctor visits to be 2.816, or 3.0 rounded to the nearest integer. Her actual number of doctor visits was zero.

Using the notion of generalized-R^2, we can get a notion of how well the model fits the data by computing the squared correlation between *DOCVIS* and the predicted number of visits. If we use the rounded values, for example, 3.0 instead of 3.33, the correlation is 0.1179 giving $R_g^2 = (0.1179)^2 = 0.0139$. The fit for this simple model is not very good as we might well expect. This model does not account for so many important factors, such as income, general health status, and so on. Different software packages report many different values, sometimes called *pseudo*-R^2,

[21] See J. Scott Long, *Regression Models for Categorical and Limited Dependent Variables* (Thousand Oaks, CA: Sage Publications, 1997), Chapter 8. A much more advanced and specialized reference is *Regression Analysis of Count Data* Second Edition by A. Colin Cameron and Pravin K. Trivedi (Cambridge, UK: Cambridge University Press, 2013).

[22] Regina T. Riphahn, Achim Wambach, and Andreas Million, "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation", *Journal of Applied Econometrics*, Vol. 18, No. 4, 2003, pp. 387–405.

with different meanings as well. We urge you to ignore all these values, including R_g^2.

Instead of an R^2-like number, it is a good idea to report a test of overall model significance, analogous to the overall F-test for the regression model. The null hypothesis is that all the model coefficients, except the intercept, are equal to zero. We recommend the likelihood ratio statistic. See Section 16.2.7 for a discussion of this test in the context of the probit model. The test statistic is $LR = 2(\ln L_U - \ln L_R)$ where $\ln L_U$ is the value of the log-likelihood function for the full and unrestricted model and $\ln L_R$ is the value of the log-likelihood function for the restricted model that assumes that the hypothesis is true. The restricted model in this case is $E(DOCVIS) = \exp(\gamma_1)$. If the null hypothesis is true, the LR test statistic has a $\chi^2_{(3)}$-distribution in large samples. In our example, $LR = 174.93$ and the 0.95 percentile of the $\chi^2_{(3)}$-distribution is 7.815. We reject the null hypothesis at the 5% level of significance, and we conclude that at least one variable makes a significant impact on the number of doctor visits.

What about the magnitudes of the effects of these variables on the number of doctor visits? Treating *AGE* as continuous we can use (16.29) to compute a marginal effect,

$$\widehat{\frac{\partial E(DOCVIS)}{\partial AGE}} = \exp(-0.0030 + 0.0116AGE + 0.1283FEMALE + 0.5726PUBLIC) \times 0.0116$$

To evaluate this effect, we must insert values for *AGE*, *FEMALE,* and *PUBLIC*. Let *FEMALE* = 1 and *PUBLIC* = 1. If *AGE* = 30, the estimate is 0.0332, with the 95% interval estimate being [0.0261, 0.0402]. That is, we estimate for a 30-year-old female with public insurance an additional year of age will increase her expected number of doctor visits in a 3-month period by 0.0332. Because the marginal effect is a nonlinear function of the estimated parameters, the interval estimate uses a standard error calculated using the delta method. For *AGE* = 70, it is 0.0528 [0.0355, 0.0702]. The effect of another year of age is greater for older individuals, as you would expect.

Both *FEMALE* and *PUBLIC* are indicator variables, taking values zero and one. For these variables, we cannot evaluate the "marginal effect" using a derivative. Instead, we estimate the difference between the expected number of doctor visits for the two cases. For example,

$$\Delta E(DOCVIS) = E(DOCVIS|PUBLIC = 1) - E(DOCVIS|PUBLIC = 0)$$

The calculated value of the difference is

$$\widehat{\Delta E(DOCVIS)} = \exp(-0.0030 + 0.0116AGE + 0.1283FEMALE + 0.5726) - \exp(-0.0030 + 0.0116AGE + 0.1283FEMALE)$$

We estimate the difference for a 30-year-old female to be 1.24 [1.00, 1.48], and for a 70-year-old female, it is 1.98 [1.59, 2.36]. Women with public insurance visit the doctor significantly more than women of the same age who do not have public insurance.

There are many generalizations of the Poisson model that are used in applied work. One generalization is called the *negative binomial model*. It can be used when an assumption implicit in the Poisson model is violated, namely that the variance of Poisson variable is equal to its expected value, that is $\text{var}(Y) = E(Y)$ for Poisson random variables. There are tests for whether this assumption holds. These are sometimes called *tests for overdispersion*. A second type of possible misspecification is illustrated by the following question: How many extramarital affairs did you have in the last year?[23] The first thing to note is that the question is relevant only for married individuals. The possible answers are zero, one, two, three, etc. However, here a "zero" might mean two different things. It might mean, "I would *never* cheat on my spouse!!" or it might mean, "Well, I have cheated in the past, but not in the last year." Statistically, in this situation, there will be "too many zeros" for the standard Poisson distribution. As a result, there are some *zero-inflated* versions of the Poisson model (*ZIP*) that may be a better choice. These extensions of the Poisson model are quite fascinating and useful but beyond the scope of this book.[24]

[23] Ray Fair (1978) "The Theory of Extramarital Affairs," *Journal of Political Economy*, 86(1), 45–61.

[24] Two excellent but advanced references are: A. Colin Cameron and Pravin K. Trivedi (2013) *Regression Analysis of Count Data*, Cambridge University Press; and Rainer Winkelman (2008) *Econometric Analysis of Count Data*, Springer.

16.7 Limited Dependent Variables

In the previous sections of this chapter, we reviewed choice behavior models that have dependent variables that are discrete variables. When a model has a discrete dependent variable, the usual regression methods we have studied must be modified. In this section, we present another case in which standard least squares estimation of a regression model fails.

16.7.1 Maximum Likelihood Estimation of the Simple Linear Regression Model

We have stressed the least squares and method of moments estimators when estimating the simple linear regression model. Another option is maximum likelihood estimation (MLE). Our discussion of the method will be in the context of the simple linear model with one explanatory variable, but the method extends to the case of multiple regression with more explanatory variables. We discuss this now because several strategies for estimating **limited dependent variable** models are tied to MLE. In this case, we make assumptions SR1–SR6, which include the assumption about the conditional normality of the random errors. When the assumption of conditionally normal errors is made, we write $e_i|x_i \sim N(0, \sigma^2)$, and also then $y_i|x_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2)$. It is a very strong assumption when it is made, and it is not necessary for least squares estimation, so we have called it an *optional* assumption. For maximum likelihood estimation, it is *not* optional. It is necessary to assume a distribution for the data so that we can form the likelihood function.

If the data (y_i, x_i) pairs are drawn independently, then the conditional joint *pdf* of the data is

$$f(y_1, y_2, \ldots, y_N|\mathbf{x}, \beta_1, \beta_2, \sigma^2) = f(y_1|x_1, \beta_1, \beta_2, \sigma^2) \times \cdots \times f(y_N|x_N, \beta_1, \beta_2, \sigma^2) \quad (16.30a)$$

where

$$f(y_i|x_i, \beta_1, \beta_2, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i - \beta_1 - \beta_2 x_i)^2}{\sigma^2}\right) \quad (16.30b)$$

Writing out the product we have

$$f(y_1, \ldots, y_N|\mathbf{x}, \beta_1, \beta_2, \sigma^2) = (2\pi)^{-N/2}(\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i)^2\right]$$

$$= L(\beta_1, \beta_2, \sigma^2|\mathbf{y}, \mathbf{x}) \quad (16.30c)$$

The likelihood function $L(\beta_1, \beta_2, \sigma^2|\mathbf{y}, \mathbf{x})$ is the joint *pdf* interpreted as function of the unknown parameters, conditional on the data. In practice, we maximize the log-likelihood,

$$\ln L(\beta_1, \beta_2, \sigma^2|\mathbf{y}, \mathbf{x}) = -(N/2)\ln(2\pi) - (N/2)\ln(\sigma^2)$$

$$- \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i)^2 \quad (16.30d)$$

This looks quite intimidating to maximize, but this is one of the times we can actually maximize the log-likelihood using calculus. See Exercise 16.1 for hints. The maximum likelihood estimators $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are the OLS estimators, which have all their usual properties including a conditionally normal distribution. The MLE of the error variance is $\tilde{\sigma}^2 = \sum \hat{e}_i^2/N$, which is the sum of the squared least squares residuals divided by the sample size, with no degrees of freedom correction. This estimator is consistent but biased.

16.7.2 Truncated Regression

The first limited dependent variable model we consider is **truncated regression**. In this model, we only observe the data (y_i, x_i) when $y_i > 0$. How can this happen? Imagine collecting survey data by waiting at the checkout station in a supermarket. As customers exit you ask "How much did you spend today?" The answer will be some positive number given that they have just paid for their purchases. If the random error is conditionally normal, then the *pdf* of $(y_i|y_i > 0, x_i)$ is *truncated normal*. The properties of the truncated normal distribution are discussed in Appendix B.3.5. In this case, the truncated normal density function is

$$f(y_i|y_i > 0, x_i, \beta_1, \beta_2, \sigma^2) = \frac{f(y_i|x_i, \beta_1, \beta_2, \sigma^2)}{P(y_i > 0|x_i, \beta_1, \beta_2, \sigma^2)} = \frac{f(y_i|x_i, \beta_1, \beta_2, \sigma^2)}{\Phi\left(\frac{\beta_1 + \beta_2 x_i}{\sigma}\right)} = \frac{f(y_i|x_i, \beta_1, \beta_2, \sigma^2)}{\Phi_i} \tag{16.31}$$

Here we use $\Phi_i = \Phi[(\beta_1 + \beta_2 x_i)/\sigma]$ as a simplifying notation. See Exercise 16.2 for hints on obtaining (16.31). The log-likelihood function is

$$\ln L(\beta_1, \beta_2, \sigma^2|\mathbf{y}, \mathbf{x}) = -\sum_{i=1}^{N} \ln \Phi_i - (N/2)\ln(2\pi) - (N/2)\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i)^2 \tag{16.32}$$

Maximization of this log-likelihood has to be done using numerical methods, but econometric software has simple commands to obtain the estimates of β_1, β_2, and σ^2. The question then becomes, what can we do with these estimates? The answer is, all the usual things. For the calculation of marginal effects, use the conditional mean function

$$E(y_i|y_i > 0, x_i) = \beta_1 + \beta_2 x_i + \sigma\frac{\phi[(\beta_1 + \beta_2 x_i)/\sigma]}{\Phi[(\beta_1 + \beta_2 x_i)/\sigma]} = \beta_1 + \beta_2 x_i + \sigma\lambda(\alpha_i) \tag{16.33}$$

where $\lambda(\alpha_i)$ is the inverse Mill's ratio (IMR) mentioned in Appendix B.3.5 and $\alpha_i = (\beta_1 + \beta_2 x_i)/\sigma$. This is a bit of a mess isn't it? If x_i is continuous, the marginal effect is the derivative of this expression, $dE(y_i|y_i > 0, x_i)/dx_i = \beta_2(1 - \delta_i)$, where $\delta_i = \lambda(\alpha_i)[\lambda(\alpha_i) - \alpha_i]$, which is even more messy.[25] Because $0 < \delta_i < 1$, the marginal effect is only a fraction of the parameter value. Once again econometricians in conjunction with computer programmers have made our lives much easier than would otherwise be true and these quantities can be calculated.

16.7.3 Censored Samples and Regression

Censored samples are similar to truncated samples but have more information. In a *truncated sample*, we observe (y_i, x_i) when $y_i > 0$. For censored samples, we observe x_i for all individuals, but the outcome values are of two different types. In a survey of households, suppose we ask "How much did you spend on major appliances, such as refrigerators or washing machines, last month?" For many households, the answer will be \$0, as they made no such purchases in the previous month. For others, the answer will be a positive value, if such a purchase was made. This is the outcome variable, y_i. On the other hand, the survey will include income and other characteristics of the household, which are explanatory variables, x_i. This is called a *censored sample*, with a substantial fraction of the observations taking a *limit value*, in this case \$0. We are

[25] See Greene (2018), page 932–933.

interested in estimating the relationship between expenditures and an x_i. What should we do? There are a number of strategies. We will mention four, two that work and two that do not work.

Strategy 1 **Delete the limit observations and apply OLS**

A simple strategy is to drop from the sample the observations with $y_i = 0$ and go ahead. This strategy does not work. The usual OLS model, for $y_i > 0$, is $y_i = \beta_1 + \beta_2 x_i + u_i$, where u_i is an error term. We usually think of this model as resulting from the sum of a systematic part, the regression function, and a random error. That is, $y_i = E(y_i|y_i > 0, x_i) + e_i$. The conditional mean function is given by (16.33), so that

$$\begin{aligned} y_i = E(y_i|y_i > 0, x_i) + e_i &= \beta_1 + \beta_2 x_i + \sigma \frac{\phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right]}{\Phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right]} + e_i \\ &= \beta_1 + \beta_2 x_i + \left\{ \sigma \frac{\phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right]}{\Phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right]} + e_i \right\} \\ &= \beta_1 + \beta_2 x_i + u_i \end{aligned} \tag{16.34}$$

The error term u_i is not simple. It consists of the random component e_i and a complicated function of x_i. The error term u_i will be correlated with x_i, making OLS biased and inconsistent, which is not the result we want.

Strategy 2 **Retain all observations and apply OLS**

This strategy does not work. Using the definition for conditional expectation,

$$\begin{aligned} E(y_i|x_i) &= E(y_i|y_i > 0, x_i) \times P(y_i > 0) + E(y_i|y_i = 0, x_i) \times P(y_i = 0) \\ &= E(y_i|y_i > 0, x_i) \times \left\{ 1 - \Phi\left[-(\beta_1 + \beta_2 x_i)/\sigma \right] \right\} \\ &= E(y_i|y_i > 0, x_i) \times \Phi\left[(\beta_1 + \beta_2 x_i)/\sigma \right] \\ &= \Phi\left[(\beta_1 + \beta_2 x_i)/\sigma \right] \beta_1 + \Phi\left[(\beta_1 + \beta_2 x_i)/\sigma \right] \beta_2 x_i + \sigma\phi\left[(\beta_1 + \beta_2 x_i)/\sigma \right] \end{aligned}$$

Simply estimating $y_i = \beta_1 + \beta_2 x_i + u_i$ by OLS clearly is inappropriate.

