17
DISCRETE CHOICE

17.1 INTRODUCTION

This is the first of three chapters that will survey models used in **microeconometrics**. The analysis of individual choice that is the focus of this field is fundamentally about modeling discrete outcomes such as purchase decisions, for example whether or not to buy insurance, voting behavior, choice among a set of alternative brands, travel modes or places to live, and responses to survey questions about the strength of preferences or about self-assessed health or well-being. In these and any number of other cases, the "dependent variable" is not a quantitative measure of some economic outcome, but rather an indicator of whether or not some outcome occurred. It follows that the regression methods we have used up to this point are largely inappropriate. We turn, instead, to modeling probabilities and using econometric tools to make probabilistic statements about the occurrence of these events. We will also examine models for counts of occurrences. These are closer to familiar regression models, but are, once again, about discrete outcomes of behavioral choices. As such, in this setting as well, we will be modeling probabilities of events, rather than conditional mean functions.

The models that are analyzed in this and the next chapter are built on a platform of preferences of decision makers. We take a **random utility** view of the choices that are observed. The decision maker is faced with a situation or set of alternatives and reveals something about their underlying preferences by the choice that he or she makes. The choice(s) made will be affected by observable influences—this is, of course, the ultimate objective of advertising—and by unobservable characteristics of the chooser. The blend of these fundamental bases for individual choice is at the core of the broad range of models that we will examine here.[1]

This chapter and Chapter 18 will describe four broad frameworks for analysis:

Binary Choice: The individual faces a pair of choices and makes that choice between the two that provides the greater utility. Many such settings involve the choice between taking an action and not taking that action, for example the decision whether or not to purchase health insurance. In other cases, the decision might be between two distinctly different choices, such as the decision whether to travel to and from work via public or private transportation. In the binary choice case, the 0/1 outcome is merely a label for "no/yes"—the numerical values are a mere convenience.

Multinomial Choice: The individual chooses among more than two choices, once again, making the choice that provides the greatest utility. In the previous example, private travel might involve a choice of being a driver or passenger while public

[1] See Greene and Hensher (2010, Chapter 4) for an historical perspective on this approach to model specification.

transport might involve a choice between bus and train. At one level, this is a minor variation of the binary choice case—the latter is, of course, a special case of the former. But, more elaborate models of multinomial choice allow a rich specification of consumer preferences. In the multinomial case, the observed response is simply a label for the selected choice; it might be a brand, the name of a place, or the type of travel mode. Numerical assignments are not meaningful in this setting.

Ordered Choice: The individual reveals the strength of his or her preferences with respect to a single outcome. Familiar cases involve survey questions about strength of feelings about a particular commodity such as a movie, or self-assessments of social outcomes such as health in general or self-assessed well-being. In the ordered choice setting, opinions are given meaningful numeric values, usually $0, 1, \ldots, J$ for some upper limit, J. For example, opinions might be labelled 0, 1, 2, 3, 4 to indicate the strength of preferences, for example, for a product, a movie, a candidate or a piece of legislation. But, in this context, the numerical values are only a ranking, not a quantitative measure. Thus a "1" is greater than a "0" in a qualitative sense, but not by one unit, and the difference between a "2" and a "1" is not the same as that between a "1" and a "0."

In these three cases, although the numerical outcomes are merely labels of some nonquantitative outcome, the analysis will nonetheless have a regresson-style motivation. Throughout, the models will be based on the idea that observed "covariates" are relevant in explaining the observed choices. For example, in the binary outcome "did or did not purchase health insurance," a conditioning model suggests that covariates such as age, income, and family situation will help to explain the choice. This chapter will describe a range of models that have been developed around these considerations. We will also be interested in a fourth application of discrete outcome models:

Event Counts: The observed outcome is a count of the number of occurrences. In many cases, this is similar to the preceding three settings in that the "dependent variable" measures an individual choice, such as the number of visits to the physician or the hospital, the number of derogatory reports in one's credit history, or the number of visits to a particular recreation site. In other cases, the event count might be the outcome of some natural process, such as incidence of a disease in a population or the number of defects per unit of time in a production process. In this setting, we will be doing a more familiar sort of regression modeling. However, the models will still be constructed specifically to accommodate the discrete nature of the observed response variable.

We will consider these four cases in turn. The four broad areas have many elements in common; however, there are also substantive differences between the particular models and analysis techniques used in each. This chapter will develop the first topic, models for binary choices. In each section, we will begin with an overview of applications and then present the single basic model that is the centerpiece of the methodology, and, finally, examine some recently developed extensions of the model. This chapter contains a very lengthy discussion of models for binary choices. This analysis is as long as it is because, first, the models discussed are used throughout microeconometrics—the central model of binary choice in this area is as ubiquitous as linear regression. Second, all the econometric issues and features that are encountered in the other areas will appear in the analysis of binary choice, where we can examine them in a fairly straightforward fashion.

It will emerge that, at least in econometric terms, the models for multinomial and ordered choice considered in Chapter 18 can be built from the two fundamental building blocks, the model of random utility and the translation of that model into a description of binary choices. There are relatively few new econometric issues that arise here. Chapter 18 will be largely devoted to suggesting different approaches to modeling choices among multiple alternatives and models for ordered choices. Once again, models of preference scales, such as movie or product ratings, or self-assessments of health or well-being, can be naturally built up from the fundamental model of random utility. Finally, Chapter 18 will develop the well-known Poisson regression model for counts of events. We will then extend the model to demonstrate some recent applications and innovations.

Chapters 17 and 18 are a lengthy but far from complete survey of topics in estimating **qualitative response (QR)** models. None of these models can consistently be estimated with linear regression methods. In most cases, the method of estimation is **maximum likelihood.** Therefore, readers interested in the mechanics of estimation may want to review the material in Appendices D and E before continuing. The various properties of maximum likelihood estimators are discussed in Chapter 14. We shall assume throughout these chapters that the necessary conditions behind the optimality properties of maximum likelihood estimators are met and, therefore, we will not derive or establish these properties specifically for the QR models. Detailed proofs for most of these models can be found in surveys by Amemiya (1981), McFadden (1984), Maddala (1983), and Dhrymes (1984). Additional commentary on some of the issues of interest in the contemporary literature is given by Manski and McFadden (1981) and Maddala and Flores-Lagunes (2001). Agresti (2002) and Cameron and Trivedi (2005) contain numerous theoretical developments and applications. Greene (2008) and Hensher and Greene (2010) provide, among many others, general surveys of discrete choice models and methods.[2]

17.2 MODELS FOR BINARY OUTCOMES

For purposes of studying individual behavior, we will construct models that link the decision or outcome to a set of factors, at least in the spirit of regression. Our approach will be to analyze each of them in the general framework of probability models:

$$\text{Prob(event } j \text{ occurs)} = \text{Prob}(Y = j) = F[\text{relevant effects, parameters}]. \quad \textbf{(17-1)}$$

The study of qualitative choice focuses on appropriate specification, estimation, and use of models for the probabilities of events, where in most cases, the "event" is an individual's choice among a set of two or more alternatives.

Example 17.1 Labor Force Participation Model

In Example 5.2 we estimated an earnings equation for the subsample of 428 married women who participated in the formal labor market taken from a full sample of 753 observations. The semilog earnings equation is of the form

$$\ln \textit{earnings} = \beta_1 + \beta_2\, \textit{age} + \beta_3\, \textit{age}^2 + \beta_4\, \textit{education} + \beta_5\, \textit{kids} + \varepsilon,$$

[2]There are dozens of book length surveys of discrete choice models. Two others that are heavily oriented to application of the methods are Train (2003) and Hensher, Rose, and Greene (2005).

where *earnings* is *hourly wage* times *hours worked, education* is measured in years of schooling, and *kids* is a binary variable which equals one if there are children under 18 in the household. What of the other 325 individuals? The underlying labor supply model described a market in which labor force participation was the outcome of a market process whereby the demanders of labor services were willing to offer a wage based on expected marginal product and individuals themselves made a decision whether or not to accept the offer depending on whether it exceeded their own reservation wage. The first of these depends on, among other things, education, while the second (we assume) depends on such variables as age, the presence of children in the household, other sources of income (husband's), and marginal tax rates on labor income. The sample we used to fit the earnings equation contains data on all these other variables. The models considered in this chapter would be appropriate for modeling the outcome $y = 1$ if in the labor force, and 0 if not.

Models for explaining a binary (0/1) dependent variable are typically motivated in two contexts. The labor force participation model in Example 17.1 describes a process of individual choice between two alternatives in which the choice is influenced by observable effects (children, tax rates) and unobservable aspects of the preferences of the individual. The relationship between voting behavior and income is another example. In other cases, the **binary choice model** arises in a setting in which the nature of the observed data dictate the special treatment of a binary dependent variable model. In these cases, the analyst is essentially interested in a regression-like model of the sort considered in Chapters 2 through 7. With data on the variable of interest and a set of covariates, they are interested in specifying a relationship between the former and the latter, more or less along the lines of the models we have already studied. For example, in a model of the demand for tickets for sporting events, in which the variable of interest is number of tickets, it could happen that the observation consists only of whether the sports facility was filled to capacity (demand greater than or equal to capacity so $Y = 1$) or not ($Y = 0$). It will generally turn out that the models and techniques used in both cases are the same. Nonetheless, it is useful to examine both of them.

17.2.1 RANDOM UTILITY MODELS FOR INDIVIDUAL CHOICE

An interpretation of data on individual choices is provided by the random utility model. Let U_a and U_b represent an individual's utility of two choices. For example, U_a might be the utility of rental housing and U_b that of home ownership. The observed choice between the two reveals which one provides the greater utility, but not the unobservable utilities. Hence, the observed indicator equals 1 if $U_a > U_b$ and 0 if $U_a \leq U_b$. A common formulation is the linear random utility model,

$$U_a = \mathbf{w}'\boldsymbol{\beta}_a + \mathbf{z}_a'\boldsymbol{\gamma}_a + \varepsilon_a \quad \text{and} \quad U_b = \mathbf{w}'\boldsymbol{\beta}_b + \mathbf{z}_b'\boldsymbol{\gamma}_b + \varepsilon_b. \tag{17-2}$$

In (17-2), the observable (measurable) vector of **characteristics** of the individual is denoted **w**; this might include gender, age, income, and other demographics. The vectors $\mathbf{z}_a$ and $\mathbf{z}_b$ denote features (**attributes**) of the two choices that might be choice specific. In a voting context, for example, the attributes might be indicators of the competing candidates' positions on important issues. The random terms, ε_a and ε_b represent the stochastic elements that are specific to and known only by the individual, but not by the observer (analyst). To continue our voting example, ε_a might represent an intangible, general "preference" for candidate a.

The completion of the model for the determination of the observed outcome (choice) is the revelation of the ranking of the preferences by the choice the individual makes. Thus, if we denote by $Y = 1$ the consumer's choice of alternative a, we infer from $Y = 1$ that $U_a > U_b$. Since the outcome is ultimately driven by the random elements in the utility functions, we have

$$\begin{aligned}\text{Prob}[Y = 1|\mathbf{w}, \mathbf{z}_a, \mathbf{z}_b] &= \text{Prob}[U_a > U_b] \\ &= \text{Prob}[(\mathbf{w}'\boldsymbol{\beta}_a + \mathbf{z}_a'\boldsymbol{\gamma}_a + \varepsilon_a) - (\mathbf{x}'\boldsymbol{\beta}_b + \mathbf{z}_b'\boldsymbol{\gamma}_b + \varepsilon_b) > 0 \,|\, \mathbf{w}, \mathbf{z}_a, \mathbf{z}_b] \\ &= \text{Prob}[(\mathbf{w}'(\boldsymbol{\beta}_a - \boldsymbol{\beta}_b) + \mathbf{z}_a'\boldsymbol{\gamma}_a - \mathbf{z}_b'\boldsymbol{\gamma}_b + \varepsilon_a - \varepsilon_b) > 0 \,|\, \mathbf{w}, \mathbf{z}_a, \mathbf{z}_b] \\ &= \text{Prob}[\mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0 \,|\, \mathbf{x}],\end{aligned}$$

where $\mathbf{x}'\boldsymbol{\beta}$ collects all the observable elements of the difference of the two utility functions and ε denotes the difference between the two random elements.

Example 17.2 Structural Equations for a Binary Choice Model

Nakosteen and Zimmer (1980) analyzed a model of migration based on the following structure:[3] For a given individual, the market wage that can be earned at the present location is

$$y_p^* = \mathbf{w}_p'\boldsymbol{\beta}_p + \varepsilon_p.$$

Variables in the equation include age, sex, race, growth in employment, and growth in per capita income. If the individual migrates to a new location, then his or her market wage would be

$$y_m^* = \mathbf{w}_m'\boldsymbol{\beta}_m + \varepsilon_m.$$

Migration entails costs that are related both to the individual and to the labor market:

$$C^* = \mathbf{z}'\boldsymbol{\alpha} + u.$$

Costs of moving are related to whether the individual is self-employed and whether that person recently changed his or her industry of employment. They migrate if the benefit $y_m^* - y_p^*$ is greater than the cost, C. The net benefit of moving is

$$\begin{aligned}M^* &= y_m^* - y_p^* - C^* \\ &= \mathbf{w}_m'\boldsymbol{\beta}_m - \mathbf{w}_p'\boldsymbol{\beta}_p - \mathbf{z}'\boldsymbol{\alpha} + (\varepsilon_m - \varepsilon_p - u) \\ &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon.\end{aligned}$$

Because M^* is unobservable, we cannot treat this equation as an ordinary regression. The individual either moves or does not. After the fact, we observe only y_m^* if the individual has moved or y_p^* if he or she has not. But we do observe that $M = 1$ for a move and $M = 0$ for no move.

[3] A number of other studies have also used variants of this basic formulation. Some important examples are Willis and Rosen (1979) and Robinson and Tomes (1982). The study by Tunali (1986) examined in Example 17.6 is another application. The now standard approach, in which "participation" equals one if wage offer $(\mathbf{x}_w'\boldsymbol{\beta}_w + \varepsilon_w)$ minus reservation wage $(\mathbf{x}_r'\boldsymbol{\beta}_r + \varepsilon_r)$ is positive, is also used in Fernandez and Rodriguez-Poo (1997). Brock and Durlauf (2000) describe a number of models and situations involving individual behavior that give rise to binary choice models.

17.2.2 A LATENT REGRESSION MODEL

Discrete dependent-variable models are often cast in the form of **index function models**. We view the outcome of a discrete choice as a reflection of an underlying regression. As an often-cited example, consider the decision to make a large purchase. The theory states that the consumer makes a marginal benefit/marginal cost calculation based on the utilities achieved by making the purchase and by not making the purchase and by using the money for something else. We model the difference between benefit and cost as an unobserved variable y^* such that

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

Note that this is the result of the "net utility" calculation in the previous section and in Example 17.2. We assume that ε has mean zero and has either a standardized logistic with variance $\pi^2/3$ or a standard normal distribution with variance one or some other specific distribution with known variance. We do not observe the net benefit of the purchase (i.e., net utility), only whether it is made or not. Therefore, our observation is

$$\begin{aligned} y &= 1 \quad \text{if } y^* > 0, \\ y &= 0 \quad \text{if } y^* \leq 0. \end{aligned} \tag{17-3}$$

In this formulation, $\mathbf{x}'\boldsymbol{\beta}$ is called the index function. The assumption of known variance of ε is an innocent normalization. Suppose the variance of ε is scaled by an unrestricted parameter σ^2. The **latent regression** will be $y^* = \mathbf{x}'\boldsymbol{\beta} + \sigma\varepsilon$. But, $(y^*/\sigma) = \mathbf{x}'(\boldsymbol{\beta}/\sigma) + \varepsilon$ is the same model with the same data. The observed data will be unchanged; y is still 0 or 1, depending only on the sign of y^* not on its scale. This means that there is no information about σ in the sample data so σ cannot be estimated. The parameter vector $\boldsymbol{\beta}$ in this model is only "identified up to scale." The assumption of zero for the threshold in (17-3) is likewise innocent if the model contains a constant term (and not if it does not).[4] Let a be the supposed nonzero threshold and α be the unknown constant term and, for the present, $\mathbf{x}$ and $\boldsymbol{\beta}$ contain the rest of the index not including the constant term. Then, the probability that y equals one is

$$\text{Prob}(y^* > a \,|\, \mathbf{x}) = \text{Prob}(\alpha + \mathbf{x}'\boldsymbol{\beta} + \varepsilon > a \,|\, \mathbf{x}) = \text{Prob}[(\alpha - a) + \mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0 \,|\, \mathbf{x}].$$

Because α is unknown, the difference $(\alpha - a)$ remains an unknown parameter. The end result is that if the model contains a constant term, it is unchanged by the choice of the threshold in (17-3). The choice of zero is a normalization with no significance. With the two normalizations, then,

$$\text{Prob}(y^* > 0 \,|\, \mathbf{x}) = \text{Prob}(\varepsilon > -\mathbf{x}'\boldsymbol{\beta} \,|\, \mathbf{x}).$$

A remaining detail in the model is the choice of the specific distribution for ε. We will consider several. The overwhelming majority of applications are based either on the normal or the logistic distribution. If the distribution is symmetric, as are the normal and logistic, then

$$\text{Prob}(y^* > 0 \,|\, \mathbf{x}) = \text{Prob}(\varepsilon < \mathbf{x}'\boldsymbol{\beta} \,|\, \mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta}), \tag{17-4}$$

[4]Unless there is some compelling reason, binomial probability models should not be estimated without constant terms.

where $F(t)$ is the cdf of the random variable, ε. This provides an underlying structural model for the probability.