Strategy 3 **Heckman's two-step estimator**

The problem with Strategy 1 is that the error term u_i includes two components and one of them is correlated with the variable x_i. This is analogous to an omitted variables problem, the solution of which is to not omit the variable, but include it in the regression. That is, we would like to estimate the model

$$y_i = \beta_1 + \beta_2 x_i + \sigma\lambda_i + e_i$$

where

$$\lambda_i = \frac{\phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right]}{\Phi\left[(\beta_1 + \beta_2 x_i)/\sigma\right]} = \frac{\phi(\beta_1^* + \beta_2^* x_i)}{\Phi(\beta_1^* + \beta_2^* x_i)}$$

where $\beta_1^* = \beta_1/\sigma$ and $\beta_2^* = \beta_2/\sigma$. Nobel Prize winner James Heckman realized that while λ_i is unknown it can be consistently estimated as $\tilde{\lambda}_i = \phi(\tilde{\beta}_1^* + \tilde{\beta}_2^* x_i)/\Phi(\tilde{\beta}_1^* + \tilde{\beta}_2^* x_i)$ where $\tilde{\beta}_1^*$ and $\tilde{\beta}_2^*$ come from a probit model with dependent variable $d_i = 1$ if $y_i > 0$, and $d_i = 0$ if $y_i = 0$, and with explanatory variable x_i. Then the model we estimate by OLS is

$$y_i = \beta_1 + \beta_2 x_i + \sigma\tilde{\lambda}_i + e_i^*$$

It is called a two-step estimator because we use estimates from a first step, probit, and then a second step, OLS. The estimator is consistent and while correct standard errors are complicated, they are known and can be obtained.

Strategy 4 Maximum likelihood estimation: Tobit

Heckman's two-step estimator is consistent but not efficient. There is a maximum likelihood estimation procedure that is preferable. It is called **Tobit** in honor of James Tobin, winner of the 1981 Nobel Prize in Economics, who first studied the model.

Tobit is a maximum likelihood procedure that recognizes that we have data of two sorts, the limit observations ($y = 0$) and the nonlimit observations ($y > 0$). The two types of observations that we observe, the limit observations and those that are positive, are generated by a latent variable y^* crossing the zero threshold or not crossing that threshold. The (**probit**) probability that $y = 0$ is

$$P(y = 0|\mathbf{x}) = P(y^* \le 0|\mathbf{x}) = 1 - \Phi\left[(\beta_1 + \beta_2 x)/\sigma\right]$$

If we observe a positive value of y_i, then the term that enters the likelihood function is the normal *pdf* with mean $\beta_1 + \beta_2 x_i$ and variance σ^2. The full likelihood function is the product of the probabilities that the limit observations occur times the *pdfs* for all the positive, nonlimit, observations. Using "large pi" notation to denote multiplication, the likelihood function is

$$L(\beta_1, \beta_2, \sigma|\mathbf{x}, \mathbf{y}) = \prod_{y_i=0}\left\{1 - \Phi\left(\frac{\beta_1 + \beta_2 x_i}{\sigma}\right)\right\} \times \prod_{y_i>0}\left\{(2\pi\sigma^2)^{-0.5}\exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_1 - \beta_2 x_i)^2\right)\right\}$$

This complicated-looking likelihood function is maximized numerically using econometric software.[26] The maximum likelihood estimator is consistent and asymptotically normal, with a known covariance matrix.[27]

16.7.4 Tobit Model Interpretation

In the **Tobit model**, the parameters β_1 and β_2 are the intercept and slope of the latent variable model (16.31). In practice, we are interested in the marginal effect of a change in x on either the regression function of the observed data $E(y|x)$ or the regression function conditional on $y > 0$, $E(y|x, y > 0)$. As we indicated earlier, these functions are not straight lines. Their graphs are shown in Figure 16.3. The slope of each changes at each value of x. The slope of $E(y|x)$ has a relatively simple form, being a scale factor times the parameter value; it is

$$\frac{\partial E(y|x)}{\partial x} = \beta_2 \Phi\left(\frac{\beta_1 + \beta_2 x}{\sigma}\right) \tag{16.35}$$

where Φ is the cumulative distribution function (*cdf*) of the standard normal random variable that is evaluated at the estimates and a particular x-value. Because the *cdf* values are positive, the sign of the coefficient tells the direction of the marginal effect, but the magnitude of the marginal effect depends on both the coefficient and the *cdf*. If $\beta_2 > 0$, as x increases, the *cdf* function approaches one, and the slope of the regression function approaches that of the latent variable

[26] Tobit requires data on both the limit values of $y = 0$ and also the nonlimit values for which $y > 0$. Sometimes, it is possible that we do not observe the limit values; in such a case, the sample is said to be truncated. In this case, Tobit does not apply; however, there is a similar maximum likelihood procedure, called **truncated regression**, for such a case. An advanced reference is William Greene (2018) *Econometric Analysis, Eighth edition*, Pearson Prentice Hall, Section 19.2.3.

[27] The asymptotic covariance matrix can be found in *Introduction to the Theory and Practice of Econometrics, 2nd edition*, by George G. Judge, R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chao Lee (John Wiley and Sons, 1988), Section 19.3.2.

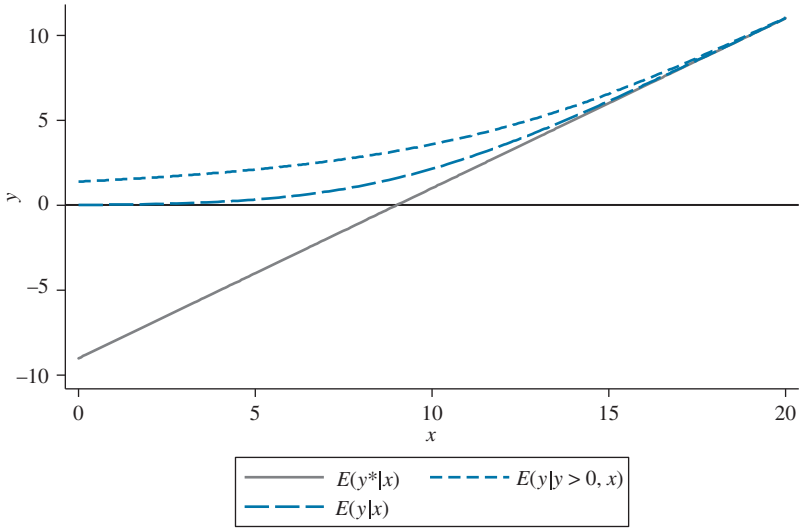

FIGURE 16.3 **Three regression functions.**

model, as is shown in Figure 16.3. The marginal effect can be decomposed into two factors called the "McDonald–Moffitt" decomposition:

$$\frac{\partial E(y|x)}{\partial x} = \text{Prob}(y > 0)\frac{\partial E(y|x, y > 0)}{\partial x} + E(y|x, y > 0)\frac{\partial \text{Prob}(y > 0)}{\partial x}$$

The first factor accounts for the marginal effect of a change in x for the portion of the population whose y-data is observed already. The second factor accounts for changes in the proportion of the population who switch from the y-unobserved category to the y-observed category when x changes.[28]

EXAMPLE 16.16 | Tobit Model of Hours Worked

An example that illustrates the situation is based on Thomas Mroz's (1987) study of married women's labor force participation and wages. The data are in the file *mroz* and consist of 753 observations on married women. Of these, 325 did not work outside the home and thus had no hours worked and no reported wages. The histogram of hours worked is shown in Figure 16.4. The histogram shows the large fraction of women with zero hours of work.

If we wish to estimate a model explaining the market hours worked by a married woman, what explanatory variables would we include? Factors that would tend to pull a woman into the labor force are her education and her prior labor market experience. Factors that may reduce her incentive to work are her age and the presence of young children in the home.[29]

Thus, we might propose the regression model

$$HOURS = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 AGE + \beta_4 KIDSL6 + e \quad (16.36)$$

[28] J. F. McDonald and R. A. Moffitt (1980) "The Uses of Tobit Analysis," *Review of Economics and Statistics*, 62, 318–321. Jeffrey M. Wooldridge (2009) *Introductory Econometrics: A Modern Approach, 5th edition*, South-Western Cengage Learning, Section 17.2 has a relatively friendly presentation.

[29] This equation does not include wages, which is jointly determined with hours. The model in (16.36) may be considered a reduced-form equation. See Section 11.2.

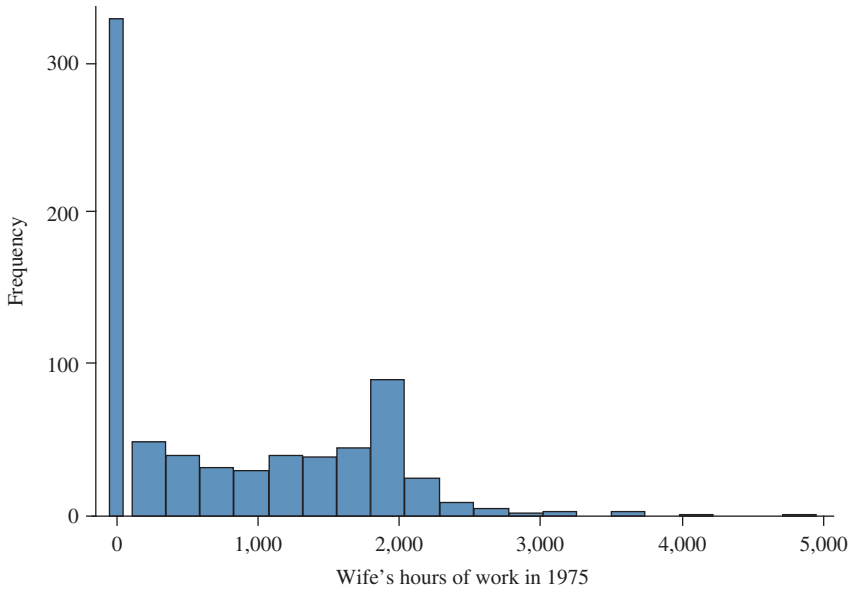

FIGURE 16.4 **Wife's hours of work in 1975.**

where *KIDSL6* is the number of children less than 6 years old in the household. Using Mroz's data, we obtain the estimates shown in Table 16.7. As previously argued, the least squares estimates are unreliable because the least squares estimator is both biased and inconsistent. The Tobit estimates have the anticipated signs and are all statistically significant at the 0.01 level. To compute the scale factor required for calculation of the marginal effects, we must choose values of the explanatory variables. We choose the sample means for *EDUC* (12.29), *EXPER* (10.63), and *AGE* (42.54) and assume one small child at home (rather than the mean value of 0.24). The calculated scale factor is $\tilde{\Phi} = 0.3630$. Thus, the marginal effect on observed hours of work of another year of education is

$$\frac{\partial E(HOURS)}{\partial EDUC} = \tilde{\beta}_2 \tilde{\Phi} = 73.29 \times 0.3630 = 26.61$$

That is, we estimate that another year of education will increase a wife's hours of work by about 27 hours, conditional upon the assumed values of the explanatory variables.

TABLE 16.7 **Estimates of Labor Supply Function**

Estimator	Variable	Estimate	Standard Error
Least squares	*INTERCEPT*	1335.31	235.65
	EDUC	27.09	12.24
	EXPER	48.04	3.64
	AGE	−31.31	3.96
	KIDSL6	−447.85	58.41
Least squares $y > 0$	*INTERCEPT*	1829.75	292.54
	EDUC	−16.46	15.58
	EXPER	33.94	5.01
	AGE	−17.11	5.46
	KIDSL6	−305.31	96.45
Tobit	*INTERCEPT*	1349.88	386.30
	EDUC	73.29	20.47
	EXPER	80.54	6.29
	AGE	−60.77	6.89
	KIDSL6	−918.92	111.66
	SIGMA	1133.70	42.06

16.7.5 Sample Selection

If you consult an econometrician concerning an estimation problem, the first question you will usually hear is, "How were the data obtained?" If the data are obtained by random sampling, then classic regression methods, such as least squares, work well. However, if the data are obtained by a sampling procedure that is not random, then standard procedures do not work well. Economists regularly face such data problems. A famous illustration comes from labor economics. If we wish to study the determinants of the wages of married women, we face a *sample selection* problem. If we collect data on married women and ask them what wage rate they earn, many will respond that the question is not relevant since they are homemakers. We only observe data on market wages when the woman chooses to enter the workforce. One strategy is to ignore the women who are not in the labor force, omit them from the sample, then use least squares to estimate a wage equation for those who work. This strategy fails, the reason for the failure being that our sample is not a random sample. The data we observe are "selected" by a systematic process for which we do not account.

A solution to this problem is a technique called **Heckit**, named after its developer, Nobel Prize winning econometrician James Heckman. This simple procedure uses two estimation steps. In the context of the problem of estimating the wage equation for married women, a probit model is first estimated explaining why a woman is in the labor force or not. In the second stage, a least squares regression is estimated relating the wage of a working woman to education, experience, and so on, and a variable called the "inverse Mills ratio," or IMR. The IMR is created from the first step probit estimation and accounts for the fact that the observed sample of working women is not random.

The econometric model describing the situation is composed of two equations. The first is the *selection equation* that determines whether the variable of interest is observed. The sample consists of N observations; however, the variable of interest is observed only for $n < N$ of these. The selection equation is expressed in terms of a latent variable z_i^* that depends on one or more explanatory variables w_i and is given by

$$z_i^* = \gamma_1 + \gamma_2 w_i + u_i, \quad i = 1, \ldots, N \tag{16.37}$$

For simplicity, we will include only one explanatory variable in the selection equation. The latent variable is not observed, but we do observe the indicator variable

$$z_i = \begin{cases} 1 & z_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{16.38}$$

The second equation is the linear model of interest. It is

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, n, \quad N > n \tag{16.39}$$

A **selectivity problem** arises when y_i is observed only when $z_i = 1$ and if the errors of the two equations are correlated. In such a situation, the usual least squares estimators of β_1 and β_2 are biased and inconsistent.

Consistent estimators are based on the conditional regression function[30]

$$E\left(y_i | z_i^* > 0\right) = \beta_1 + \beta_2 x_i + \beta_\lambda \lambda_i, \quad i = 1, \ldots, n \tag{16.40}$$

where the additional variable λ_i is the inverse Mills ratio. It is equal to

$$\lambda_i = \frac{\phi\left(\gamma_1 + \gamma_2 w_i\right)}{\Phi\left(\gamma_1 + \gamma_2 w_i\right)} \tag{16.41}$$

[30] Our Appendix B.2.6 provides a brief introduction to this important concept. See William Greene (2018) *Econometric Analysis, Eighth edition*, Pearson Prentice Hall, Chapter 19.2 for much more about the truncated normal.

While the value of λ_i is not known, the parameters γ_1 and γ_2 can be estimated using a probit model, based on the observed binary outcome z_i in (16.38). Then the estimated IMR

$$\tilde{\lambda}_i = \frac{\phi(\tilde{\gamma}_1 + \tilde{\gamma}_2 w_i)}{\Phi(\tilde{\gamma}_1 + \tilde{\gamma}_2 w_i)}$$

is inserted into the regression equation as an extra explanatory variable, yielding the estimating equation

$$y_i = \beta_1 + \beta_2 x_i + \beta_\lambda \tilde{\lambda}_i + v_i, \quad i = 1, \ldots, n \tag{16.42}$$

Least squares estimation of this equation yields consistent estimators of β_1 and β_2. A word of caution, however, as the least squares estimator is inefficient relative to the maximum likelihood estimator, and the usual standard errors and t-statistics produced after estimation of (16.42) are incorrect. Proper estimation of standard errors requires the use of specialized software for the "Heckit" model.

EXAMPLE 16.17 | Heckit Model of Wages

As an example, we will reconsider the analysis of wages earned by married women using the Mroz (1987) data in the data file *mroz*. In the sample of 753 married women, 428 have market employment and nonzero earnings. First, let us estimate a simple wage equation, explaining ln(*WAGE*) as a function of the woman's education, *EDUC*, and years of market work experience (*EXPER*), using the 428 women who have positive wages. The result is

$$\begin{aligned} \ln(WAGE) &= -0.4002 + 0.1095 EDUC \\ (t) \quad & (-2.10) \quad (7.73) \\ &+ 0.0157 EXPER \qquad R^2 = 0.1484 \qquad (16.43) \\ & (3.90) \end{aligned}$$

The estimated return to education is about 11%, and the estimated coefficients of both education and experience are statistically significant.

The Heckit procedure starts by estimating a probit model of labor force participation. As explanatory variables we use the woman's age, her years of education, an indicator variable for whether she has children, and the marginal tax rate that she would pay upon earnings if employed. The estimated probit model is

$$\begin{aligned} \widehat{P(LFP=1)} &= \Phi\big(1.1923 - 0.0206 AGE + 0.0838 EDUC \\ (t) \qquad & \qquad\quad (-2.93) \qquad\quad (3.61) \\ &\quad - 0.3139 KIDS - 1.3939 MTR\big) \\ & \quad\ (-2.54) \qquad\quad (-2.26) \end{aligned}$$

As expected, the effects of age, the presence of children, and the prospects of higher taxes significantly reduce the probability that a woman will join the labor force, while education increases it. Using the estimated coefficients, we compute the inverse Mills ratio for the 428 women with market wages

$$\tilde{\lambda} = IMR = \frac{\phi\big(1.1923 - 0.0206 AGE + 0.0838 EDUC - 0.3139 KIDS - 1.3939 MTR\big)}{\Phi\big(1.1923 - 0.0206 AGE + 0.0838 EDUC - 0.3139 KIDS - 1.3939 MTR\big)}$$

This is then included in the wage equation, and least squares estimation applied to obtain

$$\begin{aligned} \ln(WAGE) &= 0.8105 + 0.0585 EDUC + 0.0163 EXPER \\ (t) \quad & (1.64) \quad (2.45) \qquad (4.08) \\ (t-adj) \quad & (1.33) \quad (1.97) \qquad (3.88) \\ & - 0.8664 IMR \\ & (-2.65) \\ & (-2.17) \qquad\qquad (16.44) \end{aligned}$$

Two results are of note. First, the estimated coefficient of the inverse Mills ratio is statistically significant, implying that there is a **selection bias** present in the least squares results (16.43). Second, the estimated return to education has fallen from approximately 11% to approximately 6%. The upper row of t-statistics is based on standard errors as usually computed when using least squares regression. The usual standard errors do not account for the fact that the inverse Mills ratio is itself an estimated value. The correct standard errors,[31] which do account for the first stage probit

[31] The formulas are very complicated. See William Greene (2018) *Econometric Analysis, Eighth edition*, Pearson Prentice Hall, p. 954. There are several software packages, such as Stata and LIMDEP, that report correct standard errors.

estimation, are used to construct the "adjusted t-statistics" reported in (16.44). As you can see the adjusted t-statistics are slightly smaller, indicating that the adjusted standard errors are somewhat larger than the usual ones.

In most instances, it is preferable to estimate the full model, both the selection equation and the equation of interest, jointly by maximum likelihood. While the nature of this procedure is beyond the scope of this book, it is available in some software packages. The maximum likelihood estimated wage equation is

$$\underset{(t)}{\ln(WAGE)} = \underset{(2.84)}{0.6686} + \underset{(3.96)}{0.0658}EDUC + \underset{(2.87)}{0.0118}EXPER$$

The standard errors based on the full information maximum likelihood procedure are smaller than those yielded by the two-step estimation method.

16.8 Exercises

16.8.1 Problems

16.1 In Examples 16.2 and 16.4, we presented the linear probability and probit model estimates using an example of transportation choice. The logit model for the same example is $P(AUTO = 1) = \Lambda(\gamma_1 + \gamma_2 DTIME)$, where $\Lambda(\bullet)$ is the logistic *cdf* in equation (16.7). The logit model parameter estimates and their standard errors are

$$\underset{(\text{se})}{\tilde{\gamma}_1 + \tilde{\gamma}_2 DTIME} = \underset{(0.7505)}{-0.2376} + \underset{(0.2064)}{0.5311}DTIME$$

a. Calculate the estimated probability that a person will choose automobile transportation given that $DTIME = 1$.

b. Using the probit model results in Example 16.4, calculate the estimated probability that a person will choose automobile transportation given that $DTIME = 1$. How does this result compare to the logit estimate? [*Hint*: Recall that Statistical Table 1 gives cumulative probabilities for the standard normal distribution.]

c. Using the logit model results, compute the estimated marginal effect of an increase in travel time of 10 minutes for an individual whose travel time is currently 30 minutes longer by bus (public transportation). Using the linear probability model results, compute the same marginal effect estimate. How do they compare?

d. Using the logit model results, compute the estimated marginal effect of a decrease in travel time of 10 minutes for an individual whose travel time is currently 50 minutes longer by driving. Using the probit results, compute the same marginal effect estimate. How do they compare?

16.2 In Appendix 16A.1, we illustrate the calculation of a standard error for the marginal effect in a probit model of transportation, Example 16.4. In the appendix, the calculation is for the marginal effect when it currently takes 20 minutes longer to commute by bus ($DTIME = 2$).

a. Repeat the calculation for the probit model when $DTIME = 1$. [*Hint*: The values of the standard normal *pdf* are given in Statistical Table 6.]

b. Using the probit model, construct a 95% interval estimate for the marginal effect of a 10-minute increase in travel time by bus when $DTIME = 1$.

c. The logit model estimates and standard errors are

$$\underset{(\text{se})}{\tilde{\gamma}_1 + \tilde{\gamma}_2 DTIME} = \underset{(0.7505)}{-0.2376} + \underset{(0.2064)}{0.5311}DTIME$$

The estimated coefficient covariance is $\widehat{\text{cov}}(\tilde{\gamma}_1, \tilde{\gamma}_2) = -0.025498$. Calculate the standard error of the marginal effect of a 10-minute increase in travel time when $DTIME = 1$. [*Hint*: Carry through the steps in Appendix 16A.1 using equation (16.17) in place of $\Phi(\cdot)$ and equation (16.16) in place of $\phi(\cdot)$.]

d. Construct a 95% interval estimate for the marginal effect of a 10-minute increase in travel time by bus, when $DTIME = 1$ for the logit model.

16.3 In Example 16.3, we illustrate the calculation of the likelihood function for the probit model in a small example.

- **a.** Calculate the probability that $y = 1$ if $x = 1.5$, given the values of the maximum likelihood estimates.
- **b.** Using the threshold 0.5 and the result in part (a), predict the value of y if $x = 1.5$, the first observation, given the values of the maximum likelihood estimates. Does your prediction agree with the actual outcome $y = 1$?
- **c.** Calculate the value of the likelihood function, illustrated in equation (16.14), using the given $N = 3$ data pairs, if the parameter values are $\beta_1 = -1$ and $\beta_2 = 0.2$. Compare this value to the value of the likelihood function evaluated at the maximum likelihood estimates, given in Example 16.3. Which is larger?
- **d.** For the probit model, the value of the likelihood function (16.14) will always be between zero and one. True or false? Explain.
- **e.** For the probit model, the value of the log-likelihood function (16.15) will always be negative. True or false? Explain.

16.4 In Example 16.3, we illustrate the calculation of the likelihood function for the probit model in a small example. In this exercise, we will repeat that example using logit instead of probit. The logit model for the same example is $P(y = 1) = \Lambda(\gamma_1 + \gamma_2 x)$, where $\Lambda(\bullet)$ is the logistic *cdf* in equation (16.7). The maximum likelihood estimates of the parameters are $\tilde{\gamma}_1 + \tilde{\gamma}_2 x = -1.836 + 3.021x$. The maximized value of the log-likelihood function is -1.612.

- **a.** Calculate the probability that $y = 1$ if $x = 1.5$, given the values of the maximum likelihood estimates.
- **b.** Using the threshold 0.5 and the result in part (a), predict the value of y if $x = 1.5$, the first observation, given the values of the maximum likelihood estimates. Compare your prediction to the actual outcome $y = 1$ in the first observation.
- **c.** Calculate the value of the likelihood function, illustrated in equation (16.14) but substituting equation (16.17) in place of $\Phi(\bullet)$ and using the given $N = 3$ data pairs, if the parameter values are $\gamma_1 = -1$ and $\gamma_2 = 2$. Compare this value to the value of the likelihood function evaluated at the maximum likelihood estimates. Which is larger?
- **d.** For the logit model, the value of the likelihood function (16.14), with $\Lambda(\bullet)$ in place of $\Phi(\bullet)$, will always be between zero and one. True or false? Explain.
- **e.** For the logit model, the value of the log-likelihood function (16.15), with $\Lambda(\bullet)$ in place of $\Phi(\bullet)$, will always be negative. True or false? Explain.

16.5 We are given three observations on binary choice with $y_1 = 1, y_2 = 1, y_3 = 0$. Consider a logit model with only an intercept, $P(y = 1) = \Lambda(\gamma_1)$, where $\Lambda(\bullet)$ is the logistic *cdf*.

- **a.** Show that the log-likelihood function is $\ln L(\gamma_1) = 2\ln\Lambda(\gamma_1) + \ln[1 - \Lambda(\gamma_1)]$.
- **b.** Show that $d\ln L(\gamma_1)/d\gamma_1 = 2\lambda(\gamma_1)/\Lambda(\gamma_1) - \lambda(\gamma_1)/[1 - \Lambda(\gamma_1)]$, where $\lambda(\cdot)$ is the logistic *pdf* in (16.14). [*Hint*: Use Derivative Rules 8 and 9 from Appendix A.3.]
- **c.** The value of γ_1 such that $d\ln L(\gamma_1)/d\gamma_1 = 0$ is the maximum likelihood estimator $\tilde{\gamma}_1$. True, false, or maybe?
- **d.** It can be shown that for the logit model $\ln L(\gamma_1)$ is strictly concave, meaning that the second derivative is negative for all values of γ_1 or $d^2\ln L(\gamma_1)/d\gamma_1{}^2 < 0$. What is your answer to (c) now? [*Hint*: See Appendix A.3.4.]
- **e.** Setting the derivative in (c) to zero and solving, show that $\Lambda(\tilde{\gamma}_1) = 2/3$. [*Note*: This does not require you to first solve for $\tilde{\gamma}_1$.]
- **f.** Now, solve the condition in (c) to show that $\tilde{\gamma}_1 = -\ln(1/2)$.

16.6 In this exercise, we generalize the results in Exercise 16.5. Consider a logit model with only an intercept, $P(y = 1) = \Lambda(\gamma_1)$, where $\Lambda(\bullet)$ is the logistic *cdf*. Suppose in a sample of N observations, there are N_1 values $y_i = 1$ and N_0 values $y_i = 0$.

- **a.** Show that the logit log-likelihood function is $\ln L(\gamma_1) = N_1\ln\Lambda(\gamma_1) + N_0\ln[1 - \Lambda(\gamma_1)]$.
- **b.** Show that $d\ln L(\gamma_1)/d\gamma_1 = N_1\lambda(\gamma_1)/\Lambda(\gamma_1) - N_0\lambda(\gamma_1)/[1 - \Lambda(\gamma_1)]$, where $\lambda(\cdot)$ is given in (16.6). [*Hint*: Use Derivative Rules 8 and 9 from Appendix A.3.]
- **c.** Setting the derivative in (b) to zero and solving, show that $\Lambda(\tilde{\gamma}_1) = N_1/N$. What is the interpretation of N_1/N? [*Note*: This does not require you to first solve for $\tilde{\gamma}_1$, the MLE.]

d. Using (c) show that $\ln L(\tilde{\gamma}_1) = N_1 \ln(N_1/N) + N_0 \ln(N_0/N)$.

e. Show that a probit model, $P(y = 1) = \Phi(\gamma_1)$, where $\Phi(\bullet)$ is the standard normal *cdf*, results in the same value for the log-likelihood as in part (d).

16.7 Exercise 16.5 shows that given the three observations on binary choice with $y_1 = 1, y_2 = 1, y_3 = 0$ the maximum likelihood estimator of the logit model $P(y = 1) = \Lambda(\gamma_1)$ is $\tilde{\gamma}_1 = -\ln(1/2) = 0.6931472$ and that $\Lambda(\tilde{\gamma}_1) = 2/3$.

a. Using these results show that $\ln L(\tilde{\gamma}_1) = 2\ln\Lambda(\tilde{\gamma}_1) + \ln[1 - \Lambda(\tilde{\gamma}_1)] = -1.9095425$.

b. Using the data in Example 16.3, and the logit model $P(y = 1|x) = \Lambda(\gamma_1 + \gamma_2 x)$, we find that the maximum likelihood estimates of the parameters are $\tilde{\gamma}_1 + \tilde{\gamma}_2 x = -1.836 + 3.021x$, and the maximized value of the log-likelihood function is -1.612. Using these results, and those in (a), carry out the likelihood ratio test of H_0: $\gamma_2 = 0$ versus H_1: $\gamma_2 \neq 0$ at the 5% level of significance.

c. Calculate the *p*-value for the test in (b).