17.2.3 FUNCTIONAL FORM AND REGRESSION

Consider the model of labor force participation suggested in Example 17.1. The respondent either works or seeks work ($Y = 1$) or does not ($Y = 0$) in the period in which our survey is taken. We believe that a set of factors, such as age, marital status, education, and work history, gathered in a vector $\mathbf{x}$, explain the decision, so that

$$\begin{aligned}\text{Prob}(Y = 1 \mid \mathbf{x}) &= F(\mathbf{x}, \boldsymbol{\beta}) \\ \text{Prob}(Y = 0 \mid \mathbf{x}) &= 1 - F(\mathbf{x}, \boldsymbol{\beta}).\end{aligned} \tag{17-5}$$

The set of parameters $\boldsymbol{\beta}$ reflects the impact of changes in $\mathbf{x}$ on the probability. For example, among the factors that might interest us is the marginal effect of marital status on the probability of labor force participation. The problem at this point is to devise a suitable model for the right-hand side of the equation. One possibility is to retain the familiar linear regression,

$$F(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}.$$

Because $E[y \mid \mathbf{x}] = 0[1-F(\mathbf{x}, \boldsymbol{\beta})]+1[F(\mathbf{x}, \boldsymbol{\beta})] = F(\mathbf{x}, \boldsymbol{\beta})$, we can construct the regression model,

$$\begin{aligned}y &= E[y \mid \mathbf{x}] + y - E[y \mid \mathbf{x}] \\ &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon.\end{aligned} \tag{17-6}$$

The **linear probability model** has a number of shortcomings. A minor complication arises because ε is heteroscedastic in a way that depends on $\boldsymbol{\beta}$. Because $\mathbf{x}'\boldsymbol{\beta} + \varepsilon$ must equal 0 or 1, ε equals either $-\mathbf{x}'\boldsymbol{\beta}$ or $1-\mathbf{x}'\boldsymbol{\beta}$, with probabilities $1 - F$ and F, respectively. Thus, you can easily show that in this model,

$$\text{Var}[\varepsilon \mid \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}(1 - \mathbf{x}'\boldsymbol{\beta}). \tag{17-7}$$

We could manage this complication with an FGLS estimator in the fashion of Chapter 9, though this only solves the estimation problem, not the theoretical one. A more serious flaw is that without some ad hoc tinkering with the disturbances, we cannot be assured that the predictions from this model will truly look like probabilities. We cannot constrain $\mathbf{x}'\boldsymbol{\beta}$ to the 0–1 interval. Such a model produces both nonsense probabilities and negative variances. For these reasons, the linear probability model is becoming less frequently used except as a basis for comparison to some other more appropriate models.[5]

[5]The linear model is not beyond redemption. Aldrich and Nelson (1984) analyze the properties of the model at length. Judge et al. (1985) and Fomby, Hill, and Johnson (1984) give interesting discussions of the ways we may modify the model to force internal consistency. But the fixes are sample dependent, and the resulting estimator, such as it is, may have no known sampling properties. Additional discussion of weighted least squares appears in Amemiya (1977) and Mullahy (1990). Finally, its shortcomings notwithstanding, the linear probability model is applied by Caudill (1988), Heckman, and MaCurdy (1985), and Heckman and Snyder (1997). An exchange on the usefulness of the approach is Angrist (2001) and Moffitt (2001). See Angrist and Pischke (2009) for some applications.

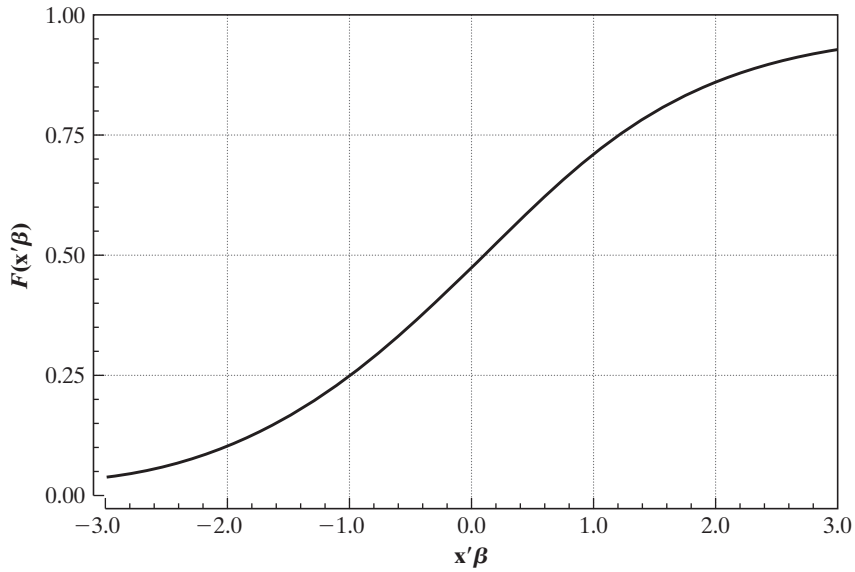

FIGURE 17.1 Model for a Probability.

Our requirement, then, is a model that will produce predictions consistent with the underlying theory in (17-4). For a given regressor vector, we would expect

$$\lim_{\mathbf{x}'\boldsymbol{\beta}\to+\infty} \text{Prob}(Y = 1 \mid \mathbf{x}) = 1$$
$$\lim_{\mathbf{x}'\boldsymbol{\beta}\to-\infty} \text{Prob}(Y = 1 \mid \mathbf{x}) = 0. \tag{17-8}$$

See Figure 17.1. In principle, any proper, continuous probability distribution defined over the real line will suffice. The normal distribution has been used in many analyses, giving rise to the **probit** model,

$$\text{Prob}(Y = 1 \mid \mathbf{x}) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \phi(t)dt = \Phi(\mathbf{x}'\boldsymbol{\beta}). \tag{17-9}$$

The function $\Phi(t)$ is a commonly used notation for the standard normal distribution function. Partly because of its mathematical convenience, the logistic distribution,

$$\text{Prob}(Y = 1 \mid \mathbf{x}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})} = \Lambda(\mathbf{x}'\boldsymbol{\beta}). \tag{17-10}$$

has also been used in many applications. We shall use the notation $\Lambda(.)$ to indicate the logistic cumulative distribution function. This model is called the **logit** model for reasons we shall discuss in the next section. Both of these distributions have the familiar bell shape of symmetric distributions. Other models which do not assume symmetry, such as the **Gumbel model**,

$$\text{Prob}(Y = 1 \mid \mathbf{x}) = \exp[-\exp(-\mathbf{x}'\boldsymbol{\beta})],$$

and **complementary log log model**,

$$\text{Prob}(Y = 1 \mid \mathbf{x}) = 1 - \exp[-\exp(\mathbf{x}'\boldsymbol{\beta})],$$

have also been employed. Still other distributions have been suggested,[6] but the probit and logit models are still the most common frameworks used in econometric applications.

The question of which distribution to use is a natural one. The logistic distribution is similar to the normal except in the tails, which are considerably heavier. (It more closely resembles a t distribution with seven degrees of freedom.) Therefore, for intermediate values of $\mathbf{x}'\boldsymbol{\beta}$ (say, between -1.2 and $+1.2$), the two distributions tend to give similar probabilities. The logistic distribution tends to give larger probabilities to $Y = 1$ when $\mathbf{x}'\boldsymbol{\beta}$ is extremely small (and smaller probabilities to $Y = 1$ when $\mathbf{x}'\boldsymbol{\beta}$ is very large) than the normal distribution. It is difficult to provide practical generalities on this basis, however, as they would require knowledge of $\boldsymbol{\beta}$. We should expect different predictions from the two models, however, if the sample contains (1) very few "responses" (Y's equal to 1) or very few "nonresponses" (Y's equal to 0) and (2) very wide variation in an important independent variable, particularly if (1) is also true. There are practical reasons for favoring one or the other in some cases for mathematical convenience, but it is difficult to justify the choice of one distribution or another on theoretical grounds. Amemiya (1981) discusses a number of related issues, but as a general proposition, the question is unresolved. In most applications, the choice between these two seems not to make much difference. However, as seen in the following example, the symmetric and asymmetric distributions can give substantively different results, and here, the guidance on how to choose is unfortunately sparse.

The probability model is a regression:

$$E[y \mid \mathbf{x}] = F(\mathbf{x}'\boldsymbol{\beta}).$$

Whatever distribution is used, it is important to note that the parameters of the model, like those of any nonlinear regression model, are not necessarily the marginal effects we are accustomed to analyzing. In general,

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = \left[\frac{dF(\mathbf{x}'\boldsymbol{\beta})}{d(\mathbf{x}'\boldsymbol{\beta})}\right] \times \boldsymbol{\beta} = f(\mathbf{x}'\boldsymbol{\beta}) \times \boldsymbol{\beta}, \tag{17-11}$$

where $f(.)$ is the density function that corresponds to the cumulative distribution, $F(.)$. For the normal distribution, this result is

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = \phi(\mathbf{x}'\boldsymbol{\beta}) \times \boldsymbol{\beta}, \tag{17-12}$$

where $\phi(t)$ is the standard normal density. For the logistic distribution,

$$\frac{d\Lambda(\mathbf{x}'\boldsymbol{\beta})}{d(\mathbf{x}'\boldsymbol{\beta})} = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{[1 + \exp(\mathbf{x}'\boldsymbol{\beta})]^2} = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})],$$

so, in the logit model,

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})]\boldsymbol{\beta}. \tag{17-13}$$

[6] See, for example, Maddala (1983, pp. 27–32), Aldrich and Nelson (1984), and Greene (2001).

It is obvious that these values will vary with the values of **x**. In interpreting the estimated model, it will be useful to calculate this value at, say, the means of the regressors and, where necessary, other pertinent values. For convenience, it is worth noting that the same scale factor applies to all the slopes in the model.

For computing **marginal effects**, one can evaluate the expressions at the sample means of the data or evaluate the marginal effects at every observation and use the sample average of the individual marginal effects—this produces the **average partial effects**. In large samples these generally give roughly the same answer (see Section 17.3.2). But that is not so in small- or moderate-sized samples. Current practice favors averaging the individual marginal effects when it is possible to do so.

Another complication for computing marginal effects in a binary choice model arises because **x** will often include dummy variables—for example, a labor force participation equation will often contain a dummy variable for marital status. Because the derivative is with respect to a small change, it is not appropriate to apply (15) for the effect of a change in a dummy variable, or a change of state. The appropriate marginal effect for a binary independent variable, say, d, would be

$$\text{Marginal effect} = \text{Prob}[Y = 1 \mid \bar{\mathbf{x}}_{(d)}, d = 1] - \text{Prob}[Y = 1 \mid \bar{\mathbf{x}}_{(d)}, d = 0], \qquad \textbf{(17-14)}$$

where $\bar{\mathbf{x}}_{(d)}$, denotes the means of all the other variables in the model. Simply taking the derivative with respect to the binary variable as if it were continuous provides an approximation that is often surprisingly accurate. In Example 17.3, for the binary variable *PSI*, the difference in the two probabilities for the probit model is $(0.5702 - 0.1057) = 0.4645$, whereas the derivative approximation reported in Table 17.1 is 0.468. Nonetheless, it might be optimistic to rely on this outcome. We will revisit this computation in the examples and discussion to follow.

17.3 ESTIMATION AND INFERENCE IN BINARY CHOICE MODELS

With the exception of the linear probability model, estimation of binary choice models is usually based on the method of maximum likelihood. Each observation is treated as a single draw from a Bernoulli distribution (binomial with one draw). The model with success probability $F(\mathbf{x}'\boldsymbol{\beta})$ and independent observations leads to the joint probability, or likelihood function,

$$\text{Prob}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \mathbf{X}) = \prod_{y_i=0} [1 - F(\mathbf{x}_i'\boldsymbol{\beta})] \prod_{y_i=1} F(\mathbf{x}_i'\boldsymbol{\beta}).$$

where $\mathbf{X}$ denotes $[\mathbf{x}_i]_{i=1,\ldots,n}$. The likelihood function for a sample of n observations can be conveniently written as

$$L(\boldsymbol{\beta} \mid \text{data}) = \prod_{i=1}^{n} [F(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i} [1 - F(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i}. \qquad \textbf{(17-15)}$$

Taking logs, we obtain

$$\ln L = \sum_{i=1}^{n} \left\{ y_i \ln F(\mathbf{x}_i'\boldsymbol{\beta}) + (1 - y_i) \ln[1 - F(\mathbf{x}_i'\boldsymbol{\beta})] \right\}.^{7} \tag{17-16}$$

The **likelihood equations** are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\frac{y_i f_i}{F_i} + (1 - y_i) \frac{-f_i}{(1 - F_i)} \right] \mathbf{x}_i = \mathbf{0}, \tag{17-17}$$

where f_i is the density, $dF_i/d(\mathbf{x}_i'\boldsymbol{\beta})$. [In (17-17) and later, we will use the subscript i to indicate that the function has an argument $\mathbf{x}_i'\boldsymbol{\beta}$.] The choice of a particular form for F_i leads to the empirical model.

Unless we are using the linear probability model, the likelihood equations in (17-17) will be nonlinear and require an iterative solution. All of the models we have seen thus far are relatively straightforward to analyze. For the logit model, by inserting (17-7) and (17-11) in (17-17), we get, after a bit of manipulation, the likelihood equations

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (y_i - \Lambda_i)\mathbf{x}_i = \mathbf{0}. \tag{17-18}$$

Note that if $\mathbf{x}_i$ contains a constant term, the first-order conditions imply that the average of the predicted probabilities must equal the proportion of ones in the sample.[8] This implication also bears some similarity to the least squares normal equations if we view the term $y_i - \Lambda_i$ as a residual.[9] For the normal distribution, the log-likelihood is

$$\ln L = \sum_{y_i=0} \ln[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})] + \sum_{y_i=1} \ln \boldsymbol{\Phi}(\mathbf{x}_i'\boldsymbol{\beta}). \tag{17-19}$$

The first-order conditions for maximizing $\ln L$ are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{y_i=0} \frac{-\phi_i}{1 - \boldsymbol{\Phi}_i}\mathbf{x}_i + \sum_{y_i=1} \frac{\phi_i}{\Phi_i}\mathbf{x}_i = \sum_{y_i=0} \lambda_{0i}\mathbf{x}_i + \sum_{y_i=1} \lambda_{1i}\mathbf{x}_i.$$

Using the device suggested in footnote 7, we can reduce this to

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\frac{q_i \phi(q_i \mathbf{x}_i'\boldsymbol{\beta})}{\boldsymbol{\Phi}(q_i \mathbf{x}_i'\boldsymbol{\beta})} \right] \mathbf{x}_i = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i = \mathbf{0}, \tag{17-20}$$

where $q_i = 2y_i - 1$.

The actual second derivatives for the logit model are quite simple:

$$\mathbf{H} = \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_{i} \Lambda_i (1 - \Lambda_i) \mathbf{x}_i \mathbf{x}_i'. \tag{17-21}$$

[7] If the distribution is symmetric, as the normal and logistic are, then $1 - F(\mathbf{x}'\boldsymbol{\beta}) = F(-\mathbf{x}'\boldsymbol{\beta})$. There is a further simplification. Let $q = 2y - 1$. Then $\ln L = \Sigma_i \ln F(q_i \mathbf{x}_i'\boldsymbol{\beta})$. See (17-21).

[8] The same result holds for the linear probability model. Although regularly observed in practice, the result has not been verified for the probit model.

[9] This sort of construction arises in many models. The first derivative of the log-likelihood with respect to the constant term produces the **generalized residual** in many settings. See, for example, Chesher, Lancaster, and Irish (1985) and the equivalent result for the tobit model in Section 19.3.4.d.

The second derivatives do not involve the random variable y_i, so Newton's method is also the **method of scoring** for the logit model. Note that the Hessian is always negative definite, so the log-likelihood is globally concave. Newton's method will usually converge to the maximum of the log-likelihood in just a few iterations unless the data are especially badly conditioned. The computation is slightly more involved for the probit model. A useful simplification is obtained by using the variable $\lambda(y_i, \boldsymbol{\beta}'\mathbf{x}_i) = \lambda_i$ that is defined in (17-20). The second derivatives can be obtained using the result that for any z, $d\phi(z)/dz = -z\phi(z)$. Then, for the probit model,

$$\mathbf{H} = \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} -\lambda_i(\lambda_i + \mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i'. \tag{17-22}$$

This matrix is also negative definite for all values of $\boldsymbol{\beta}$. The proof is less obvious than for the logit model.[10] It suffices to note that the scalar part in the summation is $\text{Var}[\varepsilon \mid \varepsilon \leq \boldsymbol{\beta}'\mathbf{x}] - 1$ when $y = 1$ and $\text{Var}[\varepsilon \mid \varepsilon \geq -\boldsymbol{\beta}'\mathbf{x}] - 1$ when $y = 0$. The unconditional variance is one. Because truncation always reduces variance—see Theorem 18.2—in both cases, the variance is between zero and one, so the value is negative.[11]

The asymptotic covariance matrix for the maximum likelihood estimator can be estimated by using the inverse of the Hessian evaluated at the maximum likelihood estimates. There are also two other estimators available. The Berndt, Hall, Hall, and Hausman estimator [see (14-18) and Example 14.4] would be

$$\mathbf{B} = \sum_{i=1}^{n} g_i^2 \mathbf{x}_i \mathbf{x}_i',$$

where $g_i = (y_i - \boldsymbol{\Lambda}_i)$ for the logit model [see (17-18)] and $g_i = \lambda_i$ for the probit model [see (17-20)]. The third estimator would be based on the expected value of the Hessian. As we saw earlier, the Hessian for the logit model does not involve y_i, so $\mathbf{H} = E[\mathbf{H}]$. But because λ_i is a function of y_i [see (17-20)], this result is not true for the probit model. Amemiya (1981) showed that for the probit model,

$$E\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'}\right]_{\text{probit}} = \sum_{i=1}^{n} \lambda_{0i}\lambda_{1i}\mathbf{x}_i\mathbf{x}_i'. \tag{17-23}$$

Once again, the scalar part of the expression is always negative [see (17-20) and note that λ_{0i} is always negative and λ_{i1} is always positive]. The estimator of the asymptotic covariance matrix for the maximum likelihood estimator is then the negative inverse of whatever matrix is used to estimate the expected Hessian. Since the actual Hessian is generally used for the iterations, this option is the usual choice. As we shall see later, though, for certain hypothesis tests, the BHHH estimator is a more convenient choice.

17.3.1 ROBUST COVARIANCE MATRIX ESTIMATION

The probit maximum likelihood estimator is often labeled a **quasi-maximum likelihood estimator (QMLE)** in view of the possibility that the normal probability model might be misspecified. White's (1982a) robust "sandwich" estimator for the asymptotic

[10] See, for example, Amemiya (1985, pp. 273–274) and Maddala (1983, p. 63).

[11] See Johnson and Kotz (1993) and Heckman (1979). We will make repeated use of this result in Chapter 19.

covariance matrix of the QMLE (see Section 14.8 for discussion),

$$\text{Est. Asy. Var}[\hat{\boldsymbol{\beta}}] = [\hat{\mathbf{H}}]^{-1}\hat{\mathbf{B}}[\hat{\mathbf{H}}]^{-1},$$

has been used in a number of ~~recent~~ studies based on the probit model [e.g., Fernandez and Rodriguez-Poo (1997), Horowitz (1993), and Blundell, Laisney, and Lechner (1993)]. If the probit model is correctly specified, then $\text{plim}(1/n)\hat{\mathbf{B}} = \text{plim}(1/n)(-\hat{\mathbf{H}})$ and either single matrix will suffice, so the robustness issue is moot ~~(of course)~~. On the other hand, the probit (Q-) maximum likelihood estimator is *not* consistent in the presence of any form of heteroscedasticity, unmeasured heterogeneity, omitted variables (even if they are orthogonal to the included ones), nonlinearity of the functional form of the index, or an error in the distributional assumption [with some narrow exceptions as described by Ruud (1986)]. Thus, in almost any case, the sandwich estimator provides an appropriate asymptotic covariance matrix for an estimator that is biased in an unknown direction. [See Section 14.8 and Freedman (2006).] White raises this issue explicitly, although it seems to receive little attention in the literature: "It is the consistency of the QMLE for the parameters of interest in a wide range of situations which insures its usefulness as the basis for robust estimation techniques" (1982a, p. 4). His very useful result is that if the quasi-maximum likelihood estimator converges to a probability limit, then the sandwich estimator can, under certain circumstances, be used to estimate the asymptotic covariance matrix of that estimator. But there is no guarantee that the QMLE *will* converge to anything interesting or useful. Simply computing a robust covariance matrix for an otherwise inconsistent estimator does not give it redemption. Consequently, the virtue of a robust covariance matrix in this setting is unclear.

17.3.2 MARGINAL EFFECTS AND AVERAGE PARTIAL EFFECTS

The predicted probabilities, $F(\mathbf{x}'\hat{\boldsymbol{\beta}}) = \hat{F}$ and the estimated partial effects $f(\mathbf{x}'\hat{\boldsymbol{\beta}}) \times \hat{\boldsymbol{\beta}} = \hat{f}\hat{\boldsymbol{\beta}}$ are nonlinear functions of the parameter estimates. To compute standard errors, we can use the linear approximation approach (delta method) discussed in Section 4.4.4. For the predicted probabilities,

$$\text{Asy. Var}[\hat{F}] = [\partial\hat{F}/\partial\hat{\boldsymbol{\beta}}]'\mathbf{V}[\partial\hat{F}/\partial\hat{\boldsymbol{\beta}}],$$

where

$$\mathbf{V} = \text{Asy. Var}[\hat{\boldsymbol{\beta}}].$$

The estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ can be any of the three described earlier. Let $z = \mathbf{x}'\hat{\boldsymbol{\beta}}$. Then the derivative vector is

$$[\partial\hat{F}/\partial\hat{\boldsymbol{\beta}}] = [d\hat{F}/dz][\partial z/\partial\hat{\boldsymbol{\beta}}] = \hat{f}\mathbf{x}.$$

Combining terms gives

$$\text{Asy. Var}[\hat{F}] = \hat{f}^2\mathbf{x}'\,\mathbf{V}\mathbf{x},$$

which depends, of course, on the particular $\mathbf{x}$ vector used. This result is useful when a marginal effect is computed for a dummy variable. In that case, the estimated effect is

$$\Delta\hat{F} = \hat{F} \mid (d = 1) - \hat{F} \mid (d = 0). \qquad \textbf{(17-24)}$$

The asymptotic variance would be

$$\text{Asy. Var}[\Delta\hat{F}] = [\partial\Delta\hat{F}/\partial\hat{\boldsymbol{\beta}}]'\mathbf{V}[\partial\Delta\hat{F}/\partial\hat{\boldsymbol{\beta}}], \tag{17-25}$$

where

$$[\partial\Delta\hat{F}/\partial\hat{\boldsymbol{\beta}}] = \hat{f}_1\begin{pmatrix}\bar{\mathbf{x}}_{(d)}\\1\end{pmatrix} - \hat{f}_0\begin{pmatrix}\bar{\mathbf{x}}_{(d)}\\0\end{pmatrix}.$$

For the other marginal effects, let $\hat{\boldsymbol{\gamma}} = \hat{f}\hat{\boldsymbol{\beta}}$. Then

$$\text{Asy. Var}[\hat{\boldsymbol{\gamma}}] = \left[\frac{\partial\hat{\boldsymbol{\gamma}}}{\partial\hat{\boldsymbol{\beta}}'}\right]\mathbf{V}\left[\frac{\partial\hat{\boldsymbol{\gamma}}}{\partial\hat{\boldsymbol{\beta}}'}\right]'.$$

The matrix of derivatives is

$$\hat{f}\left(\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\hat{\boldsymbol{\beta}}'}\right) + \hat{\boldsymbol{\beta}}\left(\frac{d\hat{f}}{dz}\right)\left(\frac{\partial z}{\partial\hat{\boldsymbol{\beta}}'}\right) = \hat{f}\,\mathbf{I} + \left(\frac{d\hat{f}}{dz}\right)\hat{\boldsymbol{\beta}}\mathbf{x}'.$$

For the probit model, $df/dz = -z\phi$, so

$$\text{Asy. Var}[\hat{\boldsymbol{\gamma}}] = \phi^2[\mathbf{I} - (\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}\mathbf{x}']\mathbf{V}[\mathbf{I} - (\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}\mathbf{x}']'.$$

For the logit model, $\hat{f} = \hat{\Lambda}(1-\hat{\Lambda})$, so

$$\frac{d\hat{f}}{dz} = (1-2\hat{\Lambda})\left(\frac{d\hat{\Lambda}}{dz}\right) = (1-2\hat{\Lambda})\hat{\Lambda}(1-\hat{\Lambda}).$$

Collecting terms, we obtain

$$\text{Asy. Var}[\hat{\boldsymbol{\gamma}}] = [\Lambda(1-\Lambda)]^2[\mathbf{I} + (1-2\Lambda)\boldsymbol{\beta}\mathbf{x}']\mathbf{V}[\mathbf{I} + (1-2\Lambda)\mathbf{x}\boldsymbol{\beta}'].$$

As before, the value obtained will depend on the $\mathbf{x}$ vector used.

Example 17.3 Probability Models

The data listed in Appendix Table F14.1 were taken from a study by Spector and Mazzeo (1980), which examined whether a new method of teaching economics, the Personalized System of Instruction (*PSI*), significantly influenced performance in later economics courses. The "dependent variable" used in our application is *GRADE*, which indicates the whether a student's grade in an intermediate macroeconomics course was higher than that in the principles course. The other variables are *GPA*, their grade point average; *TUCE*, the score on a pretest that indicates entering knowledge of the material; and *PSI*, the binary variable indicator of whether the student was exposed to the new teaching method. (Spector and Mazzeo's specific equation was somewhat different from the one estimated here.)

Table 17.1 presents four sets of parameter estimates. The slope parameters and derivatives were computed for four probability models: linear, probit, logit, and complementary log log. The last three sets of estimates are computed by maximizing the appropriate log-likelihood function. Inference is discussed in the next section, so standard errors are not presented here. The scale factor given in the last row is the density function evaluated at the means of the variables. Also, note that the slope given for *PSI* is the derivative, not the change in the function with *PSI* changed from zero to one with other variables held constant.

If one looked only at the coefficient estimates, then it would be natural to conclude that the four models had produced radically different estimates. But a comparison of the columns of slopes shows that this conclusion is clearly wrong. The models are very similar; in fact, the logit and probit models results are nearly identical.

The data used in this example are only moderately unbalanced between 0s and 1s for the dependent variable (21 and 11). As such, we might expect similar results for the probit

TABLE 17.1 Estimated Probability Models

	Linear		*Logistic*		*Probit*		*Complementary log log*	
Variable	*Coefficient*	*Slope*	*Coefficient*	*Slope*	*Coefficient*	*Slope*	*Coefficient*	*Slope*
Constant	−1.498	—	−13.021	—	−7.452	—	−10.631	—
GPA	0.464	0.464	2.826	0.534	1.626	0.533	2.293	0.477
TUCE	0.010	0.010	0.095	0.018	0.052	0.017	0.041	0.009
PSI	0.379	0.379	2.379	0.450	1.426	0.468	1.562	0.325
$f(\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}})$	1.000		0.189		0.328		0.208	

and logit models.[12] One indicator is a comparison of the coefficients. In view of the different variances of the distributions, one for the normal and $\pi^2/3$ for the logistic, we might expect to obtain comparable estimates by multiplying the probit coefficients by $\pi/\sqrt{3} \approx 1.8$. Amemiya (1981) found, through trial and error, that scaling by 1.6 instead produced better results. This proportionality result is frequently cited. The result in (17-11) may help to explain the finding. The index $\mathbf{x}'\boldsymbol{\beta}$ is not the random variable. The marginal effect in the probit model for, say, x_k is $\phi(\mathbf{x}'\boldsymbol{\beta}_p)\beta_{pk}$, whereas that for the logit is $\Lambda(1-\Lambda)\beta_{lk}$. (The subscripts p and l are for probit and logit.) Amemiya suggests that his approximation works best at the center of the distribution, where $F = 0.5$, or $\mathbf{x}'\boldsymbol{\beta} = 0$ for either distribution. Suppose it is. Then $\phi(0) = 0.3989$ and $\Lambda(0)[1-\Lambda(0)] = 0.25$. If the marginal effects are to be the same, then $0.3989\,\beta_{pk} = 0.25\beta_{lk}$, or $\beta_{lk} = 1.6\beta_{pk}$, which is the regularity observed by Amemiya. Note, though, that as we depart from the center of the distribution, the relationship will move away from 1.6. Because the logistic density descends more slowly than the normal, for unbalanced samples such as ours, the ratio of the logit coefficients to the probit coefficients will tend to be larger than 1.6. The ratios for the ones in Table 17.1 are closer to 1.7 than 1.6.

The computation of the derivatives of the conditional mean function is useful when the variable in question is continuous and often produces a reasonable approximation for a dummy variable. Another way to analyze the effect of a dummy variable on the whole distribution is to compute Prob($Y = 1$) over the range of $\mathbf{x}'\boldsymbol{\beta}$ (using the sample estimates) and with the two values of the binary variable. Using the coefficients from the probit model in Table 17.1, we have the following probabilities as a function of *GPA*, at the mean of TUCE:

$$PSI = 0\text{: Prob}(GRADE = 1) = \Phi[-7.452 + 1.626GPA + 0.052(21.938)],$$

$$PSI = 1\text{: Prob}(GRADE = 1) = \Phi[-7.452 + 1.626GPA + 0.052(21.938) + 1.426].$$

Figure 17.2 shows these two functions plotted over the range of *GPA* observed in the sample, 2.0 to 4.0. The marginal effect of *PSI* is the difference between the two functions, which ranges from only about 0.06 at $GPA = 2$ to about 0.50 at *GPA* of 3.5. This effect shows that the probability that a student's grade will increase after exposure to *PSI* is far greater for students with high *GPAs* than for those with low *GPAs*. At the sample mean of *GPA* of 3.117, the effect of *PSI* on the probability is 0.465. The simple derivative calculation of (17-9) is given in Table 17.1; the estimate is 0.468. But, of course, this calculation does not show the wide range of differences displayed in Figure 17.2.

Table 17.2 presents the estimated coefficients and marginal effects for the probit and logit models in Table 17.2. In both cases, the asymptotic covariance matrix is computed from the negative inverse of the actual Hessian of the log-likelihood. The standard errors for the estimated marginal effect of *PSI* are computed using (17-24) and (17-25) since *PSI* is a binary variable. In comparison, the simple derivatives produce estimates and standard errors of (0.449, 0.181) for the logit model and (0.464, 0.188) for the probit model. These differ only slightly from the results given in the table.

[12] One might be tempted in this case to suggest an asymmetric distribution for the model, such as the Gumbel distribution. However, the asymmetry in the model, to the extent that it is present at all, refers to the values of ε, not to the observed sample of values of the dependent variable.

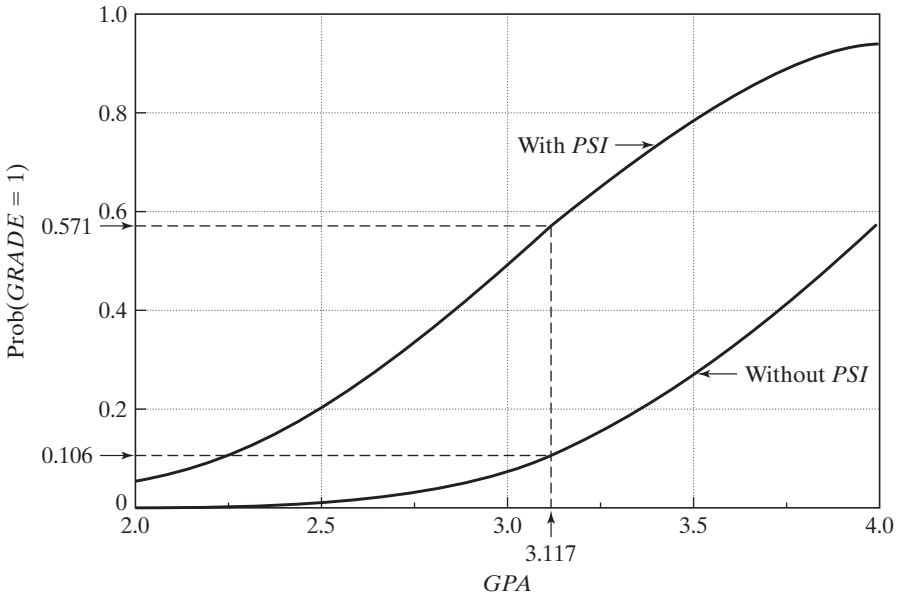

FIGURE 17.2 Effect of *PSI* on Predicted Probabilities.