16.8 Consider a probit model designed to explain the choice by homebuyers of fixed versus adjustable rate mortgages. The explanatory variables, with sample means in parentheses, are *FIXRATE* (13.25) = fixed interest rate; *MARGIN* (2.3) = the variable rate − the fixed rate; and *NETWORTH* (3.5) = borrower's net worth ($100,000 units). The dependent variable is *ADJUST* (0.41) = 1 if an adjustable mortgage is chosen. The coefficient estimates, in Table 16.8, use 78 observations over the period January, 1983 to February, 1984.

TABLE 16.8 **Estimates for Exercise 16.8**

	C	*FIXRATE*	*MARGIN*	*NETWORTH*	ln*L*(Model)	ln*L*(*C*)
Model 1	−7.0166	0.5301	−0.2675	0.0864	−42.0625	−52.8022
(se)	(3.3922)	(0.2531)	(0.1304)	(0.0354)		
Model 2	−9.8200	0.7535	−0.1945		−45.1370	−52.8022
(se)	(3.1468)	(0.2328)	(0.1249)			

a. What information is provided by the signs of the estimated coefficients of Model 1? Are the signs consistent with economic reasoning? Which coefficients are significant at the 5% level?

b. Carry out a likelihood ratio test of the model significance at the 1% level for Model 1. In Table 16.8, ln*L*(Model) is the log-likelihood of the full model and ln*L*(*C*) is the log-likelihood of the model including only the constant term.

c. What is the estimated probability of a borrower choosing an adjustable rate mortgage if *FIXRATE* = 12, *MARGIN* = 2, and *NETWORTH* = 3? What is the estimated probability of a borrower choosing an adjustable rate mortgage if *FIXRATE* = 12, *MARGIN* = 2, and *NETWORTH* = 10?

d. Carry out a likelihood ratio test of the hypothesis that *NETWORTH* has no effect on the choice of mortgage type, against the alternative that it does, at the 1% level.

e. Using Model 2, what is the marginal effect of *MARGIN* on the probability of choosing an adjustable rate mortgage if *FIXRATE* = 12 and *MARGIN* = 2?

f. Using Model 2, calculate the discrete change in the probability of choosing an adjustable rate mortgage if *MARGIN* increases from 2% to 4%, while *FIXRATE* remains 12%? Is the value twice the value found in part (e)?

16.9 Consider a probit model explaining the choice to attend college by high-school graduates. Define *COLLEGE* = 1 if a high-school graduate chooses either a 2-year or 4-year college, and zero otherwise. We use explanatory variables *GRADES*, 13 point scale with 1 indicating highest grade (A+) and 13 the lowest (F); *FAMINC*, gross family income in $1000 units; and *BLACK* = 1 if black. Using a sample of $N = 1000$ graduates the estimated model is

$$\underset{}{P(COLLEGE = 1) = \Phi(}\underset{}{2.5757} \underset{(0.0265)}{- 0.3068GRADES} \underset{(0.0017)}{+ 0.0074FAMINC} \underset{(0.2177)}{+ 0.6416BLACK}\underset{}{)}$$
$$\text{(se)}$$

a. What information is provided by the signs of the estimated coefficients? Which coefficients are statistically significant at the 5% level?

b. Estimate the probability of attending college for a white student with *GRADES* = 2 (A) and *FAMINC* = 50 ($50,000). Repeat this probability calculation if *GRADES* = 5 (B).

c. Estimate the probability of attending college for a black student with *GRADES* = 5 (B) and *FAMINC* = 50 ($50,000). Compare this probability to the comparable probability for a white student calculated in part (b).

d. Calculate the marginal effect of an increase in family income of $1000 on the probability of attending college for a white student with *GRADES* = 5 (B).

e. The log-likelihood for the model estimated above is −423.36. Omitting *FAMINC* and *BLACK* the log-likelihood of the estimated probit model is −438.26. Test the joint significance of *FAMINC* and *BLACK* at the 1% level of significance using a likelihood ratio test.

16.10 Consider a probit model explaining the choice to attend a 4-year college rather than a 2-year college by high-school graduates who chose to attend a postsecondary school. Define *FOURYR* = 1 if a high-school graduate chooses 4-year college and *FOURYR* = 0 if the high school graduate chooses a 2-year college. We use explanatory variables *GRADES*, 13 point scale with 1 indicating highest grade (A+) and 13 the lowest (F); *FAMINC*, gross family income in $1000 units; and *HSCATH* = 1 if the student attended a Catholic high school and *HSCATH* = 0 otherwise. Table 16.9 contains some probit model estimates.

TABLE 16.9 Estimates for Exercise 16.10

Model	(1)	(2)	(3)	(4)	(5)
				HSCATH = 0	*HSCATH* = 1
C	1.6395	1.6299	1.6039	1.6039	2.3143
	(23.8658)	(23.6925)	(22.5893)	(22.5893)	(8.0379)
GRADES	−0.2350	−0.2357	−0.2344	−0.2344	−0.2603
	(−25.1058)	(−25.1437)	(−24.2364)	(−24.2364)	(−6.7691)
FAMINC	0.0042	0.0040	0.0043	0.0043	0.0015
	(8.2798)	(7.6633)	(7.7604)	(7.7604)	(1.0620)
HSCATH		0.3645	0.7104		
		(5.0842)	(2.3954)		
HSCATH × *GRADES*			−0.0259		
			(−0.6528)		
HSCATH × *FAMINC*			−0.0028		
			(−1.9050)		
N	5254	5254	5254	4784	470
ln*L*	−2967.91	−2954.50	−2952.68	−2735.14	−217.54

t-statistics in parentheses.

a. Using Model (2), how large an effect on the probability of attending a 4-year college does attending a catholic high school have for a student with *GRADES* = 5 (B) and family income of $100,000.

b. Comparing Models (2) and (3), are the interaction variables *HSCATH* × *GRADES* and *HSCATH* × *FAMINC* jointly significant at 5% using a likelihood ratio test?

c. Can we interpret the Model (3) results as saying an increase in family income reduces the probability of attending a 4-year college for someone graduating from a Catholic high school? What is the marginal effect of an additional $1000 in family income for a Catholic high school student with *GRADES* = 5 (B) and family income of $50,000?

d. Using Model (3), compute the probability of attending a 4-year college for someone graduating from a Catholic high school with *GRADES* = 5 (B) and family income of $100,000. Compare this probability to a student who did not attend a Catholic high school but has *GRADES* = 5 (B) and family income of $100,000.

e. Using Models (1) and (3), test the null hypothesis that the probit model parameters are the same for students who attend and do not attend a Catholic high school. Use a likelihood ratio test at the 5% level of significance.

f. Using Models (4) and (5), estimate the probit model separately for *HSCATH* = 0 and *HSCATH* = 1. Compute the sum of the log-likelihood functions values. Compare the sum to the log-likelihood for Model (3). Algebraically show that this is not an accident.

16.11 Using data on $N = 4{,}642$ infant births, we estimate a probit model with dependent variable *LBWEIGHT* = 1 if it is a low birthweight baby and 0 otherwise, *MAGE* is the mother's age, *PRENATAL1* = 1 if first prenatal visit is in 1 trimester and 0 otherwise, and *MBSMOKE* = 1 if the mother smoked and 0 otherwise. The results are in Table 16.10.

TABLE 16.10 Probit Estimates for Exercise 16.11

	C	*MAGE*	*PRENATAL1*	*MBSMOKE*	$MAGE^2$
Model 1	−1.2581	−0.0103	−0.1568	0.3974	
(se)	(0.1436)	(0.0054)	(0.0710)	(0.0670)	
Model 2	−0.1209	−0.1012	−0.1387	0.4061	0.0017
(se)	(0.4972)	(0.0385)	(0.0716)	(0.0672)	(0.0007)

a. In Model 1, comment on estimated signs and significance of the coefficients on *PRENATAL1* and *MBSMOKE*.

b. Using Model 1, calculate the marginal effect on the probability of a low birthweight baby given an increase in the mother's age by 1 year, for a woman who is 20 years old with *PRENATAL1* = 0 and *MBSMOKE* = 0. Repeat this calculation for a woman who is 50 years old. Do the results make sense?

c. Using Model 2, calculate the marginal effect on the probability of a low birthweight baby given an increase in the mother's age by 1 year, for a woman who is 20 years old with *PRENATAL1* = 0 and *MBSMOKE* = 0. Repeat this calculation for a woman who is 50 years old. Compare these results to those in part (b).

d. Using Model 2, calculate the impact of a prenatal visit in the first trimester on the probability of having a low birthweight baby for a woman who is 30 years old and smokes.

e. Using Model 2, calculate the impact of a mother smoking on the probability having a low birthweight baby given that she is 30 years old and had a prenatal visit in the first trimester.

f. Using Model 2, calculate the age at which the probability of a low birthweight baby is a minimum.

16.12 This exercise is an extension of Example 16.12 using the larger data set *nels* with 6,649 observations. Two estimated multinomial logit models are reported in Table 16.11. In addition to the variable *GRADES*, we have *FAMINC* = family income ($1000 units) and indicator variables for sex and race. The baseline group is students who chose not to attend college.

TABLE 16.11 Estimates for Exercise 16.12

	Model 1		Model 2	
PSECHOICE	Coefficient	*t*-value	Coefficient	*t*-value
2				
C	1.7101	9.3293	1.9105	11.1727
GRADES	−0.2711	−13.1969	−0.2780	−13.9955
FAMINC	0.0124	8.3072	0.0116	8.0085
FEMALE	0.2284	3.0387		
BLACK	0.0554	0.4322		
3				
C	4.6008	25.7958	4.6111	27.8351
GRADES	−0.6895	−32.2723	−0.6628	−32.3721
FAMINC	0.0200	13.5695	0.0183	12.9450
FEMALE	0.0422	0.5594		
BLACK	0.9924	8.0766		
ln(*L*)	−5699.8023		−5751.5982	

a. Which estimated coefficients are significant in Model 1? Based on the *t*-values, should we consider dropping *FEMALE* and *BLACK* from the model?

b. Compare the results of Model 1 to Model 2 using a likelihood ratio test. Using the $\alpha = 0.01$ level of significance, can we reject the null hypothesis that the Model 1 coefficients of *FEMALE* and *BLACK* are zero?

c. Compute the estimated probability that a white male student with *GRADES* = 5 (B) and *FAMINC* of $100,000 will attend a 4-year college.

d. Compute the odds, or probability ratio, that a white male student with *GRADES* = 5 (B) and *FAMINC* of $100,000 will attend a 4-year college rather than not attend any college.

e. Compute the change in probability of attending a 4-year college for a white male student with median *FAMINC* =$100,000 whose *GRADES* change from 5 (B) to 2 (A).

16.13 This exercise is an extension of Example 16.13. It is a conditional logit model of choice among 3 brands of soda: Coke, Pepsi, and 7-Up. The data are in the data file *cola*. As in the example, we choose Coke to be the base alternative, setting its alternative specific constant (intercept) to zero. We add to the model indicator variables *FEATURE*, indicating whether the product was "featured" at the time, and *DISPLAY* for whether there was a store display at the time of purchase. The model estimates are in Table 16.12.

TABLE 16.12 Estimates for Exercise 16.13

	Model 1		Model 2	
	Coefficient	*t*-Statistic	Coefficient	*t*-Statistic
PRICE	−1.7445	−9.6951	−1.8492	−9.8017
FEATURE	−0.0106	−0.1327	−0.0409	−0.4918
DISPLAY	0.4624	4.9700	0.4727	5.0530
PEPSI			0.2841	4.5411
7-UP			0.0907	1.4173
ln(*L*)	−1822.2267		−1811.3543	

a. Using Model 1, calculate the probability ratio, or odds, of choosing Coke relative to Pepsi if Coke costs \$1.25, Pepsi costs \$1.25, Coke has a display but Pepsi does not, and neither are featured. Note that the model contains no alternative specific constants.

b. Using Model 1, calculate the probability ratio, or odds, of choosing Coke relative to Pepsi if Coke costs \$1.25, Pepsi costs \$1.00, Coke has a display but Pepsi does not, and neither are featured.

c. Compute the change in the probability of purchase of each type of soda if the price of Coke changes from \$1.25 to \$1.50, with the prices of Pepsi and 7-Up remaining at \$1.25. Assume that a display is present for Coke, but not for the others, and none of the items is featured.

d. In Model 2, we add the alternative specific "intercept" terms for Pepsi and 7-Up to the Model 1. Calculate the probability ratio, or odds, of choosing Coke relative to Pepsi if Coke costs \$1.25, Pepsi costs \$1.25, Coke has a display but Pepsi does not, and neither are featured.

e. Using Model 2, compute the change in the probability of purchase of each type of soda if the price of Coke changes from \$1.25 to \$1.50, with the prices of Pepsi and 7-Up remaining at \$1.25. Assume that a display is present for Coke, but not for the others, and none of the items is featured.

f. The value of the log-likelihood function for the model in Example 16.13 is −1824.5621. Test the null hypothesis that the coefficients of the marketing variables, *FEATURE* and *DISPLAY*, are zero, against the alternative that they are not, using a likelihood ratio test with $\alpha = 0.01$.

16.14 In Example 16.14, we described an ordinal probit model for postsecondary education choice, and estimated a simple model in which the choice depended simply on the student's *GRADES*. Expand the ordered probit model to include family income (*FAMINC*, in \$1000), family size (*FAMSIZ*), the dummy variables *BLACK* and *PARCOLL* = 1 if a parent has at least a college degree, and 0 otherwise. The estimates of this model are Model 2 in Table 16.13.

TABLE 16.13 **Estimates for Exercise 16.14**

	Model 1		Model 2	
PSECHOICE	Coefficient	Standard Error	Coefficient	Standard Error
GRADES	−0.3066	0.0192	−0.2953	0.0202
FAMINC			0.0053	0.0013
FAMSIZ			−0.0241	0.0302
BLACK			0.7131	0.1768
PARCOLL			0.4236	0.1016
$\hat{\mu}_1$	−2.9456	0.1468	−2.5958	0.2046
$\hat{\mu}_2$	−2.0900	0.1358	−1.6946	0.1971
ln*L*	−875.8217		−839.8647	

a. Using the estimates in Table 16.13, Model 1, calculate the probability that a student will choose no college, a 2-year college, and a 4-year college if the student's grades are *GRADES* = 7 (B−). Recompute these probabilities assuming that *GRADES* = 3 (A−). Discuss the probability changes. Are they what you anticipated? Explain.

b. Discuss the Model 2 estimates, their signs and significance. [*Hint*: recall that the sign indicates the direction of the effect for the highest category but is opposite for the lowest category].

c. Test the joint significance of the variables added in (b) using a likelihood ratio test at the 1% level of significance.

d. Compute the probability that a black student from a household of four members with \$100,000 income, and with at least one parent having at least a college degree, so that *PARCOLL* = 1, will attend a 4-year college if (i) *GRADES* = 7 and (ii) *GRADES* = 3.

e. Repeat (d) for a "nonblack" student and discuss the differences in your findings.