17.3.2.a Average Partial Effects

The preceding has emphasized computing the partial effects for the average individual in the sample. Current practice has many applications based, instead, on "average partial effects." [See, e.g., Wooldridge (2002a).] The underlying logic is that the quantity of interest is

$$APE = E_x\left[\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}}\right].$$

In practical terms, this suggests the computation

$$\widehat{APE} = \bar{\hat{\gamma}} = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}.$$

TABLE 17.2 Estimated Coefficients and Standard Errors (standard errors in parentheses)

	Logistic				*Probit*			
Variable	***Coefficient***	***t Ratio***	***Slope***	***t Ratio***	***Coefficient***	***t Ratio***	***Slope***	***t Ratio***
Constant	−13.021	−2.641	—	—	−7.452	−2.931	—	—
	(4.931)				(2.542)			
GPA	2.826	2.238	0.534	2.252	1.626	2.343	0.533	2.294
	(1.263)		(0.237)		(0.694)		(0.232)	
TUCE	0.095	0.672	0.018	0.685	0.052	0.617	0.017	0.626
	(0.142)		(0.026)		(0.084)		(0.027)	
PSI	2.379	2.234	0.456	2.521	1.426	2.397	0.464	2.727
	(1.065)		(0.181)		(0.595)		(0.170)	
log-likelihood		−12.890				−12.819		

This does raise two questions. Because the computation is (marginally) more burdensome than the simple marginal effects at the means, one might wonder whether this produces a noticeably different answer. That will depend on the data. Save for small sample variation, the difference in these two results is likely to be small. Let

$$\bar{\gamma}_k = APE_k = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \Pr(y_i = 1 \mid \mathbf{x}_i)}{\partial x_{ik}} = \frac{1}{n}\sum_{i=1}^{n} F'(\mathbf{x}_i'\boldsymbol{\beta})\beta_k = \frac{1}{n}\sum_{i=1}^{n}\gamma_k(\mathbf{x}_i)$$

denote the computation of the average partial effect. We compute this at the MLE, $\hat{\boldsymbol{\beta}}$. Now, expand this function in a second-order Taylor series around the point of sample means, $\bar{\mathbf{x}}$, to obtain

$$\bar{\gamma}_k = \frac{1}{n}\sum_{i=1}^{n}\left[\gamma_k(\bar{\mathbf{x}}) + \sum_{m=1}^{k}\frac{\partial \gamma_k(\bar{\mathbf{x}})}{\partial \bar{x}_m}(x_{im} - \bar{x}_m) + \frac{1}{2}\sum_{l=1}^{K}\sum_{m=1}^{K}\frac{\partial^2 \gamma_k(\bar{\mathbf{x}})}{\partial \bar{x}_l \partial \bar{x}_m}(x_{il} - \bar{x}_l)(x_{im} - \bar{x}_m)\right] + \Delta,$$

where Δ is the remaining higher-order terms. The first of the three terms is the marginal effect computed at the sample means. The second term is zero by construction. That leaves the remainder plus an average of a term that is a function of the variances and covariances of the data and the curvature of the probability function at the means. Little can be said to characterize these two terms in any particular sample, but one might guess they are likely to be small. We will examine an application in Example 17.4.

Based on the sample of observations on the partial effects, a natural estimator of the variance of the partial effects would seem to be

$$\hat{\sigma}^2_{\gamma,k} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\gamma}_k(\mathbf{x}_i) - \bar{\hat{\gamma}}_k\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\widehat{PE}_{i,k} - \widehat{APE}_k\right)^2.$$

See, for example, Contoyannis et al. (2004, p. 498), who report that they computed the "sample standard deviation of the partial effects." Since $\widehat{APE}_k = \bar{\hat{\gamma}}_k$ is the mean of a sample, notwithstanding the following consideration, the preceding estimator should be further divided by the sample size since we are computing the *standard error of the mean of a sample*. This seems not to be the norm in the literature. This estimator should not be viewed as an alternative to the delta method applied to the partial effects evaluated at the means of the data, $\hat{\boldsymbol{\gamma}}(\bar{\mathbf{x}})$. The delta method produces an estimator of the asymptotic variance of an estimator of the population parameter, $\boldsymbol{\gamma}(\boldsymbol{\mu}_{\mathbf{x}})$, that is, of a function of $\hat{\boldsymbol{\beta}}$. The asymptotic covariance matrix computed using the delta method for $\hat{\boldsymbol{\gamma}}(\bar{\mathbf{x}})$ would be $\hat{\mathbf{G}}(\bar{\mathbf{x}})\hat{\mathbf{V}}\hat{\mathbf{G}}'(\bar{\mathbf{x}})$ where $\hat{\mathbf{G}}(\bar{\mathbf{x}})$ is the matrix of partial derivatives and $\hat{\mathbf{V}}$ is the estimator of the asymptotic variance of $\hat{\boldsymbol{\beta}}$. This variance estimator converges to zero because $\hat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}$ and $\bar{\mathbf{x}}$ converges to a vector of constants. The ~~naive~~ estimator above does not converge to zero; it converges to the variance of the random variable $PE_{i,k}$.

The "asymptotic variance" of the partial effects estimator is intended to reflect the variation of the parameter estimator, $\hat{\boldsymbol{\beta}}$, whereas the naive estimator generates the variation from the heterogeneity of the sample data while holding the parameter fixed

at $\hat{\boldsymbol{\beta}}$. For example, for a logit model,

$$\hat{\gamma}_k(\mathbf{x}_i) = \hat{\beta}_k \Lambda\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)\left[1 - \Lambda\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)\right] = \hat{\boldsymbol{\beta}}_k\hat{\delta}_i,$$

and $\hat{\delta}_i$ is the same for all k. It follows that

$$\hat{\sigma}^2_{\gamma,k} = \hat{\boldsymbol{\beta}}_k^2\left[\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\delta}_i - \bar{\hat{\delta}})^2\right] = \hat{\boldsymbol{\beta}}_k^2 s_{\hat{\delta}}^2.$$

A surprising consequence is that if one computes t ratios for the average partial effects using $\hat{\sigma}^2_{\gamma,k}$, the values will all equal the same $1/s_{\hat{\delta}}$. This might signal that something is amiss. (This is somewhat apparent in the Contoyannis et al. results on page 498; however, not enough digits were reported to see the effect clearly.)

A search for applications that use the delta method to estimate standard errors for average partial effects in nonlinear models yields hundreds of occurrences. However, we could not locate any that document in detail the precise formulas used. (One author, noting the complexity of computation, recommended bootstrapping instead.) A complicated flaw with the sample variance estimator (notwithstanding all the preceding) is that the ~~naive~~ estimator (whether scaled by $1/n$ or not) neglects the fact that all n observations used to compute the estimated APE are correlated; they all use the same estimator of $\boldsymbol{\beta}$. The preceding estimator treats the estimates of PE_i as if they were a random sample. They would be if they were based on the true $\boldsymbol{\beta}$. But the estimators based on the same $\hat{\boldsymbol{\beta}}$ are not uncorrelated. The delta method will account for the asymptotic (co)variation of the terms in the sum of functions of $\hat{\boldsymbol{\beta}}$. To use the delta method to estimate the asymptotic standard errors for the average partial effects, $\widehat{APE}_k$, we should use

$$\begin{aligned}
Est.\,Asy.\,Var\left[\bar{\hat{\boldsymbol{\gamma}}}\right] &= \frac{1}{n^2} Est.\,Asy.\,Var\left[\sum_{i=1}^{n}\hat{\boldsymbol{\gamma}}_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Est.\,Asy.\,Cov\left[\hat{\boldsymbol{\gamma}}_i, \hat{\boldsymbol{\gamma}}_j\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{G}_i(\hat{\boldsymbol{\beta}})\hat{\mathbf{V}}\mathbf{G}_j'(\hat{\boldsymbol{\beta}}) \\
&= \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{G}_i(\hat{\boldsymbol{\beta}})\right]\hat{\mathbf{V}}\left[\frac{1}{n}\sum_{j=1}^{n}\mathbf{G}_j'(\hat{\boldsymbol{\beta}})\right],
\end{aligned}$$

where

$$\mathbf{G}_i(\hat{\boldsymbol{\beta}}) = \frac{\partial f\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{\beta}}}{\partial\hat{\boldsymbol{\beta}}'} = f\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)\mathbf{I} + f'\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{\beta}}\mathbf{x}_i'.$$

This treats the APE as a point estimator of a population parameter—one that converges in probability to what we assume is its population counterpart. But, it is conditioned on the sample data; convergence is with respect to $\hat{\boldsymbol{\beta}}$. This looks like a formidable amount of computation—Example 17.4 uses a sample of 27,326 observations, so it appears we need a double sum of roughly 750 million terms. However, the computation is actually

TABLE 17.3 Estimated Parameters and Partial Effects

	Parameter Estimates		*Marginal Effects*		*Average Partial Effects*		
Variable	*Estimate*	*Std.Error*	*Estimate*	*Std.Error*	*Estimate*	*Std.Error*	*Naive S.E.*
Constant	0.25112	0.09114					
Age	0.02071	0.00129	0.00497	0.00031	0.00471	0.00029	0.00043
Income	−0.18592	0.07506	−0.04466	0.01803	−0.04229	0.01707	0.00386
Kids	−0.22947	0.02954	−0.05512	0.00710	−0.05220	0.00669	0.00476
Education	−0.04559	0.00565	−0.01095	0.00136	−0.01037	0.00128	0.00095
Married	0.08529	0.03329	0.02049	0.00800	0.01940	0.00757	0.00177

linear in n, not quadratic, because the same matrix is used in the center of each product. The estimator of the asymptotic covariance matrix for the APE is simply

$$Est.Asy.Var\left[\bar{\hat{\boldsymbol{\gamma}}}\right] = \overline{\mathbf{G}\left(\hat{\boldsymbol{\beta}}\right)}\hat{\mathbf{V}}\overline{\mathbf{G}'\left(\hat{\boldsymbol{\beta}}\right)}.$$

The appropriate covariance matrix is computed by making the same adjustment as in the partial effects—the derivative matrices are averaged over the observations rather than being computed at the means of the data.

Example 17.4 Average Partial Effects

We estimated a binary logit model for $y = 1(DocVis > 0)$ using the German health care utilization data examined in Example 7.6 (and several later examples). The model is

$$\text{Prob}(DocVis_{it} > 0) = \Lambda(\beta_1 + \beta_2 Age_{it} + \beta_3 Income_{it} + \beta_4 Kids_{it} + \beta_5 Education_{it} + \beta_6 Married_{it}).$$

No account of the panel nature of the data set was taken for this exercise. The sample contains 27,326 observations, which should be large enough to reveal the large sample behavior of the computations. Table 17.3 presents the parameter estimates for the logit probability model and both the marginal effects and the average partial effects, each with standard errors computed using the results given earlier. (The partial effects for the two dummy variables, *Kids* and *Married*, are computed using the approximation, rather than using the discrete differences.) The results do suggest the similarity of the computations. The values in parentheses in the last column are based on the naive estimator that ignores the covariances and is not divided by the $1/n$ for the variance of the mean.

17.3.2.b Interaction Effects

Models with **interaction effects**, such as

$$\text{Prob}(DocVis_{it} > 0) = \Lambda(\beta_1 + \beta_2 Age_{it} + \beta_3 Income_{it} + \beta_4 Kids_{it} + \beta_5 Education_{it} + \beta_6 Married_{it} + \beta_7 Age_{it} \times Education_{it}),$$

have attracted considerable attention in recent applications of binary choice models.[13] A practical issue concerns the computation of partial effects by standard computer packages. Write the model as

$$\text{Prob}(DocVis_{it} > 0) = \Lambda(\beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + \beta_7 x_{7it}).$$

Estimation of the model parameters is routine. Rote computation of partial effects using (17-11) will produce

$$PE_7 = \partial \text{Prob}(DocVis > 0)/\partial x_7 = \beta_7 \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})],$$

[13] See, for example, Ai and Norton (2004) and Greene (2010).

which is what common computer packages will dutifully report. The problem is that $x_7 = x_2x_5$, and PE_7 in the previous equation is not the partial effect for x_7. Moreover, the partial effects for x_2 and x_5 will also be misreported by the rote computation. To revert back to our original specification,

$$\partial \text{Prob}(DocVis > 0 \mid \mathbf{x})/\partial\, Age = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})](\beta_2 + \beta_7\, Education),$$

$$\partial \text{Prob}(DocVis > 0 \mid \mathbf{x})/\partial\, Education = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})](\beta_5 + \beta_7\, Age),$$

and what is computed as "$\partial \text{Prob}(DocVis > 0 \mid \mathbf{x})/\partial Age \times Education$" is meaningless. The practical problem motivating Ai and Norton (2004) was that the computer package does not know that x_7 is x_2x_5, so it computes a partial effect for x_7 as if it could vary "partially" from the other variables. The (now) obvious solution is for the analyst to force the correct computations of the relevant partial effects by whatever software they are using, perhaps by programming the computations themselves.

The practical complication raises a theoretical question that is less clear cut. What is the "interaction effect" in the model? In a linear model based on the preceding, we would have

$$\partial^2 E[y \mid \mathbf{x}]/\partial x_2 \partial x_5 = \beta_7$$

which is unambiguous. However, in this *nonlinear* binary choice model, the correct result is

$$\begin{aligned}\partial^2 E[y \mid \mathbf{x}]/\partial x_2 \partial x_5 = {} & \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})]\beta_7 \\ & + \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})][1 - 2\Lambda(\mathbf{x}'\boldsymbol{\beta})](\beta_2 + \beta_7\, Education)(\beta_5 + \beta_7\, Age).\end{aligned}$$

Not only is β_7 not the interesting effect, but there is also a complicated additional term. Loosely, we can associate the first term as a "direct" effect—note that it is the naive term PE_7 from earlier. The second part can be attributed to the fact that we are differentiating a nonlinear model—essentially, the second part of the partial effect results from the nonlinearity of the function. The existence of an "interaction effect" in this model is inescapable—notice that the second part is nonzero (generally) even if β_7 does equal zero. Whether this is intended to represent an "interaction" in some economic sense is unclear. In the absence of the product term in the model, probably not. We can see an implication of this in Figure 17.1. At the point where $\mathbf{x}'\boldsymbol{\beta} = 0$, where the probability equals one half, the probability function is linear. At that point, $(1 - 2\Lambda)$ will equal zero and the functional form effect will be zero as well. When $\mathbf{x}'\boldsymbol{\beta}$ departs from zero, the probability becomes nonlinear. (These same effects can be shown for the probit model—at $\mathbf{x}'\boldsymbol{\beta} = 0$, the second derivative of the probit probability is $-\mathbf{x}'\boldsymbol{\beta}\phi(\mathbf{x}'\boldsymbol{\beta}) = 0$.)

We developed an extensive application of interaction effects in a nonlinear model in Example 7.6. In that application, using the same data for the numerical exercise, we analyzed a nonlinear regression $E[y \mid \mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$. The results obtained in that study were general, and will apply to the application here, where the nonlinear regression is $E[y \mid \mathbf{x}] = \Lambda(\mathbf{x}'\boldsymbol{\beta})$ or $\Phi(\mathbf{x}'\boldsymbol{\beta})$.

Example 17.5 Interaction Effect

We added the interaction term, Age × Education, to the model in Example 17.4. The model is now

$$\begin{aligned}\text{Prob}(DocVis_{it} > 0) = {} & \Lambda(\beta_1 + \beta_2\, Age_{it} + \beta_3\, Income_{it} + \beta_4\, Kids_{it} \\ & + \beta_5\, Education_{it} + \beta_6\, Married_{it} + \beta_7\, Age_{it} \times Education_{it}).\end{aligned}$$

Estimation of the model produces an estimate of β_7 of -0.00112. The naive average partial effect for x_7 is -0.000254. This is the first part in the earlier decomposition. The second, functional form term (averaged over the sample observations) is 0.0000634, so the estimated interaction effect, the sum of the two terms is -0.000191. The naive calculation errs by about $(-0.000254/-0.000191-1)\times 100$ percent $=33$ percent.

17.3.3 MEASURING GOODNESS OF FIT

There have been many fit measures suggested for QR models.[14] At a minimum, one should report the maximized value of the log-likelihood function, ln L. Because the hypothesis that all the slopes in the model are zero is often interesting, the log-likelihood computed with only a constant term, ln L_0 [see (17-29)], should also be reported. An analog to the R^2 in a conventional regression is McFadden's (1974) likelihood ratio index,

$$\text{LRI} = 1 - \frac{\ln L}{\ln L_0}.$$

This measure has an intuitive appeal in that it is bounded by zero and one. (See Section 14.6.5.) If all the slope coefficients are zero, then it equals zero. There is no way to make LRI equal 1, although one can come close. If F_i is always one when y equals one and zero when y equals zero, then ln L equals zero (the log of one) and LRI equals one. It has been suggested that this finding is indicative of a "perfect fit" and that LRI increases as the fit of the model improves. To a degree, this point is true. Unfortunately, the values between zero and one have no natural interpretation. If $F(\mathbf{x}_i'\boldsymbol{\beta})$ is a proper pdf, then even with many regressors the model cannot fit perfectly unless $\mathbf{x}_i'\boldsymbol{\beta}$ goes to $+\infty$ or $-\infty$. As a practical matter, it does happen. But when it does, it indicates a flaw in the model, not a good fit. If the range of one of the independent variables contains a value, say, x^*, such that the sign of $(x - x^*)$ predicts y perfectly and vice versa, then the model will become a perfect predictor. This result also holds in general if the sign of $\mathbf{x}'\boldsymbol{\beta}$ gives a perfect predictor for some vector $\boldsymbol{\beta}$.[15] For example, one might mistakenly include as a regressor a dummy variable that is identical, or nearly so, to the dependent variable. In this case, the maximization procedure will break down precisely because $\mathbf{x}'\boldsymbol{\beta}$ is diverging during the iterations. [See McKenzie (1998) for an application and discussion.] Of course, this situation is not at all what we had in mind for a good fit.