16.15 Consider a Poisson regression explaining the number of Olympic Games medals won using data from 1988 (in Seoul, South Korea) and 1992 (in Barcelona, Spain) by various

countries as a function of $LPOP = \ln(POP)$ = the logarithm of population in millions, and $LGDP = \ln(GDP)$ = the logarithm of gross domestic product (in billions of 1995 dollars). That is, $E(MEDALTOT|\mathbf{X}) = \exp[\beta_1 + \beta_2 \ln(POP) + \beta_3 \ln(GDP)]$. The estimated coefficients, using 316 observations, are in Table 16.14, Model 1.

TABLE 16.14 **Estimates for Exercise 16.15**

	Model 1		Model 2	
	Coefficient	Standard Error	Coefficient	Standard Error
C	−1.4442	0.0826	−1.4664	0.0835
LPOP	0.2143	0.0217	0.2185	0.0219
LGDP	0.5556	0.0164	0.5536	0.0165
HOST			0.6620	0.1375

a. Using Model 1 results, what is the estimated impact on the number of medals won if *GDP* increases by 1%? [*Hint*: It can be shown (can you?) that β_3 is an elasticity.]

b. In 1996, Bulgaria had *GDP* = 11.8 billion and a population of 8.356 million. Estimate the expected number of medals that Bulgaria would win in the Olympics, held in Atlanta, USA. They did win 15 medals.

c. Calculate the probability that Bulgaria in 1996 would win one or fewer medals.

d. In 1996, Switzerland had *GDP* = 306 billion and a population of 6.875 million. Estimate the expected number of medals that Switzerland would win. They did win 1 medal.

e. Calculate the probability that Switzerland in 1996 would win one or fewer medals.

f. *HOST* is an indicator variable = 1 for the country hosting the Olympics. This variable is added in Model 2. Interpret its coefficient. [*Hint*: What is the estimated percentage change in the conditional mean?] Is the estimated effect large or small? Is the coefficient statistically significant at the 1% level?

g. In 1996, the Olympic games were held in the U.S. city of Atlanta, GA. In that year, the U.S. population was 265 million and its *GDP* was 7280 billion. Estimate the expected number of medals the United States would win using Model 1 and again using Model 2. The United States won 101 medals that year. Which model's estimated value was closer to the true outcome?

16.16 Consider a regression explaining the share of Olympic Games medals won by each country in 1988 (in Seoul, South Korea), 1992 (in Barcelona, Spain), and 1996 (in Atlanta, GA, USA) as a function of $LPOP = \ln(POP)$ = the logarithm of population in millions, $LGDP = \ln(GDP)$ = the logarithm of gross domestic product (in billions of 1995 dollars), and *HOST*, an indicator variable = 1 for the country hosting the Olympics. The total number of medals awarded in 1988 was 738; in 1992, there were 815 medals awarded, and in 1996, 842 medals were awarded. Using the total number of medals awarded, we compute the percentage share of medals (*SHARE*) won by each country.

a. The least squares estimates of $SHARE = \beta_1 + \beta_2 \ln(POP) + \beta_3 \ln(GDP) + \beta_4 HOST + e$ are in Table 16.15. Are the signs and significance of the coefficient estimates reasonable?

TABLE 16.15 **Estimates for Exercise 16.16**

	OLS			Tobit	
	Coefficient	Standard Error	HCE	Coefficient	Standard Error
C	−0.2929	0.1000	0.0789	−4.2547	0.3318
LPOP	−0.0058	0.0496	0.0352	0.1707	0.1135
LGDP	0.3656	0.0454	0.0579	0.9605	0.0973
HOST	4.1723	0.9281	2.0770	3.2475	1.4611
$\tilde{\sigma}$				2.4841	0.1273

b. Using the OLS estimates, what is the predicted effect of *GDP* on the expected share of medals won? That is, how much do we predict the share of medals won will change if *GDP* increases by 1%? Construct a 95% interval estimate of this effect.

c. For the model estimated by OLS, the robust Breusch-Pagan LM test statistic for heteroskedasticity as a function of ln(*GDP*) is $NR^2 = 32.80$. What can we conclude about the OLS estimator and the usual standard errors based on this test?

d. We also report the OLS heteroskedasticity robust standard errors (HCE) in Table 16.15. Construct a 95% interval estimate for the predicted effect of a 1% increase in *GDP* on the share of medals won using the robust standard errors.

e. Among the 508 countries competing in these summer Olympics, almost 62% won no medals. Does this cause any potential problems for the least squares estimator? By using robust standard errors in part (c), we have solved any problems with the OLS estimator. True or false? Explain your choice.

f. Compare the Tobit parameter estimates reported in Table 16.15 to the OLS estimates and standard errors. What are the differences? Is Tobit a reasonable estimator for the share of medals won in this example? Why?

g. Using the Tobit estimates, what is the estimated effect of *GDP* on the expected share of medals won for a nonhost country with $GDP = 150$ billion and $POP = 30$ million? That is, how much do we estimate the expected share of medals won will change if *GDP* increases by one percent? [*Hint*: In equation (16.35), let $y = SHARE$ and $x = \ln(GDP)$. Then

$$\partial E(SHARE|\mathbf{X})/\partial \ln(GDP) = \beta_3 \Phi\left[\frac{\beta_1 + \beta_2 \ln(POP) + \beta_3 \ln(GDP) + \beta_4 HOST}{\sigma}\right]$$

Also, $\partial \ln(GDP)/\partial GDP = 1/GDP$. Then refer to the analysis of the linear-log model in Section 4.3.3.]

16.8.2 Computer Exercises

16.17 In Chapter 7, we examined the Tennessee's Project STAR. In the experiment, children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular sized classes with 22–25 students, and regular sized classes with a full-time teacher aide to assist the teacher. In Example 7.11, we checked for random assignment of children to the three types of classes using a linear probability model, regressing the indicator *SMALL* (small class) on student characteristics. Let us reconsider this regression using logit rather than the linear probability model. If there is random assignment of children to types of classes, then we should not find any significant relationships. Use data file *star5_small2* for this exercise. The data file *star5* contains more observations.

a. Estimate a logit model with outcome variable *SMALL* and explanatory variables *BOY* and *BLACK*. Individually test the coefficients of these variables for significance. What do you find? Test the coefficients jointly for significance using the likelihood ratio test. What do you find? Can we reject the null hypothesis that assignment to small classes is done randomly?

b. Repeat the estimation and testing in part (a) using outcome variables *AIDE* and *REGULAR*. Do you find any evidence that students were not randomly assigned?

c. Add the variable *FREELUNCH* to the models in (a) and (b) and reestimate them. Do you find any evidence that there is a systematic pattern between class assignment and this variable?

d. Add the two variables *TCHWHITE* and *TCHMASTERS* to the models in (c) and reestimate them. In each, carry out a likelihood ratio test for the joint significance of *TCHWHITE* and *TCHMASTERS*. What do you conclude? In the experiment students were randomized within schools but not across schools. Does this offer any explanation of your findings? If so, how?

16.18 Mortgage lenders are interested in determining borrower and loan characteristics that may lead to delinquency or foreclosure. In the data file *lasvegas* are 1000 observations on mortgages for single family homes in Las Vegas, Nevada during 2008. The variable of interest is *DELINQUENT*, an indicator variable = 1 if the borrower missed at least three payments (90 + days late), but 0 otherwise. Explanatory variables are *LVR* = the ratio of the loan amount to the value of the property; *REF* = 1 if purpose of the loan was a "refinance" and = 0 if loan was for a purchase; *INSUR* = 1 if mortgage carries mortgage insurance, 0 otherwise; *RATE* = initial interest rate of the mortgage;

AMOUNT = dollar value of mortgage (in \$100,000); *CREDIT* = credit score, *TERM* = number of years between disbursement of the loan and the date it is expected to be fully repaid, *ARM* = 1 if mortgage has an adjustable rate, and = 0 if mortgage has a fixed rate.

a. Estimate the linear probability (regression) model explaining *DELINQUENT* as a function of the remaining variables. Use White heteroskedasticity robust standard errors. Are the signs of the estimated coefficients reasonable?

b. Use logit to estimate the model in (a). Are the signs and significance of the estimated coefficients the same as for the linear probability model?

c. Compute the predicted value of *DELINQENT* for the 500th and 1000th observations using both the linear probability model and the logit model. Interpret the values.

d. Construct a histogram of *CREDIT*. Using both linear probability and logit models, calculate the probability of delinquency for *CREDIT* = 500, 600, and 700 for a loan of \$250,000 (*AMOUNT* = 2.5). For the other variables, let the loan to value ratio (*LVR*) be 80%, the initial interest rate is 8%, all indicator variables take the value 0, and *TERM* = 30. Discuss similarities and differences among the predicted probabilities from the two models.

e. Using both linear probability and logit models, compute the marginal effect of *CREDIT* on the probability of delinquency for *CREDIT* = 500, 600, and 700, given that the other explanatory variables take the values in (d). Discuss the interpretation of the marginal effect.

f. Construct a histogram of *LVR*. Using both linear probability and logit models, calculate the probability of delinquency for *LVR* = 20 and *LVR* = 80, with *CREDIT* = 600 and other variables set as they are in (d). Compare and contrast the results.

g. Compare the percentage of correct predictions from the linear probability model and the logit model using a predicted probability of 0.5 as the threshold.

h. As a loan officer, you wish to provide loans to customers who repay on schedule and are not delinquent. Suppose you have available to you the first 500 observations in the data on which to base your loan decision on the second 500 applications (501–1,000). Is using the logit model with a threshold of 0.5 for the predicted probability the best decision rule for deciding on loan applications? If not, what is a better rule?

16.19 Mortgage lenders are interested in determining borrower and loan factors that may lead to delinquency or foreclosure. In the data file *vegas5_small* are 1000 observations on mortgages for single family homes in Las Vegas, Nevada during 2010. (The data file *vegas5* contains 10,000 observations.) The variable of interest is *DEFAULT*, an indicator variable = 1 if the borrower's payment was 90 + days late, but 0 otherwise. Explanatory variables are *ARM* = 1 if it's an adjustable rate mortgage, 0 if fixed; *REFINANCE* = 1 if loan is for a refinance of any type, 0 if for purchase; *LIEN2* = 1 if there is a second lien mortgage, 0 if it is the first lien; *TERM30* = 1 if it is a 30-year mortgage, 0 if 15-year mortgage; *UNDERWATER* = 1 if borrower estimated to owe more than the property is worth at time of origination, 0 otherwise; *LTV* = loan to value ratio of property at origination, percent; *RATE* = current interest rate on loan, percent; *AMOUNT* = loan amount in \$10,000 units; and *FICO* = borrower's credit score at origination.

a. Estimate the linear probability (regression) model explaining *DEFAULT* as a function of the remaining variables. Use White robust standard errors. Are the signs of the estimated coefficients reasonable?

b. Use probit to estimate the model in (a). Are the signs and significance of the estimated coefficients the same as for the linear probability model?

c. Compute the predicted value of *DEFAULT* for the 500th and 1000th observations using both the linear probability model and the probit model. Interpret the values.

d. Construct a histogram of *FICO*. Using both linear probability and probit models, calculate the probability of default for *FICO* = 500, 600, and 700 for a loan of \$250,000 (*AMOUNT* = 25). For the other variables, the loan to value ratio (*LTV*) is 80%, initial interest rate is 8%, indicator variables take the value 0 except for *TERM30* = 1. Discuss similarities and differences among the predicted probabilities from the two models.

e. Using both linear probability and probit models, compute the marginal effect of *FICO* on the probability of delinquency for *FICO* = 500, 600, and 700, given that the other explanatory variables take the values in (d). Discuss the interpretation of the marginal effect.

f. Construct a histogram of *LTV*. Using both linear probability and probit models, calculate the probability of delinquency for *LVR* = 20 and *LVR* = 80, with *FICO* = 600 and other variables set as they are in (d). Compare and contrast the results.

g. Compare the percentage of correct predictions from the linear probability model and the probit model using a predicted probability of 0.5 as the threshold.

h. As a loan officer, you wish to provide loans to customers who repay on schedule and are not delinquent. Suppose you have available to you the first 500 observations in the data on which to base your loan decision on the second 500 applications (501-1,000). Is using the probit model with a threshold of 0.5 for the predicted probability the best decision rule for deciding on loan applications? If not, what is a better rule? [*Note*: for *vegas5* use the first 5000 observations for the estimation sample and the second 5000 observations for prediction.]

16.20 This exercise deals with the loan data in the data file *lasvegas* described in Exercise 6.18. The "Chow" test was introduced in Section 7.2.3 for testing the equality of coefficients in two regressions on subsets of observations. Here we ask a similar question concerning the parameters of the logit model for delinquency for the two subpopulations of borrowers who either have mortgage insurance ($INSUR = 1$) or not ($INSUR = 0$).

a. Using all observations, estimate the logit model for *DELINQUENT* using all explanatory variables except *INSUR*. Call the value of the log-likelihood function evaluated at the maximum likelihood estimates $\ln L_R$.

b. Reestimate the model in (a) using the sample observations for which $INSUR = 0$. Call the value of the log-likelihood function evaluated at the maximum likelihood estimates $\ln L_0$.

c. Reestimate the model in (b) using the sample observations for which $INSUR = 1$. Call the value of the log-likelihood function evaluated at the maximum likelihood estimates $\ln L_1$.

d. Compare the estimates from the models in (a–c). What major differences in coefficient signs, magnitudes, and significance do you observe?

e. Reestimate the model in (a) including each explanatory variable, as well as *INSUR*, and its interactions with all the other variables. Compare the value of the log-likelihood function from the fully interacted model, call it $\ln L_U$, to $\ln L_0 + \ln L_1$. If you have done things correctly, then $\ln L_U$ should equal $\ln L_0 + \ln L_1$. Can you explain why this must be so?

f. Carry out a likelihood ratio version of the Chow test by computing $LR = 2(\ln L_U - \ln L_R)$. What is the appropriate critical value for a test at the 5% level of significance? What conclusion do you draw about the subgroups of individuals who do and do not have mortgage insurance? Do the two groups behave in the same way?