Other fit measures have been suggested. Ben-Akiva and Lerman (1985) and Kay and Little (1986) suggested a fit measure that is keyed to the prediction rule,

$$R^2_{\text{BL}} = \frac{1}{n}\sum_{i=1}^{n}\left[y_i\hat{F}_i + (1-y_i)(1-\hat{F}_i)\right],$$

which is the average probability of correct prediction by the prediction rule. The difficulty in this computation is that in unbalanced samples, the less frequent outcome will usually be predicted very badly by the standard procedure, and this measure does not pick up that point. Cramer (1999) has suggested an alternative measure that directly

[14] See, for example, Cragg and Uhler (1970), Amemiya (1981), Maddala (1983), McFadden (1974), Ben-Akiva and Lerman (1985), Kay and Little (1986), Veall and Zimmermann (1992), Zavoina and McKelvey (1975), Efron (1978), and Cramer (1999). A survey of techniques appears in Windmeijer (1995).

[15] See McFadden (1984) and Amemiya (1985). If this condition holds, then gradient methods *will* find that $\boldsymbol{\beta}$.

measures this failure,

$$\begin{aligned}\lambda &= (\text{average } \hat{F} \mid y_i = 1) - (\text{average } \hat{F} \mid y_i = 0) \\ &= (\text{average}(1 - \hat{F}) \mid y_i = 0) - (\text{average}(1 - \hat{F}) \mid y_i = 1).\end{aligned}$$

Cramer's measure heavily penalizes the incorrect predictions, and because each proportion is taken within the subsample, it is not unduly influenced by the large proportionate size of the group of more frequent outcomes.

A useful summary of the predictive ability of the model is a 2×2 table of the hits and misses of a prediction rule such as

$$\hat{y} = 1 \quad \text{if } \hat{F} > F^* \text{ and 0 otherwise.} \tag{17-26}$$

The usual threshold value is 0.5, on the basis that we should predict a one if the model says a one is more likely than a zero. It is important not to place too much emphasis on this measure of goodness of fit, however. Consider, for example, the naive predictor

$$\hat{y} = 1 \quad \text{if } P > 0.5 \text{ and 0 otherwise,} \tag{17-27}$$

where P is the simple proportion of ones in the sample. This rule will always predict correctly $100P$ percent of the observations, which means that the naive model does not have zero fit. In fact, if the proportion of ones in the sample is very high, it is possible to construct examples in which the second model will generate more correct predictions than the first! Once again, this flaw is not in the model; it is a flaw in the fit measure.[16] The important element to bear in mind is that the coefficients of the estimated model are not chosen so as to maximize this (or any other) fit measure, as they are in the linear regression model where **b** maximizes R^2.

Another consideration is that 0.5, although the usual choice, may not be a very good value to use for the threshold. If the sample is **unbalanced**—that is, has many more ones than zeros, or vice versa—then by this prediction rule it might never predict a one (or zero). To consider an example, suppose that in a sample of 10,000 observations, only 1,000 have $Y = 1$. We know that the average predicted probability in the sample will be 0.10. As such, it may require an extreme configuration of regressors even to produce an F of 0.2, to say nothing of 0.5. In such a setting, the prediction rule may fail every time to predict when $Y = 1$. The obvious adjustment is to reduce F^*. Of course, this adjustment comes at a cost. If we reduce the threshold F^* so as to predict $y = 1$ more often, then we will increase the number of correct classifications of observations that do have $y = 1$, but we will also increase the number of times that we *incorrectly* classify as ones observations that have $y = 0$.[17] In general, any prediction rule of the form in (17-26) will make two types of errors: It will incorrectly classify zeros as ones and ones as zeros. In practice, these errors need not be symmetric in the costs that result. For example, in a credit scoring model [see Boyes, Hoffman, and Low (1989)], incorrectly classifying an applicant as a bad risk is not the same as incorrectly classifying a bad risk as a good one. Changing F^* will always reduce the probability of one type of error

[16] See Amemiya (1981).

[17] The technique of **discriminant analysis** is used to build a procedure around this consideration. In this setting, we consider not only the number of correct and incorrect classifications, but also the cost of each type of misclassification.

while increasing the probability of the other. There is no correct answer as to the best value to choose. It depends on the setting and on the criterion function upon which the prediction rule depends.

The likelihood ratio index and various modifications of it are obviously related to the likelihood ratio statistic for testing the hypothesis that the coefficient vector is zero. Cramer' measure is oriented more toward the relationship between the fitted probabilities and the actual values. It is usefully tied to the standard prediction rule $\hat{y} = \mathbf{1}[\hat{F} > 0.5]$. Whether these it has have a close relationship to any type of fit in the familiar sense is a question that needs to be studied. In some cases, it appears so. But the maximum likelihood estimator, on which all the fit measures are based, is not chosen so as to maximize a fitting criterion based on prediction of y as it is in the classical regression (which maximizes R^2). It is chosen to maximize the joint density of the observed dependent variables. It remains an interesting question for research whether fitting y well or obtaining good parameter estimates is a preferable estimation criterion. Evidently, they need not be the same thing.

Example 17.6 Prediction with a Probit Model

Tunali (1986) estimated a probit model in a study of migration, subsequent remigration, and earnings for a large sample of observations of male members of households in Turkey. Among his results, he reports the summary shown here for a probit model: The estimated model is highly significant, with a likelihood ratio test of the hypothesis that the coefficients (16 of them) are zero based on a chi-squared value of 69 with 16 degrees of freedom.[18] The model predicts 491 of 690, or 71.2 percent, of the observations correctly, although the likelihood ratio index is only 0.083. A naive model, which always predicts that $y = 0$ because $P < 0.5$, predicts 487 of 690, or 70.6 percent, of the observations correctly. This result is hardly suggestive of no fit. The maximum likelihood estimator produces several significant influences on the probability but makes only four more correct predictions than the naive predictor.[19]

		Predicted		
		***D* = 0**	***D* = 1**	***Total***
Actual	$D = 0$	471	16	487
	$D = 1$	183	20	203
	Total	654	36	690

17.3.4 HYPOTHESIS TESTS

For testing hypotheses about the coefficients, the full menu of procedures is available. The simplest method for a single restriction would be based on the usual t tests, using the standard errors from the information matrix. Using the normal distribution of the estimator, we would use the standard normal table rather than the t table for critical points. For more involved restrictions, it is possible to use the Wald test. For a set of restrictions $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, the statistic is

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'\{\mathbf{R}(\text{Est. Asy. Var}[\hat{\boldsymbol{\beta}}])\mathbf{R}'\}^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}).$$

[18]This view actually understates slightly the significance of his model, because the preceding predictions are based on a bivariate model. The likelihood ratio test fails to reject the hypothesis that a univariate model applies, however.

[19]It is also noteworthy that nearly all the correct predictions of the maximum likelihood estimator are the zeros. It hits only 10 percent of the ones in the sample.

For example, for testing the hypothesis that a subset of the coefficients, say, the last M, are zero, the Wald statistic uses $\mathbf{R} = [\mathbf{0} \mid \mathbf{I}_M]$ and $\mathbf{q} = \mathbf{0}$. Collecting terms, we find that the test statistic for this hypothesis is

$$W = \hat{\boldsymbol{\beta}}_M' \mathbf{V}_M^{-1} \hat{\boldsymbol{\beta}}_M, \tag{17-28}$$

where the subscript M indicates the subvector or submatrix corresponding to the M variables and $\mathbf{V}$ is the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$.

Likelihood ratio and Lagrange multiplier statistics can also be computed. The likelihood ratio statistic is

$$\text{LR} = -2[\ln \hat{L}_R - \ln \hat{L}_U],$$

where $\hat{L}_R$ and $\hat{L}_U$ are the log-likelihood functions evaluated at the restricted and unrestricted estimates, respectively. A common test, which is similar to the F test that all the slopes in a regression are zero, is the **likelihood ratio test** that all the slope coefficients in the probit or logit model are zero. For this test, the constant term remains unrestricted. In this case, the restricted log-likelihood is the same for both probit and logit models,

$$\ln L_0 = n[P \ln P + (1 - P) \ln(1 - P)], \tag{17-29}$$

where P is the proportion of the observations that have dependent variable equal to 1.

It might be tempting to use the likelihood ratio test to choose between the probit and logit models. But there is no restriction involved, and the test is not valid for this purpose. To underscore the point, there is nothing in its construction to prevent the chi-squared statistic for this "test" from being negative.

The **Lagrange multiplier test** statistic is $\text{LM} = \mathbf{g}'\mathbf{V}\mathbf{g}$, where $\mathbf{g}$ is the first derivatives of the *unrestricted* model evaluated at the *restricted* parameter vector and $\mathbf{V}$ is any of the three estimators of the asymptotic covariance matrix of the maximum likelihood estimator, once again computed using the restricted estimates. Davidson and MacKinnon (1984) find evidence that $E[\mathbf{H}]$ is the best of the three estimators to use, which gives

$$\text{LM} = \left(\sum_{i=1}^{n} g_i \mathbf{x}_i \right)' \left[\sum_{i=1}^{n} E[-h_i] \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \left(\sum_{i=1}^{n} g_i \mathbf{x}_i \right), \tag{17-30}$$

where $E[-h_i]$ is defined in (17-21) for the logit model and in (17-23) for the probit model.

For the logit model, when the hypothesis is that all the slopes are zero,

$$\text{LM} = nR^2,$$

where R^2 is the uncentered coefficient of determination in the regression of $(y_i - \bar{y})$ on $\mathbf{x}_i$ and $\bar{y}$ is the proportion of 1s in the sample. An alternative formulation based on the BHHH estimator, which we developed in Section 14.6.3 is also convenient. For any of the models (probit, logit, Gumbel, etc.), the first derivative vector can be written as

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} g_i \mathbf{x}_i = \mathbf{X}'\mathbf{G}\mathbf{i},$$

where $\mathbf{G}(n \times n) = \text{diag}[g_1, g_2, \ldots, g_n]$ and $\mathbf{i}$ is an $n \times 1$ column of 1s. The BHHH estimator of the Hessian is $(\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X})$, so the LM statistic based on this estimator is

$$\text{LM} = n\left[\frac{1}{n}\mathbf{i}'(\mathbf{G}\mathbf{X})(\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{G}')\mathbf{i}\right] = nR_{\mathbf{i}}^2, \tag{17-31}$$

where $R_{\mathbf{i}}^2$ is the uncentered coefficient of determination in a regression of a column of ones on the first derivatives of the logs of the individual probabilities.

All the statistics listed here are asymptotically equivalent and under the null hypothesis of the restricted model have limiting chi-squared distributions with degrees of freedom equal to the number of restrictions being tested. We consider some examples in the next section.

Example 17.7 Testing for Structural Break in a Logit Model

The model in Example 17.4, based on Riphahn, Wambach, and Million (2003), is

$$\text{Prob}(DocVis_{it} > 0) = \Lambda(\beta_1 + \beta_2 Age_{it} + \beta_3 Income_{it} + \beta_4 Kids_{it} + \beta_5 Education_{it} + \beta_6 Married_{it}).$$

In the original study, the authors split the sample on the basis of gender, and fit separate models for male and female headed households. We will use the preceding results to test for the appropriateness of the sample splitting. This test of the pooling hypothesis is a counterpart to the **Chow test** of structural change in the linear model developed in Section 6.4.1. Since we are not using least squares (in a linear model), we use the likelihood based procedures rather than an F test as we did earlier. Estimates of the three models are shown in Table 17.4. The chi-squared statistic for the likelihood ratio test is

$$\text{LR} = -2[-17673.09788 - (-9541.77802 - 7855.96999)] = 550.69744.$$

The 95 percent critical value for six degrees of freedom is 12.592. To carry out the Wald test for this hyothesis there are two numerically identical ways to proceed. First, using the estimates for Male and Female samples separately, we can compute a chi-squared statistic to test the hypothesis that the difference of the two coefficients is zero. This would be

$$W = [\hat{\beta}_{Male} - \hat{\beta}_{Female}]'[Est.Asy.Var(\hat{\beta}_{Male}) + Est.Asy.Var(\hat{\beta}_{Female})]^{-1}[\hat{\beta}_{Male} - \hat{\beta}_{Female}] = 538.13629.$$

Another way to obtain the same result is to add to the pooled model the original 6 variables now multiplied by the *Female* dummy variable. We use the augmented $\mathbf{X}$ matrix

TABLE 17.4 Estimated Models for Pooling Hypothesis

	Pooled Sample		*Male*		*Female*	
Variable	*Estimate*	*Std.Error*	*Estimate*	*Std.Error*	*Estimate*	*Std.Error*
Constant	0.25112	0.09114	−0.20881	0.11475	0.44767	0.16016
Age	0.02071	0.00129	0.02375	0.00178	0.01331	0.00202
Income	−0.18592	0.07506	−0.23059	0.10415	−0.17182	0.11225
Kids	−0.22947	0.02954	−0.26149	0.04054	−0.27153	0.04539
Education	−0.04559	0.00565	−0.04251	0.00737	−0.00170	0.00970
Married	0.08529	0.03329	0.17451	0.04833	0.03621	0.04864
ln L	−17673.09788		−9541.77802		−7855.96999	

$\mathbf{X}^* = [\mathbf{X}, \textit{female} \times \mathbf{X}]$. The model with 12 variables is now estimated, and a test of the pooling hypothesis is done by testing the joint hypothesis that the coefficients on these 6 additional variables are zero. The Lagrange multiplier test is carried out by using this augmented model as well. To apply (17-31), the necessary derivatives are in (17-18). For the logit model, the derivative matrix is simply $\mathbf{G}^* = \text{diag}[y_i - \Lambda(\mathbf{x}_i^{*\prime}\boldsymbol{\beta})]$. For the LM test, the vector $\boldsymbol{\beta}$ that is used is the one for the restricted model. Thus, $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}'_{Pooled}, 0, 0, 0, 0, 0, 0)'$. The estimated probabilities that appear in $\mathbf{G}^*$ are simply those obtained from the pooled model. Then,

$$\text{LM} = \mathbf{i}'\mathbf{G}^*\mathbf{X}^* \times [(\mathbf{X}^{*\prime}\mathbf{G}'^*)(\mathbf{G}^*\mathbf{X})]^{-1}\mathbf{X}^{*\prime}\mathbf{G}^{*\prime}\mathbf{i} = 548.17052.$$

The pooling hypothesis is rejected by all three procedures.

17.3.5 ENDOGENOUS RIGHT-HAND-SIDE VARIABLES IN BINARY CHOICE MODELS

The analysis in Example 17.8 (Labor Supply Model) suggests that the presence of endogenous right-hand-side variables in a binary choice model presents familiar problems for estimation. The problem is made worse in nonlinear models because even if one has an instrumental variable readily at hand, it may not be immediately clear what is to be done with it. The instrumental variable estimator described in Chapter 8 is based on moments of the data, variances, and covariances. In this binary choice setting, we are not using any form of least squares to estimate the parameters, so the IV method would appear not to apply. Generalized method of moments is a possibility.

$$\begin{aligned} y_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \gamma w_i + \varepsilon_i, \\ y_i &= 1(y_i^* > 0), \\ E[\varepsilon_i \mid w_i] &= g(w_i) \neq 0. \end{aligned}$$

Thus, w_i is endogenous in this model. The maximum likelihood estimators considered earlier will not consistently estimate $(\boldsymbol{\beta}, \gamma)$. [Without an additional specification that allows us to formalize $\text{Prob}(y_i = 1 \mid \mathbf{x}_i, w_i)$, we cannot state what the MLE will, in fact, estimate.] Suppose that we have a "relevant" (see Section 8.2) instrumental variable, z_i such that

$$\begin{aligned} E[\varepsilon_i \mid z_i, \mathbf{x}_i] &= 0, \\ E[w_i z_i] &\neq 0. \end{aligned}$$

A natural instrumental variable estimator would be based on the "moment" condition

$$E\left[\left(y_i^* - \mathbf{x}_i'\boldsymbol{\beta} - \gamma w_i\right)\begin{pmatrix}\mathbf{x}_i \\ z_i\end{pmatrix}\right] = \mathbf{0}.$$

However, y_i^* is not observed, y_i is. But the "residual," $y_i - \mathbf{x}_i'\boldsymbol{\beta} - \gamma w_i$, would have no meaning even if the true parameters were known.[20] One approach that was used in Avery et al. (1983), Butler and Chatterjee (1997), and Bertschek and Lechner (1998) is to assume that the instrumental variable is orthogonal to the residual $[y - \Phi(\mathbf{x}_i'\boldsymbol{\beta} + \gamma w_i)]$;

[20] One would proceed in precisely this fashion if the central specification were a linear probability model (LPM) to begin with. See, for example, Eisenberg and Rowe (2006) or Angrist (2001) for an application and some analysis of this case.

that is,

$$E\left[[y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta} + \gamma w_i)]\begin{pmatrix}\mathbf{x}_i\\ z_i\end{pmatrix}\right] = \mathbf{0}.$$

This form of the moment equation, based on observables, can form the basis of a straightforward two-step GMM estimator. (See Chapter 13 for details.)