16.21 Data on 1500 purchases of canned lite tuna are in the data file *tunafish*. There are four brands of tuna (Starkist – water, Starkist – oil, Chicken of the Sea – water, Chicken of the Sea – oil). The A.C. Nielsen data were made available through the University of Chicago's Graduate School of Business. The data file *tunafish_small* is a smaller dataset with 250 purchases. The data are in "stacked" format with four data lines per purchase, one for each tuna brand. The consumer choice is indicated by the indicator variable *CHOICE*. Relevant variables are *NETPRICE* = price minus coupon value, if used; $DISPLAY = 1$ if product is on display, $FEATURE = 1$ if item is featured, and *INCOME* = household income.

a. What is the primary variable-type distinction between *NETPRICE* and *INCOME*?

b. What is the sample percentage of purchases for each brand? What do you observe about consumer preferences for these product choices?

c. Using the conditional logit model, write the probability of choosing each brand using as explanatory variables *NETPRICE*, *DISPLAY*, and *FEATURE*, plus an alternative specific constant using Starkist packed in water as the base category.

d. Estimate the model specified in part (c).

e. For the model in (d) find the marginal effect of *NETPRICE* on the probability of choice of each brand, using for all brands $DISPLAY = FEATURE = 1$. Do these marginal effects have the signs you anticipate? Are the marginal effects statistically significant?

f. Add the variable *INCOME* to the model specified in (c). Perform a likelihood ratio test of its significance.

g. For the model in (f) find the marginal effect of *NETPRICE* on the probability of choice of each brand, using for all brands $DISPLAY = FEATURE = 1$ and $INCOME = 30$.

16.22 How do age, education, and other personal characteristics predict our assessment of our health status? Use the data file *rwm88* to answer the following.

a. Tabulate the values of the variable *HSAT3*, which is a self-rating of health satisfaction, with 1 being the lowest and 3 being highest. What percentages fall into each of the health status categories?

b. Estimate an ordered probit model predicting *HSAT3* using *AGE*, AGE^2, *EDUC2* = years of education, *FEMALE* = 1 if female, *MARRIED* = 1 if married, *HHKIDS* =1 if there are children under age 16 in the household, and *WORKING* = 1 if employed, 0 otherwise. Which variables have coefficients that are statistically significant at the 5% level?

c. Estimate the probability that an employed, unmarried male, age 40 with 16 years of education, and no children, will have health satisfaction *HSAT3* = 2.

d. Estimate the probability that an employed, unmarried male, age 50 with 16 years of education, and no children, will have health satisfaction *HSAT3* = 2.

e. Estimate the probability that an employed, unmarried male, age 40 with 16 years of education, and no children, will have health satisfaction *HSAT3* = 3.

f. Estimate the probability that an employed, unmarried male, age 50 with 16 years of education, and no children, will have health satisfaction *HSAT3* = 3.

g. Estimate the probability that an unemployed, unmarried male, age 50 with 16 years of education, and no children, will have health satisfaction *HSAT3* = 2. Compare this probability to the result in part (d).

h. Estimate the probability that an unemployed, unmarried male, age 50 with 16 years of education, and no children, will have health satisfaction *HSAT3* = 3. Compare this probability to the result in part (f).

16.23 How well do age, education, and other personal characteristics predict our assessment of our health status? Use the data file *rwm88* to answer the following.

a. Tabulate the variable *HSAT3*, which is a self-rating of health satisfaction, with 1 being the lowest and 3 being highest. What percent of the sample assess their health status as *HSAT3* = 1, 2, or 3?

b. Estimate an ordered probit model predicting *HSAT3* using *AGE*, AGE^2, *EDUC2* = years of education, and *WORKING* = 1 if employed, 0 otherwise. Which variables have coefficients that are statistically significant at the 5% level?

c. Estimate the marginal impact of age on the probabilities of health satisfactions *HSAT3* = 1, 2, or 3 for someone age 40, with 16 years of education, and who is working.

d. Estimate the marginal impact of age on the probabilities of health satisfactions *HSAT3* = 1, 2, or 3 for someone age 70, with 16 years of education, and who is working.

e. Estimate the marginal impact of *WORKING* on the probabilities of health satisfactions *HSAT3* = 1, 2, or 3 for someone age 40, with 16 years of education.

16.24 Consider household expenditures per person on apparel. Use the data file *cex5* for this exercise.

a. What percentage of the households spent nothing on apparel in the previous quarter?

b. Estimate a linear regression with *APPAR* as dependent variable and use as explanatory variables *INCOME*, *SMSA* (*S*tandard *M*etropolitan *S*tatistical *A*rea = 1 if household lives in an urban area, and = 0 otherwise), *ADVANCED*, *COLLEGE*, and *OLDER* (= 1 if someone in the household is 65 years of age or older). Discuss the signs and significance of the estimated coefficients. Interpret the coefficient of *INCOME*. Interpret the coefficient of *ADVANCED*.

c. Repeat the estimation in (b) using only observations for which *APPAR* > 0. What are your answers to the questions in (b) now?

d. Create the variable *SHOP* = 1 if *APPAR* > 0, and *SHOP* = 0 otherwise. Estimate a probit model with dependent variable *SHOP* as a function of the variables in (b). What factors significantly affect the decision to buy clothing?

e. Estimate a Tobit model with dependent variable *APPAR*. Compare the coefficient estimates signs and significance to those in (b) and (c). Calculate the marginal effect of income on the expected amount spent on *APPAREL* for a household living in an urban area, with income $65,000, containing someone with an advanced degree and no one 65 or older in the household. Repeat the calculation for a household with $125,000 income.

16.25 Consider using data file *mroz* to estimate a model explaining a married woman's hours of work, *HOURS*, as a function of her education, *EDUC*, her experience, *EXPER*, and her husband's hours of work, *HHOURS*.

a. Use all observations to estimate the regression model

$$HOURS = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 HHOURS + e$$

Is OLS a consistent estimator in this case?

b. Use only the observations for which $HOURS > 0$ to estimate the regression model in (a). Is OLS a consistent estimator in this case?

c. Estimate a probit model for the woman's decision to be in the labor force, $LFP = 1$, $LFP = \Phi(\gamma_1 + \gamma_2 EXPER + \gamma_3 KIDSL6 + \gamma_4 KIDS618 + \gamma_5 MTR + \gamma_6 LARGECITY)$. Which if any of the variables help explain the woman's labor force participation decision?

d. Using the estimates from the probit model, obtain

$$\tilde{w} = \tilde{\gamma}_1 + \tilde{\gamma}_2 EXPER + \tilde{\gamma}_3 KIDSL6 + \tilde{\gamma}_4 KIDS618 + \tilde{\gamma}_5 MTR + \tilde{\gamma}_6 LARGECITY$$

Create the inverse Mills ratio $\tilde{\lambda} = \phi(\tilde{w})/\Phi(\tilde{w})$. What are the sample mean and variance of $\tilde{\lambda}$?

e. Estimate the model $HOURS = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 HHOURS + \beta_\lambda \tilde{\lambda} + e$ using the observations for which $HOURS > 0$. Compare these estimates to those in parts (a) and (b). Are the standard errors from this estimation correct?

f. Estimate the model in (e) using heteroskedasticity robust standard errors. Use the option HC3 if it is available. These standard errors are not absolutely correct but an improvement over the ones in (e).

g. Estimate the model in (e) using bootstrap standard errors, with $B = 400$ bootstrap replications. Compare these standard errors to those in (e) and (f).

h. Estimate the model in (e) using proper econometric software for this Heckit model. Compare the results to those in (e)–(g). Be sure to identify whether your software is using a two-step estimator, like part (e), or full information maximum likelihood.

16.26 In Example 7.11, we used the linear probability model to check whether students were assigned randomly to small classes in Project STAR. In this exercise, we use multinomial logit and the data file *star* to explore the issue.

a. Create the variable $CLASS = 1$ for a regular sized class, $CLASS = 2$ for a small class, and $CLASS = 3$ for a regular sized class with a teacher aide. What percentage of the students in the sample were assigned to each type of class?

b. Estimate a multinomial logit model explaining *CLASS* with explanatory variables *BOY*, *WHITE_ASIAN*, *BLACK*, *FREELUNCH*, *SCHURBAN*, and *SCHRURAL*. Use $CLASS = 1$, the regular class, as the base group. If students are assigned randomly what values should the model coefficients take? Are any of the estimated coefficients significantly different from zero at the 5% level?

c. Find the ratio of the probability of being in a small class for a white boy who receives lunch if his school is in a rural area, relative to the probability of him being in a regular sized class.

d. Find the ratio of the probability of being in a regular sized class with a teacher aide for a white boy who receives lunch if his school is in a rural area, relative to the probability of him being in a regular sized class.

e. Carry out a likelihood ratio test that the coefficients of *BOY*, *WHITE_ASIAN*, *BLACK*, *FREELUNCH*, and *SCHURBAN* are zero, against the alternative that they are not, at the 5% level. What is the 5% critical value for this test?

f. Carry out a likelihood ratio test that the coefficients of *BOY*, *WHITE_ASIAN*, *BLACK*, *FREELUNCH*, *SCHURBAN,* and *SCHRURAL* are zero, against the alternative that they are not, at the 5% level. What is the 5% critical value for this test?

g. Based on the outcomes of parts (a)–(f), what do you conclude about random assignment of students in Project STAR?

16.27 In Example 16.15, we considered a count data model for the number of doctor visits by an individual as a function of a few explanatory variables. In this exercise, we expand the analysis using a larger data set in the data file, *rwm88*, and more explanatory variables. Adjust the data in the following ways: (i) omit individuals for whom $HHNINC2 = 0$; (ii) create the variable $LINC = ln(HHNINC2)$; (iii) create $AGE2 = AGE^2$; (iv) create the variable $POST = 1$ (a postsecondary degree indicator variable) if $FACHHS = 1$ or if $UNIV = 1$, and $POST = 0$ otherwise.

a. Using the first 3000 observations estimate a Poisson model explaining *DOCVIS* as a function of *FEMALE*, *AGE*, *AGE2*, *SELF*, *LINC*, *POST*, and *PUBLIC*. Discuss the signs and the significance of the coefficients on *FEMALE*, *SELF*, *POST*, and *PUBLIC*. Calculate the percentage increase in the expected number of doctor visits for each factor represented by these indicator variables.

b. Compute the estimated percentage change in the expected number of doctor visits associated with another year of age for a person who is 30 years old; who is 50 years old; and who is 70 years old.

c. Interpret the estimated coefficient of *LINC*.

d. Calculate the expected number of doctor visits for each person, *EDOCVIS*, and round this value to the nearest integer to obtain *NVISITS*, the predicted number of visits for each person. Create a variable that indicates a successful prediction. Let *SUCCESS* = 1 if *NVISITS* = *DOCVIS* and *SUCCESS* = 0 otherwise. What is the percentage of successful predictions for observations 1–3000? What is the percentage of successful predictions for the remaining 979 observations?

e. Create *SUCCESS1* which indicates a successful prediction of more than one doctor visit. That is, create a variable *DOCVIS1* = 1 if an individual has more than one doctor visit, and *PREDICT1* = 1 if the model has predicted more than one doctor visit. Let *SUCCESS1* = 1 if *DOCVIS1* = *PREDICT1* and *SUCCESS1* = 0 otherwise. What is the percentage of successful predictions of more than one doctor visit for observations 1–3000? What is the percentage of successful predictions of more than one doctor visit for the remaining 979 observations?

16.28 We have used Ray Fair's voting data, (data file *fair5*, throughout the book to predict presidential election outcomes with the linear regression model. Here we apply probit to predict the outcome of the 2016 U.S. Presidential election. Create the variable *DEMWIN* = 1 if *VOTE* $\geq$ 50.0 and *DEMWIN* = 0 otherwise. As of October 28, 2016, the values for the key economic variables were *GROWTH* = 0.97, *INFLAT* = 1.42, and *GOODNEWS* = 2.

a. Estimate a probit model for *DEMWIN* as a function of *GROWTH*, *INFLAT*, *GOODNEWS* using data for years prior to 2016. Comment on the signs and significance of the estimated coefficients.

b. Using the probit model in part (a), and the given values of *GROWTH*, *INFLAT*, and *GOODNEWS*, predict the election outcome in 2016. What is the estimated probability that a democrat will win?

c. Add *DPER*, *DUR*, *WAR*, and *INCUMB* to the model used in (a). Reestimate the probit model. What happens to the signs and significance of the estimated coefficients?

d. Using the model in (c), obtain the estimated probability, *PHAT*, of a democrat winning for the sample period 1916–2012. Are any of the predicted values very close to 1.0 or 0.0? For how many observations is *PHAT* > 0.99999? For how many observations is *PHAT* < 0.00001?

e. Examine the values of *DEMWIN* when the following four-variable pattern exists in the data: *DPER* = −1, *DUR* = 0, *WAR* = 0, *INCUMB* = −1. How many such observations are there? [*Note*: Some software will indicate probit failure when the dependent variable does not vary for a value of an independent variable, or in this case a particular combination of values. You may think of this as something like "perfect collinearity." When this happens maximum likelihood estimation including the particular pattern of observations fails.]

16.29 In this exercise, we illustrate some features of instrumental variables estimation, and two-stage least squares, when the potential endogenous variable is binary. Use the data file *rwm88* for this problem, and do not worry too much about the economic reasoning behind the model.

a. Estimate by OLS the regression of *DOCVIS* on *AGE*, *FEMALE*, *WORKING*, *HHNINC2*, and *ADDON*. Use heteroskedasticity robust standard errors. Does it appear that having add-on insurance is a significant factor affecting the number of doctor visits?

b. *ADDON* might be endogenous. Estimate a first stage equation using OLS with *ADDON* as dependent variable and *AGE*, *FEMALE*, *WORKING*, *HHNINC2*, *WHITEC*, and *SELF* as explanatory variables. Since the dependent variable is binary use heteroskedasticity robust standard errors. Are *WHITEC* and *SELF* jointly significant? Why does this matter if our objective is two-stage least squares estimation?

c. Obtain the fitted value from part (b), $\widehat{ADDON}$, and reestimate the model in (a) using $\widehat{ADDON}$ in place of *ADDON*. Use heteroskedasticity robust standard errors. Does it appear that having add-on insurance is a significant factor affecting the number of doctor visits?

d. Use your software command designed for two-stage least squares and estimate the model in (a) using external instruments *WHITEC* and *SELF*. Use heteroskedasticity robust standard errors. How do these estimates compare to those in part (c)? Has two-stage least squares performed as expected?

e. Since *ADDON* is binary, estimate the first stage equation in (b) using probit. Compute the estimated probability that *ADDON* = 1, *PHAT*. Reestimate the model in (a) using *PHAT* in place of *ADDON*. Use heteroskedasticity robust standard errors. Are the results the same as in part (d)? Why not?

f. Use your software command designed for two-stage least squares and estimate the model in (a) using external instrument *PHAT*. Use heteroskedasticity robust standard errors. How do these estimates compare to those in part (e)? Has two-stage least squares performed as expected?

16.30 In this exercise, we use multinomial logit to describe factors leading an individual to fall into one of three categories. Use data file *rwm88* for this exercise.

a. Create a variable called *INSURED* = 1, if a person does not have public insurance or add-on insurance (*PUBLIC* = 0 and *ADDON* = 0). Let *INSURED* = 2 if (*PUBLIC* = 1 and *ADDON* = 0). Let *INSURED* = 3 if (*PUBLIC* = 1 and *ADDON* = 1). Tabulate the number of individuals falling into each category. How many individuals are accounted for?

b. Estimate a multinomial logit model with outcome variable *INSURED* and explanatory variables *AGE*, *FEMALE*, *WORKING*, and *HHNINC2*. Use *INSURED* = 1 as the base category. What information is provided by the signs and significance of the estimated coefficients?

c. Obtain the predicted probabilities of falling into each category for each person in the sample, calling them *P1*, *P2*, and *P3*. Find the sample averages of *P1*, *P2*, and *P3* and compare these to the percentages of the sample for whom *INSURED* = 1, 2, and 3, respectfully.

d. Obtain the predicted probabilities of falling into each category for a person who is 50 years old, female, working and with a household income, *HHNINC2* = 2400.

e. Repeat the calculations in (d) for *HHNINC2* = 4200.

f. Calculate the 25th and 75th percentiles of *HHNINC2*. Comment on the changes in probabilities computed in parts (d) and (e).

Appendix 16A Probit Marginal Effects: Details

16A.1 Standard Error of Marginal Effect at a Given Point

Consider the probit model $p = \Phi(\beta_1 + \beta_2 x)$. The marginal effect of a continuous x, evaluated at a specific point $x = x_0$, is

$$\left.\frac{dp}{dx}\right|_{x=x_0} = \phi(\beta_1 + \beta_2 x_0)\beta_2 = g(\beta_1, \beta_2)$$

The estimator of the marginal effect is $g(\tilde{\beta}_1, \tilde{\beta}_2)$, where $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are the maximum likelihood estimators of the unknown parameters. The variance of this estimator was developed in Appendix 5B.2, in (5B.4), and is given by

$$\begin{aligned}\operatorname{var}\left[g(\tilde{\beta}_1, \tilde{\beta}_2)\right] \cong & \left[\frac{\partial g(\beta_1, \beta_2)}{\partial \beta_1}\right]^2 \operatorname{var}(\tilde{\beta}_1) + \left[\frac{\partial g(\beta_1, \beta_2)}{\partial \beta_2}\right]^2 \operatorname{var}(\tilde{\beta}_2) \\ & + 2\left[\frac{\partial g(\beta_1, \beta_2)}{\partial \beta_1}\right]\left[\frac{\partial g(\beta_1, \beta_2)}{\partial \beta_2}\right] \operatorname{cov}(\tilde{\beta}_1, \tilde{\beta}_2) \qquad (16A.1)\end{aligned}$$

The variances and covariances of the estimators come from maximum likelihood estimation. The essence of these calculations is given in Appendix C.8.2. To implement the delta method, we require the derivative

$$\begin{aligned}\frac{\partial g(\beta_1, \beta_2)}{\partial \beta_1} &= \frac{\partial\left[\phi(\beta_1 + \beta_2 x_0)\beta_2\right]}{\partial \beta_1} \\ &= \left\{\frac{\partial \phi(\beta_1 + \beta_2 x_0)}{\partial \beta_1} \times \beta_2\right\} + \phi(\beta_1 + \beta_2 x_0) \times \frac{\partial \beta_2}{\partial \beta_1} \\ &= -\phi(\beta_1 + \beta_2 x_0) \times (\beta_1 + \beta_2 x_0) \times \beta_2\end{aligned}$$

The second line above uses the product rule, Derivative Rule 6. To obtain the final result, we used $\partial\beta_2/\partial\beta_1 = 0$ and

$$\begin{aligned}\frac{\partial\phi(\beta_1+\beta_2x_0)}{\partial\beta_1} &= \frac{\partial}{\partial\beta_1}\left[\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\beta_1+\beta_2x_0)^2}\right]\\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\beta_1+\beta_2x_0)^2}\left(2\times-\frac{1}{2}\times(\beta_1+\beta_2x_0)\right)\\ &= -\phi(\beta_1+\beta_2x_0)\times(\beta_1+\beta_2x_0)\end{aligned}$$

The second step uses Derivative Rule 7 for exponential functions. Using similar steps, we obtain the other key derivative,

$$\frac{\partial g(\beta_1,\beta_2)}{\partial\beta_2} = \phi(\beta_1+\beta_2x_0)\left[1-(\beta_1+\beta_2x_0)\times\beta_2x_0\right]$$

From the maximum likelihood estimation results using the transportation data example, we obtain the estimator variances and covariances[32]

$$\begin{bmatrix}\widehat{\text{var}}(\tilde\beta_1) & \widehat{\text{cov}}(\tilde\beta_1,\tilde\beta_2)\\ \widehat{\text{cov}}(\tilde\beta_1,\tilde\beta_2) & \widehat{\text{var}}(\tilde\beta_2)\end{bmatrix} = \begin{bmatrix}0.1593956 & 0.0003261\\ 0.0003261 & 0.0105817\end{bmatrix}$$

The derivatives must be evaluated at the maximum likelihood estimates. For the transportation data used in Examples 16.4 and 16.5 for $DTIME = 2$ ($x_0 = 2$), the calculated values of the derivatives are

$$\widehat{\frac{\partial g(\beta_1,\beta_2)}{\partial\beta_1}} = -0.055531 \quad \text{and} \quad \widehat{\frac{\partial g(\beta_1,\beta_2)}{\partial\beta_2}} = 0.2345835$$

Using (16A.1), and carrying out the required multiplication, we obtain the estimated variance and standard error of the marginal effect

$$\widehat{\text{var}}\left[g(\tilde\beta_1,\tilde\beta_2)\right] = 0.0010653 \quad \text{and} \quad \text{se}\left[g(\tilde\beta_1,\tilde\beta_2)\right] = 0.0326394$$

16A.2 Standard Error of Average Marginal Effect

Consider the probit model $p = \Phi(\beta_1+\beta_2x)$. For the transportation data example, the explanatory variable $x = DTIME$. The average marginal effect of this continuous variable is

$$AME = \frac{1}{N}\sum_{i=1}^{N}\phi(\beta_1+\beta_2DTIME_i)\beta_2 = g_2(\beta_1,\beta_2)$$

The estimator of the average marginal effect is $g_2(\tilde\beta_1,\tilde\beta_2)$. To apply the delta method to find $\text{var}\left[g_2(\tilde\beta_1,\tilde\beta_2)\right]$, we require the derivatives

$$\begin{aligned}\frac{\partial g_2(\beta_1,\beta_2)}{\partial\beta_1} &= \frac{\partial}{\partial\beta_1}\left[\frac{1}{N}\sum_{i=1}^{N}\phi(\beta_1+\beta_2DTIME_i)\beta_2\right]\\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\beta_1}\left[\phi(\beta_1+\beta_2DTIME_i)\beta_2\right]\\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(\beta_1,\beta_2)}{\partial\beta_1}\end{aligned}$$

[32] Using minus the inverse matrix of second derivatives.

The term $\dfrac{\partial g(\beta_1, \beta_2)}{\partial \beta_1}$ we evaluated in the previous section. Similarly, the derivative

$$\begin{aligned}\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_2} &= \frac{\partial}{\partial \beta_2}\left[\frac{1}{N}\sum_{i=1}^{N}\phi(\beta_1 + \beta_2 DTIME_i)\beta_2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial \beta_2}\left[\phi(\beta_1 + \beta_2 DTIME_i)\beta_2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial g(\beta_1, \beta_2)}{\partial \beta_2}\end{aligned}$$

For the transportation data, we compute

$$\widehat{\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_1}} = -0.00185 \quad \text{and} \quad \widehat{\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_2}} = -0.032366$$

Using (16A.1) with g replaced by g_2, and carrying out the required multiplication, we obtain the estimated variance and standard error of the average marginal effect

$$\widehat{\text{var}}\left[g_2(\tilde{\beta}_1, \tilde{\beta}_2)\right] = 0.0000117 \quad \text{and} \quad \text{se}\left[g_2(\tilde{\beta}_1, \tilde{\beta}_2)\right] = 0.003416$$

Appendix 16B Random Utility Models

Economics is a general theory of choice behavior. Individuals make choices that maximize their wellbeing, or welfare, or, as economists term it, "utility." Observers cannot measure utility directly, and we cannot compare the utility, or satisfaction, that Jane enjoys while eating ice cream to Bill's satisfaction. But when a person is confronted with two or more choices, we assume that they make the choice that maximizes their welfare, however that might be defined. If a person must choose between taking a bus to work or driving to work, then, after considering the various costs and benefits, the person's choice reveals their utility maximizing outcome. We can imagine that the utility they receive depends on the attributes of the alternatives. As modelers we can select some such attributes as explanatory variables, but we must recognize that we will never truly understand choices completely; there is a random unexplained component, or random error, in any model.

Choice models, both binary and multinomial, as well as other limited dependent variable models, are often developed using a random utility model framework. Utility, or satisfaction, is unobservable and consequently it is called a **latent variable**, one that must be present but which is unseen. We will illustrate this approach to modeling by developing the probit model of binary choice in the random utility framework.

16B.1 Binary Choice Model

Assume that an individual must choose between two alternatives. Let U_{i1} be the utility derived from alternative one and let U_{i0} be the utility derived from alternative two. Let z_{i1} be the attributes of alternative one as perceived by the ith individual, and let z_{i0} be the attributes of alternative two as perceived by the ith individual. Let w_i represent the attributes of the ith individual. There may be several attributes of the alternatives that are relevant, and several individual characteristics that matter as well, but for simplicity, we will assume that there is but one attribute of each alternative and one individual characteristic. Then, a linear random utility model for each alternative is

$$\begin{aligned}U_{i1} &= \alpha_1 + z_{i1}\delta + w_i\gamma_1 + e_{i1} \\ U_{i0} &= \alpha_0 + z_{i0}\delta + w_i\gamma_0 + e_{i0}\end{aligned} \tag{16B.1}$$

In each model, there is a random error component, e_{i1} and e_{i0}. Assuming strict exogeneity, $E(e_{i1}|z_{i1}, z_{i0}, w_i) = 0$ and the same for e_{i0}, we can write

$$U_{i1} = E(U_{i1}|\cdot) + e_{i1} \quad \text{and} \quad U_{i0} = E(U_{i0}|\cdot) + e_{i0}$$

so that the utility from each part consists of a systematic part and a random part, as we are used to. Each of the expected utility terms is conditional, but we suppress the notation for convenience. Also, note that the individual characteristics w_i have coefficients that are unique to each alternative but that the attributes of alternatives, z_{i1} and z_{i0}, have a common parameter, δ. The logic of this specification will become clear soon.

As in equation (16.1), let the outcome variable be

$$y_i = \begin{cases} 1 & \text{if alternative one is chosen} \\ 0 & \text{if alternative two is chosen} \end{cases} \tag{16B.2}$$

Based on our model of random utility, alternative one will be chosen, and $y_i = 1$, if $U_{i1} \geq U_{i0}$, or if $U_{i1} - U_{i0} \geq 0$, where

$$\begin{aligned} U_{i1} - U_{i0} &= E(U_{i1}|\cdot) + e_{i1} - \left[E(U_{i0}|\cdot) + e_{i0}\right] \\ &= (\alpha_1 - \alpha_0) + (z_{i1} - z_{i0})\delta + w_i(\gamma_1 - \gamma_0) + (e_{i1} - e_{i0}) \end{aligned} \tag{16B.3}$$

The left-hand side variable $U_{i1} - U_{i0}$ is unobservable, but we know the difference in utilities determines an individual's choice. Let $y_i^* = U_{i1} - U_{i0}$ denote the latent variable which is the difference in utilities. Observe what would happen if the characteristics of the individual had the same coefficient in the random utility models (16B.1). Then the individual characteristics would fall out of (16B.3) and would have no effect on the choice. Equation (16B.3) becomes a regression specification by writing it as

$$\begin{aligned} y_i^* &= (\alpha_1 - \alpha_0) + (z_{i1} - z_{i0})\delta + w_i(\gamma_1 - \gamma_0) + (e_{i1} - e_{i0}) \\ &= \beta_1 + \beta_2(z_{i1} - z_{i0}) + \beta_3 w_i + e_i \\ &= \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i \end{aligned} \tag{16B.4}$$

We observe $y_i = 1$ if $y_i^* = U_{i1} - U_{i0} \geq 0$. The probability of an individual choosing alternative one is

$$\begin{aligned} p(\mathbf{x}_i) &= P(y_i = 1|\cdot) = P(y_i^* \geq 0|\cdot) = P\left[(U_{i1} \geq U_{i0})|\cdot\right] \\ &= P\left[E(U_{i1}|\cdot) + e_{i1} \geq E(U_{i0}|\cdot) + e_{i0}\right] \\ &= P\left[e_{i0} - e_{i1} \leq E(U_{i1}|\cdot) - E(U_{i0}|\cdot)\right] \\ &= P\left[e_{i0} - e_{i1} \leq \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}\right] \\ &= F(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}) \end{aligned} \tag{16B.5}$$

In the last line of (16B.5), $F(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3})$ is the cumulative distribution function of the random variable $e_{i0} - e_{i1}$. In Section 16.2, we used the *cdf* as a convenient device for keeping the probabilities between zero and one, but here it arises quite naturally from the random utility framework.

16B.2 Probit or Logit?