The GMM estimator is not less parametric than the full information maximum likelihood estimator described later because the probit model based on the normal distribution is still invoked to specify the moment equation.[21] Nothing is gained in simplicity or robustness of this approach to full information maximum likelihood estimation, which we now consider. (As Bertschek and Lechner argue, however, the gains might come in terms of practical implementation and computation time. The same considerations motivated Avery et al.)

This maximum likelihood estimator requires a full specification of the model, including the assumption that underlies the endogeneity of w_i. This becomes essentially a simultaneous equations model. The model equations are

$$\begin{aligned}
&y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \gamma w_i + \varepsilon_i, y_i = 1[y_i^* > 0],\\
&w_i = \mathbf{z}_i'\boldsymbol{\alpha} + u_i,\\
&(\varepsilon_i, u_i) \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\sigma_u\\ \rho\sigma_u & \sigma_u^2\end{pmatrix}\right].
\end{aligned}$$

(We are assuming that there is a vector of instrumental variables, $\mathbf{z}_i$.) Probit estimation based on y_i and $(\mathbf{x}_i, w_i)$ will not consistently estimate $(\boldsymbol{\beta}, \gamma)$ because of the correlation between w_i and ε_i induced by the correlation between u_i and ε_i. Several methods have been proposed for estimation of this model. One possibility is to use the partial reduced form obtained by inserting the second equation in the first. This becomes a probit model with probability $\text{Prob}(y_i = 1 \mid \mathbf{x}_i, \mathbf{z}_i) = \Phi(\mathbf{x}_i'\boldsymbol{\beta}^* + \mathbf{z}_i'\boldsymbol{\alpha}^*)$. This will produce consistent estimates of $\boldsymbol{\beta}^* = \boldsymbol{\beta}/(1 + \gamma^2\sigma_u^2 + 2\gamma\sigma_u\rho)^{1/2}$ and $\boldsymbol{\alpha}^* = \gamma\boldsymbol{\alpha}/(1 + \gamma^2\sigma_u^2 + 2\gamma\sigma_u\rho)^{1/2}$ as the coefficients on $\mathbf{x}_i$ and $\mathbf{z}_i$, respectively. (The procedure will estimate a mixture of $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^*$ for any variable that appears in both $\mathbf{x}_i$ and $\mathbf{z}_i$.) In addition, linear regression of w_i on $\mathbf{z}_i$ produces estimates of $\boldsymbol{\alpha}$ and σ_u^2, but there is no method of moments estimator of ρ or γ produced by this procedure, so this estimator is incomplete. Newey (1987) suggested a "minimum chi-squared" estimator that does estimate all parameters. A more direct, and actually simpler approach is full information maximum likelihood.

The log-likelihood is built up from the joint density of y_i and w_i, which we write as the product of the conditional and the marginal densities,

$$f(y_i, w_i) = f(y_i \mid w_i)\, f(w_i).$$

To derive the conditional distribution, we use results for the bivariate normal, and write

$$\varepsilon_i \mid u_i = \left[(\rho\sigma_u)/\sigma_u^2\right] u_i + v_i,$$

[21] This is precisely the platform that underlies the GLIM/GEE treatment of binary choice models in, for example, the widely used programs *SAS* and *Stata*.

where v_i is normally distributed with $\text{Var}[v_i] = (1 - \rho^2)$. Inserting this in the first equation, we have

$$y_i^* \mid w_i = \mathbf{x}_i'\boldsymbol{\beta} + \gamma w_i + (\rho/\sigma_u)u_i + v_i.$$

Therefore,

$$\text{Prob}[y_i = 1 \mid \mathbf{x}_i, w_i] = \boldsymbol{\Phi}\left[\frac{\mathbf{x}_i'\boldsymbol{\beta} + \gamma w_i + (\rho/\sigma_u)u_i}{\sqrt{1-\rho^2}}\right]. \tag{17-32}$$

Inserting the expression for $u_i = (w_i - \mathbf{z}_i'\boldsymbol{\alpha})$, and using the normal density for the marginal distribution of w_i in the second equation, we obtain the log-likelihood function for the sample,

$$\ln L = \sum_{i=1}^{n} \ln \boldsymbol{\Phi}\left[(2y_i - 1)\left(\frac{\mathbf{x}_i'\boldsymbol{\beta} + \gamma w_i + (\rho/\sigma_u)(w_i - \mathbf{z}_i'\boldsymbol{\alpha})}{\sqrt{1-\rho^2}}\right)\right] + \ln\left[\frac{1}{\sigma_u}\boldsymbol{\phi}\left(\frac{w_i - \mathbf{z}_i'\boldsymbol{\alpha}}{\sigma_u}\right)\right].$$

Example 17.8 Labor Supply Model

In Examples 5.2 and 17.1, we examined a labor suppy model for married women using Mroz's (1987) data on labor supply. The wife's labor force participation equation suggested in Example 17.1 is

$$\text{Prob}(LFP_i = 1) = \Phi\left(\beta_1 + \beta_2 Age_i + \beta_3 Age_i^2 + \beta_4 Education_i + \beta_5 Kids_i\right).$$

A natural extension of this model would be to include the husband's hours in the equation,

$$\text{Prob}\left(LFP_i = 1\right) = \Phi(\beta_1 + \beta_2 Age_i + \beta_3 Age_i^2 + \beta_4 Education_i + \beta_5 Kids_i + \gamma\, HHrs_i\right).$$

It would also be natural to assume that the husband's hours would be correlated with the determinants (observed and unobserved) of the wife's labor force participation. The auxiliary equation might be

$$HHrs_i = \alpha_1 + \alpha_2 HAge_i + \alpha_3 HEducation_i + \alpha_4 Family\ Income_i + u_i.$$

As before, we use the Mroz (1987) labor supply data described in Example 5.2. Table 17.5 reports the single-equation and maximum likelihood estimates of the parameters of the two equations. Comparing the two sets of probit estimates, it appears that the (assumed) endogeneity of the husband's hours is not substantially affecting the estimates. There are two

TABLE 17.5 Estimated Labor Supply Model

	Probit		*Regression*		*Maximum Likelihood*	
Constant	−3.86704	(1.41153)			−5.08405	(1.43134)
Age	0.18681	(0.065901)			0.17108	(0.063321)
Age^2	−0.00243	(0.000774)			−0.00219	(0.0007629)
Education	0.11098	(0.021663)			0.09037	(0.029041)
Kids	−0.42652	(0.13074)			−0.40202	(0.12967)
Husband hours	−0.000173	(0.0000797)			0.00055	(0.000482)
Constant			2325.38	(167.515)	2424.90	(158.152)
Husband age			−6.71056	(2.73573)	−7.3343	(2.57979)
Husband education			9.29051	(7.87278)	2.1465	(7.28048)
Family income			55.72534	(19.14917)	63.4669	(18.61712)
σ_u			588.2355		586.994	
ρ			0.0000		−0.4221	(0.26931)
ln L	−489.0766		−5868.432		−6357.093	

simple ways to test the hypothesis that ρ equals zero. The FIML estimator produces an estimated asymptotic standard error with the estimate of ρ, so a Wald test can be carried out. For the preceding results, the Wald statistic would be $(-0.4221/0.26921)^2 = 2.458$. The critical value from the chi-squared table for one degree of freedom would be 3.84, so we would not reject the hypothesis. The second approach would use the likelihood ratio test. Under the null hypothesis of exogeneity, the probit model and the regression equation can be estimated independently. The log-likelihood for the full model would be the sum of the two log-likelihoods, which would be −6357.508 based on the following results. Without the restriction $\rho = 0$, the combined log likelihood is −6357.093. Twice the difference is 0.831, which is also well under the 3.84 critical value, so on this basis as well, we would not reject the null hypothesis that $\rho = 0$.

Blundell and Powell (2004) label the foregoing the **control function** approach to accommodating the endogeneity. As noted, the estimator is fully parametric. They propose an alternative semiparametric approach that retains much of the functional form specification, but works around the specific distributional assumptions. Adapting their model to our earlier notation, their departure point is a general specification that produces, once again, a control function,

$$E[y_i \mid \mathbf{x}_i, w_i, u_i] = F(\mathbf{x}_i'\boldsymbol{\beta} + \gamma w_i, u_i).$$

Note that (17-32) satisfies the assumption; however, they reach this point without assuming either joint or marginal normality. The authors propose a three-step, semiparametric approach to estimating the structural parameters. In an application somewhat similar to Example 17.8, they apply the technique to a labor force participation model for British men in which a variable of interest is a dummy variable for education greater than 16 years, the endogenous variable in the participation equation, also of interest, is earned income of the spouse, and an instrumental variable is a welfare benefit entitlement. Their findings are rather more substantial than ours; they find that when the endogeneity of other family income is accommodated in the equation, the education coefficient increases by 40 percent and remains significant, but the coefficient on other income increases by more than tenfold.

In the control function model noted earlier, where $E[y_i \mid \mathbf{x}_i, w_i, u_i] = F(\mathbf{x}_i'\boldsymbol{\beta} + \gamma w_i, u_i)$ and $w_i = \mathbf{z}_i'\boldsymbol{\alpha} + u_i$, since the covariance of w_i and u_i is the issue, it might seem natural to solve the problem by replacing w_i with $\mathbf{z}_i'\mathbf{a}$ where $\mathbf{a}$ is an estimator of $\boldsymbol{\alpha}$, or some other prediction of w_i based only on exogenous variables. The earlier development shows that the appropriate approach is to add the estimated residual to the equation, instead. The issue is explored in detail by Terza, Basu, and Rathouz (2008), who reach the same conclusion in a general model.

The residual inclusion method also suggests a two-step approach. Rewrite the log-likelihood function as

$$\ln L = \sum_{i=1}^{n} \ln \Phi\left[(2y_i - 1)(\mathbf{x}_i'\boldsymbol{\beta}^* + \gamma^* w_i + \tau\tilde{\varepsilon}_i)\right] + \sum_{i=1}^{n} \ln\left[\frac{1}{\sigma_u}\boldsymbol{\phi}(\tilde{\varepsilon}_i)\right],$$

where $\boldsymbol{\beta}^* = (1/\sqrt{1-\rho^2})\boldsymbol{\beta}$, $\gamma^* = (1/\sqrt{1-\rho^2})\gamma$, $\tau = (\rho/\sqrt{1-\rho^2})$ and $\tilde{\varepsilon}_i = (w_i - \mathbf{z}_i'\boldsymbol{\alpha})/\sigma_u$.

The parameters in the regression, $\boldsymbol{\alpha}$ and σ_u, can be consistently estimated by a linear regression of w on $\mathbf{z}$. The scaled residual $\tilde{e}_i = (w_i - \mathbf{z}_i'\mathbf{a})/s_u$ can now be computed and inserted into the log-likelihood. Note that the second term in the log-likelihood involves parameters that have already been estimated at the first step. The second-step

log-likelihood is, then,

$$\ln L = \sum_{i=1}^{n} \ln \Phi \left[(2y_i - 1)(\mathbf{x}_i'\boldsymbol{\beta}^* + \gamma^* w_i + \tau \tilde{e}_i) \right].$$

This can be maximized using the methods developed in Section 17.3. The estimator of ρ can be recovered from $\rho = \tau/(1 + \tau^2)^{1/2}$. Estimators of $\boldsymbol{\beta}$ and γ follow, and the delta method can be used to construct standard errors. Since this is a two-step estimator, the resulting estimator of the asymptotic covariance matrix would be further adjusted using the Murphy and Topel (2002) results in Section 14.7. Bootstrapping the entire apparatus (see Section 15.4) would be an alternative way to estimate an asymptotic covariance matrix. The original (one-step) log-likelihood is not very complicated, and full information estimation is fairly straightforward. The preceding demonstrates how the alternative two-step method would proceed and emphasizes once again, the appropriateness of the "residual inclusion" method.

The case in which the endogenous variable in the main equation is, itself, a binary variable occupies a large segment of the recent literature. Consider the model

$$\begin{aligned}
&T_i^* = \mathbf{z}'_i\boldsymbol{\alpha} + u_i,\ T_i = 1[w_i^* > 0], \\
&y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i + \varepsilon_i,\ y_i = 1[y_i^* > 0], \\
&\begin{pmatrix} \varepsilon_i \\ u_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right],
\end{aligned}$$

where T_i is a binary variable indicating some kind of program participation (e.g., graduating from high school or college, receiving some kind of job training, purchasing health insurance, etc.). The model in this form (and several similar ones) is a "treatment effects" model. The subject of treatment effects models is surveyed in many studies, including Angrist (2001) and Angrist and Pischke (2009, 2010). The main object of estimation is γ (at least superficially). In these settings, the observed outcome may be y_i* (e.g., income or hours) or y_i (e.g., labor force participation). We have considered the first case in Chapter 8, and will revisit it in Chapter 19. The case just examined is that in which y_i and T_i^* are the observed variables. The preceding analysis has suggested that problems of endogeneity will intervene in all cases. We will examine this model in some detail in Section 17.5.5 and in Chapter 19.

17.3.6 ENDOGENOUS CHOICE-BASED SAMPLING

In some studies [e.g., Boyes, Hoffman, and Low (1989), Greene (1992)], the mix of ones and zeros in the observed sample of the dependent variable is deliberately skewed in favor of one outcome or the other to achieve a more balanced sample than random sampling would produce. The sampling is said to be **choice based**. In the studies noted, the dependent variable measured the occurrence of loan default, which is a relatively uncommon occurrence. To enrich the sample, observations with $y = 1$ (default) were oversampled. Intuition should suggest (correctly) that the bias in the sample should be transmitted to the parameter estimates, which will be estimated so as to mimic the sample, not the population, which is known to be different. Manski and Lerman (1977) derived the weighted endogenous sampling maximum likelihood (WESML) estimator for this situation. The estimator requires that the true population proportions, ω_1

and ω_0, be known. Let p_1 and p_0 be the sample proportions of ones and zeros. Then the estimator is obtained by maximizing a weighted log-likelihood,

$$\ln L = \sum_{i=1}^{n} w_i \ln F(q_i \mathbf{x}_i' \boldsymbol{\beta}),$$

where $w_i = y_i(\omega_1/p_1) + (1 - y_i)(\omega_0/p_0)$. Note that w_i takes only two different values. The derivatives and the Hessian are likewise weighted. A final correction is needed after estimation; the appropriate estimator of the asymptotic covariance matrix is the sandwich estimator discussed in Section 17.3.1, $\mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}$ (with weighted $\mathbf{B}$ and $\mathbf{H}$), instead of $\mathbf{B}$ or $\mathbf{H}$ alone. (The weights are not squared in computing $\mathbf{B}$.)[22]

Example 17.9 Credit Scoring

In Example 7.9, we examined the spending patterns of a sample of 10,499 cardholders for a major credit card vendor. The sample of cardholders is a subsample of 13,444 applicants for the credit card. Applications for credit cards, then (1992) and now are processed by a major nationwide processor, Fair Isaacs, Inc. The algorithm used by the processors is proprietary. However, conventional wisdom holds that a few variables are important in the process, such as Age, Income, whether the applicant owns their home, whether they are self-employed, and how long they have lived at their current address. The number of major and minor derogatory reports (60-day and 30-day delinquencies) are very influential variables in credit scoring. The probit model we will use to 'model the model' is

$$\begin{aligned}\text{Prob}(\textit{Cardholder} = 1) &= \text{Prob}(C = 1 \mid \mathbf{x}) \\ &= \Phi(\beta_1 + \beta_2\, \textit{Age} + \beta_3\, \textit{Income} + \beta_4\, \textit{OwnRent} \\ &\quad + \beta_5\, \textit{Months Living at Current Address} \\ &\quad + \beta_6\, \textit{Self-Employed} \\ &\quad + \beta_7\, \textit{Number of major derogatory reports} \\ &\quad + \beta_8\, \textit{Number of minor derogatory reports}.\end{aligned}$$

In the data set, 78.1 percent of the applicants are cardholders. In the population, at that time, the true proportion was roughly 23.2 percent, so the sample is substantially choice based on this variable. The sample was deliberately skewed in favor of cardholders for purposes of the original study [Greene (1992)]. The weights to be applied for the WESML estimator are $0.232/0.781 = 0.297$ for the observations with $C = 1$ and $0.768/0.219 = 3.507$ for observations with $C = 0$. Table 17.6 presents the unweighted and weighted estimates for this application. The change in the estimates produced by the weighting is quite modest, save for the constant term. The results are consistent with the conventional wisdom that *Income* and *OwnRent* are two important variables in a credit application and self-employment receives a substantial negative weight. But, as might be expected, the single most significant influence on cardholder status is major derogatory reports. Since lenders are strongly focused on default probability, past evidence of default behavior will be a major consideration.

17.3.7 SPECIFICATION ANALYSIS

In his survey of qualitative response models, Amemiya (1981) reports the following widely cited approximations for the linear probability (LP) model: Over the range of

[22] WESML and the choice-based sampling estimator are not the free lunch they may appear to be. That which the biased sampling does, the weighting undoes. It is common for the end result to be very large standard errors, which might be viewed as unfortunate, insofar as the purpose of the biased sampling was to balance the data precisely to avoid this problem.

TABLE 17.6 Estimated Card Application Equation (*t* ratios in parentheses)

	Unweighted			*Weighted*		
Variable	***Estimate***	***Standard Error***		***Estimate***	***Standard Error***	
Constant	0.31783	0.05094	(6.24)	−1.13089	0.04725	(−23.94)
Age	0.00184	0.00154	(1.20)	0.00156	0.00145	(1.07)
Income	0.00095	0.00025	(3.86)	0.00094	0.00024	(3.92)
OwnRent	0.18233	0.03061	(5.96)	0.23967	0.02968	(8.08)
CurrentAddress	0.02237	0.00120	(18.67)	0.02106	0.00109	(19.40)
SelfEmployed	−0.43625	0.05585	(−7.81)	−0.47650	0.05851	(−8.14)
Major Derogs	−0.69912	0.01920	(−36.42)	−0.64792	0.02525	(−25.66)
Minor Derogs	−0.04126	0.01865	(−2.21)	−0.04285	0.01778	(−2.41)

probabilities of 30 to 70 percent,

$$\hat{\boldsymbol{\beta}}_{LP} \approx 0.4\boldsymbol{\beta}_{probit} \text{ for the slopes,}$$

$$\hat{\boldsymbol{\beta}}_{LP} \approx 0.25\boldsymbol{\beta}_{logit} \text{ for the slopes.}$$

Aside from confirming our intuition that least squares approximates the nonlinear model and providing a quick comparison for the three models involved, the practical usefulness of the formula is somewhat limited. Still, it is a striking result.[23] A series of studies has focused on reasons why the least squares estimates should be proportional to the probit and logit estimates. A related question concerns the problems associated with assuming that a probit model applies when, in fact, a logit model is appropriate or vice versa.[24] The approximation would seem to suggest that with this type of misspecification, we would once again obtain a scaled version of the correct coefficient vector. (Amemiya also reports the widely observed relationship $\hat{\boldsymbol{\beta}}_{logit} \approx 1.6\hat{\boldsymbol{\beta}}_{probit}$, which follows from the results for the linear probability model. This result is apparent in Table 17.1 where the ratios of the three slopes range from 1.6 to 1.9.)

In the linear regression model, we considered two important specification problems: the effect of omitted variables and the effect of heteroscedasticity. In the classical model, $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, when least squares estimates $\mathbf{b}_1$ are computed omitting $\mathbf{X}_2$,

$$E[\mathbf{b}_1] = \boldsymbol{\beta}_1 + [\mathbf{X}_1'\mathbf{X}_1]^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2.$$

Unless $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal or $\boldsymbol{\beta}_2 = \mathbf{0}$, $\mathbf{b}_1$ is biased. If we ignore heteroscedasticity, then although the least squares estimator is still unbiased and consistent, it is inefficient and the usual estimate of its sampling covariance matrix is inappropriate. Yatchew and Griliches (1984) have examined these same issues in the setting of the probit and logit models. Their general results are far more pessimistic. In the context of a binary choice model, they find the following:

[23] This result does not imply that it is useful to report 2.5 times the linear probability estimates with the probit estimates for comparability. The linear probability estimates are already in the form of marginal effects, whereas the probit coefficients must be scaled *downward*. If the sample proportion happens to be close to 0.5, then the right scale factor will be roughly $\phi[\Phi^{-1}(0.5)] = 0.3989$. But the density falls rapidly as P moves away from 0.5.

[24] See Ruud (1986) and Gourieroux et al. (1987).

1. If x_2 is omitted from a model containing x_1 and x_2, (i.e. $\boldsymbol{\beta}_2 \neq 0$) then
$$\text{plim}\,\hat{\boldsymbol{\beta}}_1 = c_1\boldsymbol{\beta}_1 + c_2\boldsymbol{\beta}_2,$$
where c_1 and c_2 are complicated functions of the unknown parameters. The implication is that even if the omitted variable is uncorrelated with the included one, the coefficient on the included variable will be inconsistent.
2. If the disturbances in the underlying regression are heteroscedastic, then the maximum likelihood estimators are inconsistent and the covariance matrix is inappropriate.

The second result is particularly troubling because the probit model is most often used with microeconomic data, which are frequently heteroscedastic.

Any of the three methods of hypothesis testing discussed here can be used to analyze these specification problems. The Lagrange multiplier test has the advantage that it can be carried out using the estimates from the restricted model, which sometimes brings a large saving in computational effort. This situation is especially true for the test for **heteroscedasticity.**[25]

To reiterate, the Lagrange multiplier statistic is computed as follows. Let the null hypothesis, H_0, be a specification of the model, and let H_1 be the alternative. For example, H_0 might specify that only variables $\mathbf{x}_1$ appear in the model, whereas H_1 might specify that $\mathbf{x}_2$ appears in the model as well. The statistic is
$$\text{LM} = \mathbf{g}_0'\mathbf{V}_0^{-1}\mathbf{g}_0,$$
where $\mathbf{g}_0$ is the vector of derivatives of the log-likelihood as specified by H_1 but evaluated at the maximum likelihood estimator of the parameters assuming that H_0 is true, and $\mathbf{V}_0^{-1}$ is any of the three consistent estimators of the asymptotic variance matrix of the maximum likelihood estimator under H_1, also computed using the maximum likelihood estimators based on H_0. The statistic is asymptotically distributed as chi-squared with degrees of freedom equal to the number of restrictions.

17.3.7.a Omitted Variables

The hypothesis to be tested is
$$\begin{aligned} H_0&: y^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon, \\ H_1&: y^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon, \end{aligned} \tag{17-33}$$
so the test is of the null hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$. The Lagrange multiplier test would be carried out as follows:

1. Estimate the model in H_0 by maximum likelihood. The restricted coefficient vector is $[\hat{\boldsymbol{\beta}}_1, \mathbf{0}]$.
2. Let $\mathbf{x}$ be the compound vector, $[\mathbf{x}_1, \mathbf{x}_2]$.

The statistic is then computed according to (17-30) or (17-31). It is noteworthy that in this case as in many others, the Lagrange multiplier is the coefficient of determination in a regression. The likelihood ratio test is equally straightforward. Using the estimates of the two models, the statistic is simply $2(\ln L_1 - \ln L_0)$.

[25] The results in this section are based on Davidson and MacKinnon (1984) and Engle (1984). A symposium on the subject of specification tests in discrete choice models is Blundell (1987).

17.3.7.b Heteroscedasticity

We use the general formulation analyzed by Harvey (1976) (see Section 14.9.2.a),[26]

$$\text{Var}[\varepsilon] = [\exp(\mathbf{z}'\boldsymbol{\gamma})]^2.$$

This model can be applied equally to the probit and logit models. We will derive the results specifically for the probit model; the logit model is essentially the same. Thus,

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

$$\text{Var}[\varepsilon \mid \mathbf{x}, \mathbf{z}] = [\exp(\mathbf{z}'\boldsymbol{\gamma})]^2. \tag{17-34}$$

The presence of heteroscedasticity makes some care necessary in interpreting the coefficients for a variable w_k that could be in $\mathbf{x}$ or $\mathbf{z}$ or both,

$$\frac{\partial \text{Prob}(Y = 1 \mid \mathbf{x}, \mathbf{z})}{\partial w_k} = \phi\left[\frac{\mathbf{x}'\boldsymbol{\beta}}{\exp(\mathbf{z}'\boldsymbol{\gamma})}\right]\frac{\beta_k - (\mathbf{x}'\boldsymbol{\beta})\gamma_k}{\exp(\mathbf{z}'\boldsymbol{\gamma})}.$$

Only the first (second) term applies if w_k appears only in $\mathbf{x}$ ($\mathbf{z}$). This implies that the simple coefficient may differ radically from the effect that is of interest in the estimated model. This effect is clearly visible in the next example.

The log-likelihood is

$$\ln L = \sum_{i=1}^{n}\left\{y_i \ln F\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\right) + (1 - y_i)\ln\left[1 - F\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\right)\right]\right\}. \tag{17-35}$$

To be able to estimate all the parameters, $\mathbf{z}$ cannot have a constant term. The derivatives are

$$\begin{aligned}\frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}\left[\frac{f_i(y_i - F_i)}{F_i(1 - F_i)}\right]\exp(-\mathbf{z}_i'\boldsymbol{\gamma})\mathbf{x}_i,\\ \frac{\partial \ln L}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^{n}\left[\frac{f_i(y_i - F_i)}{F_i(1 - F_i)}\right]\exp(-\mathbf{z}_i'\boldsymbol{\gamma})\mathbf{z}_i(-\mathbf{x}_i'\boldsymbol{\beta}),\end{aligned} \tag{17-36}$$

which implies a difficult log-likelihood to maximize. But if the model is estimated assuming that $\boldsymbol{\gamma} = \mathbf{0}$, then we can easily test for homoscedasticity. Let

$$\mathbf{w}_i = \begin{bmatrix}\mathbf{x}_i\\ (-\mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{z}_i\end{bmatrix}, \tag{17-37}$$

computed at the maximum likelihood estimator, assuming that $\boldsymbol{\gamma} = \mathbf{0}$. Then (17-30) or (17-31) can be used as usual for the Lagrange multiplier statistic.

Davidson and MacKinnon carried out a Monte Carlo study to examine the true sizes and power functions of these tests. As might be expected, the test for omitted variables is relatively powerful. The test for heteroscedasticity may well pick up some other form of misspecification, however, including perhaps the simple omission of $\mathbf{z}$ from the index function, so its power may be problematic. It is perhaps not surprising that the same problem arose earlier in our test for heteroscedasticity in the linear regression model.

[26] See Knapp and Seaks (1992) for an application. Other formulations are suggested by Fisher and Nagin (1981), Hausman and Wise (1978), and Horowitz (1993).

Example 17.10 Specification Tests in a Labor Force Participation Model

Using the data described in Example 17.1, we fit a probit model for labor force participation based on the specification

$$\text{Prob}[LFP = 1] = F(constant, age, age^2, family\ income, education, kids).$$

For these data, $P = 428/753 = 0.568393$. The restricted (all slopes equal zero, free constant term) log-likelihood is $325 \times \ln(325/753) + 428 \times \ln(428/753) = -514.8732$. The unrestricted log-likelihood for the probit model is -490.8478. The chi-squared statistic is, therefore, 48.05072. The critical value from the chi-squared distribution with five degrees of freedom is 11.07, so the joint hypothesis that the coefficients on *age*, age^2, *family income,* and *kids* are all zero is rejected.

Consider the alternative hypothesis, that the constant term and the coefficients on *age*, age^2, *family income*, and *education* are the same whether *kids* equals one or zero, against the alternative that an altogether different equation applies for the two groups of women, those with *kids* $= 1$ and those with *kids* $= 0$. To test this hypothesis, we would use a counterpart to the **Chow test** of Section 6.4 and Example 6.9. The restricted model in this instance would be based on the pooled data set of all 753 observations. The log-likelihood for the pooled model—which has a constant term, *age*, age^2, *family income*, and *education* is -496.8663. The log-likelihoods for this model based on the 524 observations with *kids* $= 1$ and the 229 observations with *kids* $= 0$ are -347.87441 and -141.60501, respectively. The log-likelihood for the unrestricted model with separate coefficient vectors is thus the sum, -489.47942. The chi-squared statistic for testing the five restrictions of the pooled model is twice the difference, $\text{LR} = 2[-489.47942 - (-496.8663)] = 14.7738$. The 95 percent critical value from the chi-squared distribution with 5 degrees of freedom is 11.07, so at this significance level, the hypothesis that the constant terms and the coefficients on *age*, age^2, *family income*, and *education* are the same is rejected. (The 99 percent critical value is 15.09.)

Table 17.7 presents estimates of the probit model with a correction for heteroscedasticity of the form

$$\text{Var}[\varepsilon_i] = \exp(\gamma_1 kids + \gamma_2 family\ income).$$

The three tests for homoscedasticity give

$$\text{LR} = 2[-487.6356 - (-490.8478)] = 6.424,$$

$$\text{LM} = 2.236 \text{ based on the BHHH estimator,}$$

$$\text{Wald} = 6.533 \text{ (2 restrictions)}.$$

The 95 percent critical value for two restrictions is 5.99, so the LM statistic conflicts with the other two.

TABLE 17.7 Estimated Coefficients

		Estimate (Std. Er)	*Marg. Effect**	*Estimate (St. Er.)*	*Marg. Effect**
Constant	β_1	−4.157(1.402)	−0.00837(0.0028)	−6.030(2.498)	−0.00825(.00649)
Age	β_2	0.185(0.0660)	−0.0079(0.0027)	0.264(0.118)	−0.0088(0.00251)
Age^2	β_3	−0.0024(0.00077)	—	−0.0036(0.0014)	—
Income	β_4	0.0458(0.0421)	0.0180(0.0165)	0.424(0.222)	0.0552(0.0240)
Education	β_5	0.0982(0.0230)	0.0385(0.0090)	0.140(0.0519)	0.0289(0.00869)
Kids	β_6	−0.449(0.131)	−0.171(0.0480)	−0.879(0.303)	−0.167(0.0779)
Kids	γ_1	0.000	—	−0.141(0.324)	—
Income	γ_2	0.000	—	0.313(0.123)	—
ln L		−490.8478		−487.6356	
Correct Preds.		0s: 106, 1s: 357		0s: 115, 1s: 358	

*Marginal effect and estimated standard error include both mean ($\boldsymbol{\beta}$) and variance ($\boldsymbol{\gamma}$) effects.

17.4 BINARY CHOICE MODELS FOR PANEL DATA

Qualitative response models have been a growth industry in econometrics. The recent literature, particularly in the area of panel data analysis, has produced a number of new techniques. The availability of high-quality panel data sets on microeconomic behavior has maintained an interest in extending the models of Chapter 11 to binary (and other discrete choice) models. In this section, we will survey a few results from this rapidly growing literature.

The structural model for a possibly unbalanced panel of data would be written

$$\begin{aligned} y_{it}^* &= \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T_i, \\ y_{it} &= 1 \quad \text{if } y_{it}^* > 0, \text{ and 0 otherwise,} \end{aligned} \tag{17-38}$$

The second line of this definition is often written

$$y_{it} = \mathbf{1}(\mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} > 0)$$

to indicate a variable that equals one when the condition in parentheses is true and zero when it is not. Ideally, we would like to specify that ε_{it} and ε_{is} are freely correlated within a group, but uncorrelated across groups. But doing so will involve computing joint probabilities from a T_i variate distribution, which is generally problematic.[27] (We will return to this issue later.) A more promising approach is an effects model,

$$\begin{aligned} y_{it}^* &= \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it} + u_i, \quad i = 1, \ldots, n, t = 1, \ldots, T_i, \\ y_{it} &= 1 \quad \text{if } y_{it}^* > 0, \text{ and 0 otherwise,} \end{aligned} \tag{17-39}$$

where, as before (see Sections 11.4 and 11.5), u_i is the unobserved, individual specific heterogeneity. Once again, we distinguish between "random" and "fixed" effects models by the relationship between u_i and $\mathbf{x}_{it}$. The assumption that u_i is unrelated to $\mathbf{x}_{it}$, so that the conditional distribution $f(u_i \mid \mathbf{x}_{it})$ is not dependent on $\mathbf{x}_{it}$, produces the **random effects model.** Note that this places a restriction on the distribution of the heterogeneity.

If that distribution is unrestricted, so that u_i and $\mathbf{x}_{it}$ may be correlated, then we have what is called the **fixed effects model**. The distinction does not relate to any intrinsic characteristic of the effect itself.