In binary choice problems, economists tend to use probit over logit. The reason follows from assumptions about the random utility models. Suppose that $e_{i1} \sim N(0, \sigma_1^2)$, $e_{i0} \sim N(0, \sigma_0^2)$, and $\text{cov}(e_{i1}, e_{i0}) = \sigma_{10}$. Then $(e_{i0} - e_{i1}) \sim N(0, \sigma^2 = \sigma_0^2 + \sigma_1^2 - 2\sigma_{10})$. Then

$$
\begin{aligned}
p(\mathbf{x}_i) &= P(y_i = 1|\cdot) \\
&= P\left[e_{i0} - e_{i1} \le \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}\right] \\
&= P\left[\frac{e_{i0} - e_{i1}}{\sigma} \le \frac{\beta_1}{\sigma} + \frac{\beta_2}{\sigma} x_{i2} + \frac{\beta_3}{\sigma} x_{i3}\right] \\
&= \Phi\left(\beta_1^* + \beta_2^* x_{i2} + \beta_3^* x_{i3}\right)
\end{aligned}
$$

The parameters in the probit model are actually $\beta_k^* = \beta_k/\sigma$. The parameter scaling is usually ignored in notation with the explanation that we choose $\sigma = 1$ as a normalization.[33] Then the probit model is $p(\mathbf{x}_i) = \Phi\left(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}\right)$.

On the other hand, to obtain a logit model, the random errors e_{i1} and e_{i0} must be statistically independent and identically distributed with an *extreme value distribution*.[34] In this case, $\left(e_{i0} - e_{i1}\right) = v_1$ has a logistic distribution. The details are a fun exercise (see Example B.7 for part of it) and outlined in Dhrymes (1986, page 1574).[35]

The bottom line is that there is no reason to assume that the random utility errors are statistically independent, nor to have the asymmetrical extreme value distribution. It is a mathematically convenient assumption because the end result, the logistic distribution, has a *cdf* of convenient form. Assuming that the random utility errors are normally distributed, and correlated, is not at all a stretch of the imagination.

Appendix 16C Using Latent Variables

Using latent variables, we can develop a variety of models that involve observed and partially observed variables. We will illustrate a few using simple models. Others can be found in Amemiya (1984, "Tobit Models: A Survey," *Journal of Econometrics*, 24, pages 3–61).

16C.1 Tobit (Tobit Type I)

Amemiya called the standard Tobit model "Type I Tobit." Let $y_i^* = \beta_1 + \beta_2 x_i + e_i$ be a latent variable with $e_i \sim N\left(0, \sigma^2\right)$. The Tobit model then arises by specifying the observed outcome value y_i to be,

$$
y_i = \begin{cases} y_i^* = \beta_1 + \beta_2 x_i + e_i & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \le 0 \end{cases}
$$

Three possible regression functions are then

$$
E\left(y_i^*|x_i\right) = \beta_1 + \beta_2 x_i
$$

$$
E\left(y_i|x_i, y_i > 0\right) = \beta_1 + \beta_2 x_i + \frac{\phi\left[\left(\beta_1 + \beta_2 x_i\right)/\sigma\right]}{\Phi\left[\left(\beta_1 + \beta_2 x_i\right)/\sigma\right]}
$$

$$
E\left(y_i|x_i\right) = \Phi\left[\left(\beta_1 + \beta_2 x_i\right)/\sigma\right] E\left(y_i|x_i, y_i > 0\right)
$$

[33] The issue of this normalization comes into play in the discussion of Heckman's two-step estimator, discussed in Section 16.7.5.

[34] https://en.wikipedia.org/wiki/Gumbel_distribution

[35] http://www.sciencedirect.com/science/handbooks/15734412/3

The marginal effects for a continuous variable x_i are

$$\partial E(y_i^*|x_i)/\partial x_i = \beta_2$$
$$\partial E(y_i|x_i, y_i > 0)/\partial x_i = \left\{1 - \alpha_i\lambda(\alpha_i) - [\lambda(\alpha_i)]^2\right\}\beta_2$$
$$\partial E(y_i|x_i)/\partial x_i = \Phi(\alpha_i)\beta_2$$

where $\alpha_i = (\beta_1 + \beta_2 x_i)/\sigma$ and $\lambda(\alpha_i) = \phi(\alpha_i)/\Phi(\alpha_i)$.

16C.2 Heckit (Tobit Type II)

The famous model of self-selection (Tobit Type II) developed by James Heckman is called "Heckit." In this model, there are two equations. The selection equation, that describes a person's participation decision, and an intensity, or amount, equation, which is the equation of interest. In the latent variable formulation, the equations are

$$z_i^* = \gamma_1 + \gamma_2 w_i + u_i \quad \text{selection equation}$$
$$y_i^* = \beta_1 + \beta_2 x_i + e_i \quad \text{amount equation, the equation of interest}$$

The equations are connected through their error terms. Let $u_i \sim N(0, \sigma_u^2)$ and $e_i \sim N(0, \sigma_e^2)$, with the covariance between these two random errors being σ_{ue}. The latent variables z_i^* and y_i^* are not observed. We do observe the binary variable

$$z_i = \begin{cases} 1 & z_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$y_i = \begin{cases} y_i^* = \beta_1 + \beta_2 x_i + e_i & \text{if } z_i^* > 0 \text{ or } z_i = 1 \\ 0 & \text{if } z_i^* \le 0 \text{ or } z_i = 0 \end{cases}$$

Using a theorem about bivariate normal random variables, similar to Appendix B.3.5, it can be shown that

$$E(y_i|x_i, w_i, y_i > 0) = \beta_1 + \beta_2 x_i + \sigma_{ue}\frac{\phi\left[(\gamma_1 + \gamma_2 w_i)/\sigma_u\right]}{\Phi\left[(\gamma_1 + \gamma_2 w_i)/\sigma_u\right]} = \beta_1 + \beta_2 x_i + \sigma_{ue}\frac{\phi(\gamma_1^* + \gamma_2^* w_i)}{\Phi(\gamma_1^* + \gamma_2^* w_i)}$$

Heckman's two-step estimator first estimates the selection model's scaled parameters $\gamma_1^* = \gamma_1/\sigma_u$ and $\gamma_2^* = \gamma_2/\sigma_u$ by probit using all observations. Then, using *only positive* observations, estimates by OLS the equation of interest

$$y_i = \beta_1 + \beta_2 x_i + \sigma_{ue}\frac{\phi(\tilde{\gamma}_1^* + \tilde{\gamma}_2^* w_i)}{\Phi(\tilde{\gamma}_1^* + \tilde{\gamma}_2^* w_i)} + v_i$$

The two-step estimator is consistent and asymptotically normally distributed, but the usual OLS standard errors are incorrect. The corrected ones are complicated but available in econometric software. An alternative is to estimate by maximum likelihood the two equations jointly, which is a more efficient estimation option. The MLE is often the default in econometric software, so check your documentation.

Appendix 16D A Tobit Monte Carlo Experiment

Let the latent variable be

$$y_i^* = \beta_1 + \beta_2 x_i + e_i = -9 + x_i + e_i \tag{16D.1}$$

with the error term assumed to have a normal distribution, $e_i \sim N(0, \sigma^2 = 16)$. The observable outcome y_i takes the value zero if $y_i^* \leq 0$, but $y_i = y_i^*$ if $y_i^* > 0$. In the simulation, we

- Create $N = 200$ random values of x_i that are spread evenly (or uniformly) over the interval [0, 20].
- Obtain $N = 200$ random values e_i from a normal distribution with mean 0 and variance 16.
- Create $N = 200$ values of the latent variable $y_i^* = -9 + x_i + e_i$.
- Obtain $N = 200$ values of the observed y_i using

$$y_i = \begin{cases} 0 & \text{if } y_i^* \leq 0 \\ y_i^* & \text{if } y_i^* > 0 \end{cases}$$

The 200 observations obtained this way constitute a sample that is **censored** with a lower limit of zero. The latent data are plotted in Figure 16D.1. In this figure, the line labeled $E(y_i^*|x_i)$ has intercept -9 and slope 1. The values of the latent variable y_i^* (triangle and hollow circle, △ and ○) are scattered along this regression function; if we observed these data we could estimate the parameters using the least squares principle, by fitting a line through the center of the data.

However, we do not observe all the latent data. When the values of y_i^* are zero or less then we observe $y_i = 0$ (●). We observe the y_i^* when they are positive. These observable data, along with the fitted least squares regression, are shown in Figure 16D.2.

The least squares principle will fail to estimate $\beta_1 = -9$ and $\beta_2 = 1$ because the observed data do not fall along the underlying regression function $E(y^*|x) = \beta_1 + \beta_2 x = -9 + x$.

To illustrate, the results from the first Monte Carlo sample, data file *tobit5*, are contained in Table 16D.1. In the first column (y^*) are the OLS estimates using the simulated latent data. In the second column $(y > 0)$ are the OLS estimates using only the 118 observations for which the

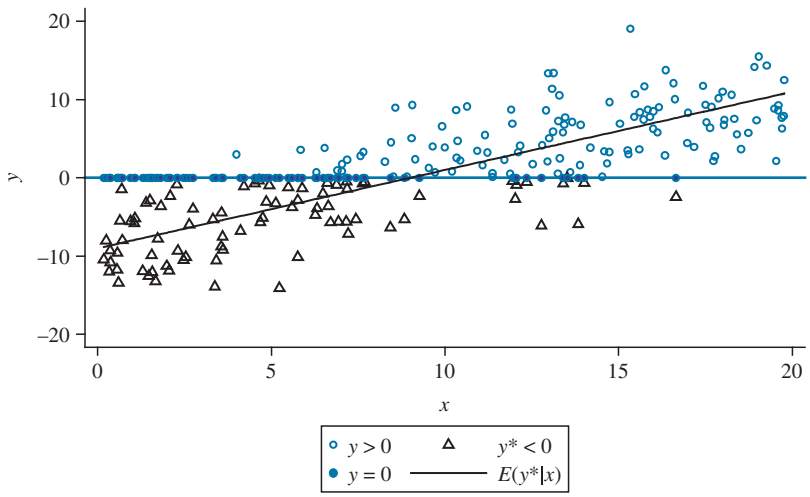

FIGURE 16D.1 **Latent and censored data.**

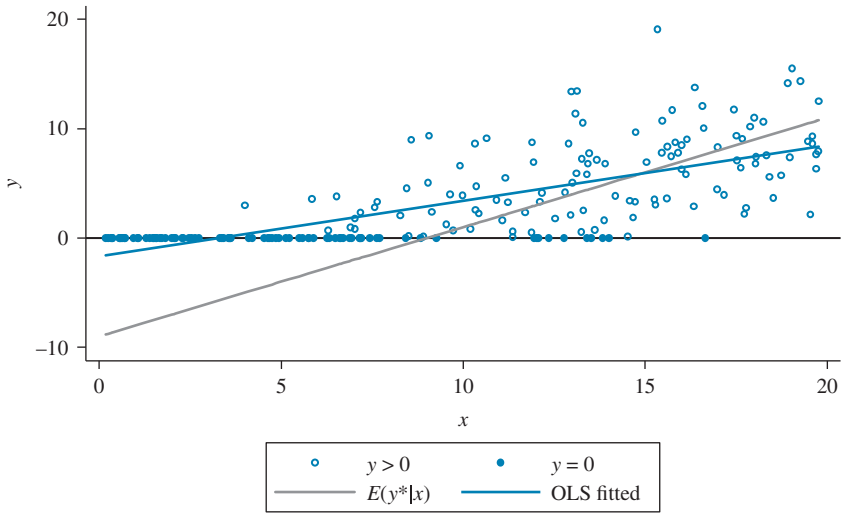

FIGURE 16D.2 **Observed data and OLS fitted line.**

TABLE 16D.1 **Simulated Censored Data (*tobit5*)**

	y^*	$y > 0$	y	Tobit
C	−8.6611	−1.1891	−1.6515	−8.0007
	(0.5842)	(1.1777)	(0.4290)	(0.9802)
x	0.9690	0.5176	0.5075	0.9215
	(0.0499)	(0.0823)	(0.0366)	(0.0722)
$\hat{\sigma}$	4.1050	3.4340	3.0146	3.9884
				(0.2670)
N	200	118	200	200

(Standard errors in parentheses)

observed value of y is positive; in the third column (y), are the OLS estimates on the 200 observed values of y, and in the fourth column are the Tobit estimates. The Tobit estimates are relatively close to the true value, while the estimates based only on the positive y values, or on all the y values, are far from the mark. An added benefit of the ML method is that there is a standard error for the estimated value of σ.

In the Monte Carlo simulation, we repeat this process of creating $N = 200$ observations, and applying least squares estimation, many times. This is analogous to "repeated sampling" in the context of experimental statistics. In this case, we repeat the process $NSAM = 1000$ times, drawing new x-values and error values e, recording each time the values of the estimates we obtain. At the end, we can compute the average values of the estimates which is the Monte Carlo "expected value,"

$$E_{MC}(b_k) = \frac{1}{NSAM}\sum_{m=1}^{NSAM} b_{k(m)}$$

where $b_{k(m)}$ is the estimate of β_k in the mth Monte Carlo sample. We also compute the Monte Carlo average of the usual, or "nominal" standard error, and the standard deviation of the estimates. The standard deviation measures the true sampling variability of the estimates. It is our hope that the usual standard error captures the actual sampling variation so that the average nominal standard error and the standard deviation of the estimates are close. The results are in Table 16D.2.

TABLE 16D.2 **Monte Carlo Simulation Results**

	Intercept = −9			Slope = 1		
	Mean	Standard Error	Standard Deviation	Mean	Standard Error	Standard Deviation
y^*	−9.0021	0.5759	0.5685	1.0000	0.0498	0.0492
$y > 0$	−2.1706	0.9518	1.1241	0.6087	0.0729	0.0779
y	−2.2113	0.2928	0.4185	0.5632	0.0389	0.0362
Tobit	−9.0571	1.0116	0.9994	1.0039	0.0740	0.0733

The results of applying OLS to the latent data (y^*) produce estimates that are on average very close to the true values for both the intercept and the slope. The average of the nominal standard error is close to the standard deviation of the estimates. If we discard the $y = 0$ observations and apply least squares to just the positive y observations, $y > 0$, these averages are −2.1706 and 0.6087, respectively. If we apply the least squares estimation procedure to all the observed censored data (y, including observations $y = 0$), the average value of the estimated intercept is −2.2113, and the average value of the estimated slope is 0.5632. The least squares estimates are biased by a substantial amount, compared to the true values $\beta_1 = -9$ and $\beta_2 = 1$. This bias will not disappear no matter how large the sample size we consider because the least squares estimators are inconsistent when data are censored or truncated. On the other hand, the Tobit estimates on average are very close to the true values.

A word of caution is in order about commercial software packages. There are many algorithms available for obtaining maximum likelihood estimates, and different packages use different ones, which may lead to slight differences (in perhaps the 3rd or 4th decimal) in the parameter estimates and their standard errors. When carrying out important research, it is a good tip to confirm empirical results with a second software package, just to be sure that they give essentially the same numbers.