As we shall see shortly, this is a modeling framework that is fraught with difficulties and unconventional estimation problems. Among them are the following: Estimation of the random effects model requires very strong assumptions about the heterogeneity;

[27] A "limited information" approach based on the GMM estimation method has been suggested by Avery, Hansen, and Hotz (1983). With recent advances in simulation-based computation of multinormal integrals (see Section 15.6.2.b), some work on such a panel data estimator has appeared in the literature. See, for example, Geweke, Keane, and Runkle (1994, 1997). The GEE estimator of Diggle, Liang, and Zeger (1994) [see also, Liang and Zeger (1986) and Stata (2006)] seems to be another possibility. However, in all these cases, it must be remembered that the procedure specifies estimation of a correlation matrix for a T_i vector of unobserved variables based on a dependent variable that takes only two values. We should not be too optimistic about this if T_i is even moderately large.

the fixed effects model encounters an **incidental parameters problem** that renders the maximum likelihood estimator inconsistent.

17.4.1 THE POOLED ESTIMATOR

To begin, it is useful to consider the pooled estimator that results if we simply ignore the heterogeneity, u_i in (17-39) and fit the model as if the cross-section specification of Section 17.2.2 applies. In this instance, the adage that "ignoring the heterogeneity does not make it go away," applies even more forcefully than in the linear regression case.

If the fixed effects model is appropriate, then all the preceding results for omitted variables, including the Yatchew and Griliches result (1984) apply. The pooled MLE that ignores fixed effects will be inconsistent—possibly wildly so. (Note that since the estimator is ML, not least squares, converting the data to deviations from group means is not a solution—converting the binary dependent variable to deviations will produce a continuous variable with unknown properties.)

The random effects case is more benign. From (17-39), the marginal probability implied by the model is

$$\begin{aligned}\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}) &= \text{Prob}(v_{it} + u_i > -\mathbf{x}_{it}'\boldsymbol{\beta}) \\ &= F\left[\mathbf{x}_{it}'\boldsymbol{\beta}/\left(1+\sigma_u^2\right)^{1/2}\right] \\ &= F(\mathbf{x}_{it}'\boldsymbol{\delta}).\end{aligned}$$

The implication is that based on the marginal distributions, we can consistently estimate $\boldsymbol{\delta}$ (but not $\boldsymbol{\beta}$ or σ_u separately) by pooled MLE. [This result is explored at length in Wooldridge (2002).] This would be a "pseudo MLE" since the log-likelihood function is not the true log-likelihood for the full set of observed data, but it is the correct product of the marginal distributions for $y_{it} \mid \mathbf{x}_{it}$. (This would be the binary choice case counterpart to consistent estimation of $\boldsymbol{\beta}$ in a linear random effects model by pooled ordinary least squares.) The implication, which is absent in the linear case is that ignoring the random effects in a pooled model produces an attenuated (inconsistent—downward biased) estimate of $\boldsymbol{\beta}$; the scale factor that produces $\boldsymbol{\delta}$ is $1/(1+\sigma_u^2)^{1/2}$ which is between zero and one. The implication for the partial effects is less clear. In the model specification, the partial effect is

$$PE(\mathbf{x}_{it}, u_i) = \partial E[y_{it} \mid \mathbf{x}_{it}, u_i]/\partial \mathbf{x}_{it} = \boldsymbol{\beta} \times f(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i),$$

which is not computable. The useful result would be

$$E_u[PE(\mathbf{x}_{it}, u_i)] = \boldsymbol{\beta}\, E_u[\, f(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)].$$

Wooldridge (2002a) shows that the end result, assuming normality of both v_{it} and u_i is $E_u[PE(\mathbf{x}_{it}, u_i)] = \delta\phi(\mathbf{x}_{it}'\boldsymbol{\delta})$. Thus far, surprisingly, it would seem that simply pooling the data and using the simple MLE "works." The estimated standard errors will be incorrect, so a correction such as the cluster estimator shown in Section 14.8.4 would be appropriate. Three considerations suggest that one might want to proceed to the full MLE in spite of these results: (1) The pooled estimator will be inefficient compared to the full MLE; (2) the pooled estimator does not produce an estimator of σ_u which might be of interest in its own right; (3) the FIML estimator is available in contemporary

software and is no more difficult to estimate then the pooled estimator. Note that the pooled estimator is not justified (over the FIML approach) on robustness considerations because the same normality and random effects assumptions that are needed to obtain the FIML estimator will be needed to obtain the preceding results for the pooled estimator.

17.4.2 RANDOM EFFECTS MODELS

A specification that has the same structure as the random effects model of Section 11.5 has been implemented by Butler and Moffitt (1982). We will sketch the derivation to suggest how random effects can be handled in discrete and limited dependent variable models such as this one. Full details on estimation and inference may be found in Butler and Moffitt (1982) and Greene (1995a). We will then examine some extensions of the Butler and Moffitt model.

The random effects model specifies

$$\varepsilon_{it} = v_{it} + u_i,$$

where v_{it} and u_i are independent random variables with

$$E[v_{it} \mid \mathbf{X}] = 0; \text{Cov}[v_{it}, v_{js} \mid \mathbf{X}] = \text{Var}[v_{it} \mid \mathbf{X}] = 1, \quad \text{if } i = j \text{ and } t = s; 0 \text{ otherwise},$$

$$E[u_i \mid \mathbf{X}] = 0; \text{Cov}[u_i, u_j \mid \mathbf{X}] = \text{Var}[u_i \mid \mathbf{X}] = \sigma_u^2, \quad \text{if } i = j; 0 \text{ otherwise},$$

$$\text{Cov}[v_{it}, u_j \mid \mathbf{X}] = 0 \text{ for all } i, t, j,$$

and $\mathbf{X}$ indicates all the exogenous data in the sample, $\mathbf{x}_{it}$ for all i and t.[28] Then,

$$E[\varepsilon_{it} \mid \mathbf{X}] = 0,$$

$$\text{Var}[\varepsilon_{it} \mid \mathbf{X}] = \sigma_v^2 + \sigma_u^2 = 1 + \sigma_u^2,$$

and

$$\text{Corr}[\varepsilon_{it}, \varepsilon_{is} \mid \mathbf{X}] = \rho = \frac{\sigma_u^2}{1 + \sigma_u^2}.$$

The new free parameter is $\sigma_u^2 = \rho/(1 - \rho)$.

Recall that in the cross-section case, the marginal probability associated with an observation is

$$P(y_i \mid \mathbf{x}_i) = \int_{L_i}^{U_i} f(\varepsilon_i) d\varepsilon_i, (L_i, U_i) = (-\infty, -\mathbf{x}_i'\boldsymbol{\beta}) \quad \text{if } y_i = 0 \text{ and } (-\mathbf{x}_i'\boldsymbol{\beta}, +\infty) \quad \text{if } y_i = 1.$$

This simplifies to $\Phi[(2y_i - 1)\mathbf{x}_i'\boldsymbol{\beta}]$ for the normal distribution and $\Lambda[(2y_i - 1)\mathbf{x}_i'\boldsymbol{\beta}]$ for the logit model. In the fully general case with an unrestricted covariance matrix, the contribution of group i to the likelihood would be the joint probability for all T_i observations;

$$L_i = P(y_{i1}, \ldots, y_{iT_i} \mid \mathbf{X}) = \int_{L_{iT_i}}^{U_{iT_i}} \cdots \int_{L_{i1}}^{U_{i1}} f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i}) d\varepsilon_{i1} d\varepsilon_{i2} \ldots d\varepsilon_{iT_i}. \quad \textbf{(17-40)}$$

[28] See Wooldridge (1999) for discussion of this assumption.

The integration of the joint density, as it stands, is impractical in most cases. The special nature of the random effects model allows a simplification, however. We can obtain the joint density of the v_{it}'s by integrating u_i out of the joint density of $(\varepsilon_{i1}, \ldots, \varepsilon_{iT_i}, u_i)$ which is

$$f(\varepsilon_{i1}, \ldots, \varepsilon_{iT_i}, u_i) = f(\varepsilon_{i1}, \ldots, \varepsilon_{iT_i} \mid u_i) f(u_i).$$

So,

$$f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i}) = \int_{-\infty}^{+\infty} f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i} \mid u_i) f(u_i)\, du_i.$$

The advantage of this form is that conditioned on u_i, the ε_{it}'s are independent, so

$$f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i}) = \int_{-\infty}^{+\infty} \prod_{t=1}^{T_i} f(\varepsilon_{it} \mid u_i) f(u_i)\, du_i.$$

Inserting this result in (17-40) produces

$$L_i = P[y_{i1}, \ldots, y_{iT_i} \mid \mathbf{X}] = \int_{L_{iT_i}}^{U_{iT_i}} \cdots \int_{L_{i1}}^{U_{i1}} \int_{-\infty}^{+\infty} \prod_{t=1}^{T_i} f(\varepsilon_{it} \mid u_i) f(u_i)\, du_i\, d\varepsilon_{i1}\, d\varepsilon_{i2} \ldots d\varepsilon_{iT_i}.$$

This may not look like much simplification, but in fact, it is. Because the ranges of integration are independent, we may change the order of integration;

$$L_i = P[y_{i1}, \ldots, y_{iT_i} \mid \mathbf{X}] = \int_{-\infty}^{+\infty} \left[\int_{L_{iT_i}}^{U_{iT_i}} \cdots \int_{L_{i1}}^{U_{i1}} \prod_{t=1}^{T_i} f(\varepsilon_{it} \mid u_i)\, d\varepsilon_{i1}\, d\varepsilon_{i2} \ldots d\varepsilon_{iT_i} \right] f(u_i)\, du_i.$$

Conditioned on the common u_i, the ε's are independent, so the term in square brackets is just the product of the individual probabilities. We can write this as

$$L_i = P[y_{i1}, \ldots, y_{iT_i} \mid \mathbf{X}] = \int_{-\infty}^{+\infty} \left[\prod_{t=1}^{T_i} \left(\int_{L_{it}}^{U_{it}} f(\varepsilon_{it} \mid u_i)\, d\varepsilon_{it} \right) \right] f(u_i)\, du_i. \qquad \textbf{(17-41)}$$

Now, consider the individual densities in the product. Conditioned on u_i, these are the now-familiar probabilities for the individual observations, computed now at $\mathbf{x}_{it}'\boldsymbol{\beta} + u_i$. This produces a general model for random effects for the binary choice model. Collecting all the terms, we have reduced it to

$$L_i = P[y_{i1}, \ldots, y_{iT_i} \mid \mathbf{X}] = \int_{-\infty}^{+\infty} \left[\prod_{t=1}^{T_i} \text{Prob}(Y_{it} = y_{it} \mid \mathbf{x}_{it}'\boldsymbol{\beta} + u_i) \right] f(u_i)\, du_i. \qquad \textbf{(17-42)}$$

It remains to specify the distributions, but the important result thus far is that the entire computation requires only one-dimensional integration. The inner probabilities may be any of the models we have considered so far, such as probit, logit, Gumbel, and so on. The intricate part that remains is to determine how to do the outer integration. **Butler and Moffitt's method** assuming that u_i is normally distributed is detailed in Section 14.9.6.c.

A number of authors have found the Butler and Moffitt formulation to be a satisfactory compromise between a fully unrestricted model and the cross-sectional variant that ignores the correlation altogether. An application that includes both group and time effects is Tauchen, Witte, and Griesinger's (1994) study of arrests and criminal

behavior. The Butler and Moffitt approach has been criticized for the restriction of equal correlation across periods. But it does have a compelling virtue that the model can be efficiently estimated even with fairly large T_i using conventional computational methods. [See Greene (2007b).]

A remaining problem with the Butler and Moffitt specification is its assumption of normality. In general, other distributions are problematic because of the difficulty of finding either a closed form for the integral or a satisfactory method of approximating the integral. An alternative approach that allows some flexibility is the method of **maximum simulated likelihood** (MSL), which was discussed in Section 15.6. The transformed likelihood we derived in (17-42) is an expectation:

$$L_i = \int_{-\infty}^{+\infty} \left[\prod_{t=1}^{T_i} \text{Prob}(Y_{it} = y_{it} \,|\, \mathbf{x}_{it}'\boldsymbol{\beta} + u_i) \right] f(u_i)\, du_i$$

$$= E_{u_i} \left[\prod_{t=1}^{T_i} \text{Prob}(Y_{it} = y_{it} \,|\, \mathbf{x}_{it}'\boldsymbol{\beta} + u_i) \right].$$

This expectation can be approximated by simulation rather than **quadrature.** First, let θ now denote the scale parameter in the distribution of u_i. This would be σ_u for a normal distribution, for example, or some other scaling for the logistic or uniform distribution. Then, write the term in the likelihood function as

$$L_i = E_{u_i} \left[\prod_{t=1}^{T_i} F(y_{it}, \mathbf{x}_{it}'\boldsymbol{\beta} + \theta u_i) \right] = E_{u_i}[h(u_i)].$$

The function is smooth, continuous, and continuously differentiable. If this expectation is finite, then the conditions of the law of large numbers should apply, which would mean that for a sample of observations $u_{i1}, \ldots, u_{iR}$,

$$\text{plim}\, \frac{1}{R} \sum_{r=1}^{R} h(u_{ir}) = E_u[h(u_i)].$$

This suggests, based on the results in Chapter 15, an alternative method of maximizing the log-likelihood for the random effects model. A sample of person-specific draws from the population u_i can be generated with a random number generator. For the Butler and Moffitt model with normally distributed u_i, the simulated log-likelihood function is

$$\ln L_{Simulated} = \sum_{i=1}^{n} \ln \left\{ \frac{1}{R} \sum_{r=1}^{R} \left[\prod_{t=1}^{T_i} F\left[(2y_{ik} - 1)(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma_u u_{ir}) \right] \right] \right\}. \qquad \textbf{(17-43)}$$

This function is maximized with respect $\boldsymbol{\beta}$ and σ_u. Note that in the preceding, as in the quadrature approximated log-likelihood, the model can be based on a probit, logit, or any other functional form desired.

We have examined two approaches to estimation of a probit model with random effects. GMM estimation is another possibility. Avery, Hansen, and Hotz (1983), Bertschek and Lechner (1998), and Inkmann (2000) examine this approach; the latter two offer some comparison with the quadrature and simulation-based estimators considered here. (Our application in Example 17.23 will use the Bertschek and Lechner data.)

17.4.3 FIXED EFFECTS MODELS

The fixed effects model is

$$\begin{aligned} &y_{it}^* = \alpha_i d_{it} + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \ldots, n,\ t = 1, \ldots, T_i, \\ &y_{it} = 1 \quad \text{if } y_{it}^* > 0, \text{ and 0 otherwise,} \end{aligned} \tag{17-44}$$

where d_{it} is a dummy variable that takes the value one for individual i and zero otherwise. For convenience, we have redefined $\mathbf{x}_{it}$ to be the nonconstant variables in the model. The parameters to be estimated are the K elements of $\boldsymbol{\beta}$ and the n individual constant terms. Before we consider the several virtues and shortcomings of this model, we consider the practical aspects of estimation of what are possibly a huge number of parameters, $(n + K) - n$ is not limited here, and could be in the thousands in a typical application. The log-likelihood function for the fixed effects model is

$$\ln L = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \ln P(y_{it} \mid \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}), \tag{17-45}$$

where $P(.)$ is the probability of the observed outcome, for example, $\Phi[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]$ for the probit model or $\Lambda[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]$ for the logit model, where $q_{it} = 2y_{it} - 1$. What follows can be extended to any index function model, but for the present, we'll confine our attention to symmetric distributions such as the normal and logistic, so that the probability can be conveniently written as $\text{Prob}(Y_{it} = y_{it} \mid \mathbf{x}_{it}) = P[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]$. It will be convenient to let $z_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}$ so $\text{Prob}(Y_{it} = y_{it} \mid \mathbf{x}_{it}) = P(q_{it} z_{it})$.

In our previous application of this model, in the linear regression case, we found that estimation of the parameters was made possible by a transformation of the data to deviations from group means which eliminated the person specific constants from the estimator. (See Section 11.4.1.) Save for the special case discussed later, that will not be possible here, so that if one desires to estimate the parameters of this model, it will be necessary actually to compute the possibly huge number of constant terms at the same time. This has been widely viewed as a practical obstacle to estimation of this model because of the need to invert a potentially large second derivatives matrix, but this is a misconception. [See, e.g., Maddala (1987), p. 317.] The method for estimation of nonlinear fixed effects models such as the probit and logit models is detailed in Section 14.9.6.d.

The problems with the fixed effects estimator are statistical, not practical. The estimator relies on T_i increasing for the constant terms to be consistent—in essence, each α_i is estimated with T_i observations. But, in this setting, not only is T_i fixed, it is likely to be quite small. As such, the estimators of the constant terms are not consistent (not because they converge to something other than what they are trying to estimate, but because they do not converge at all). The estimator of $\boldsymbol{\beta}$ is a function of the estimators of α, which means that the MLE of $\boldsymbol{\beta}$ is not consistent either. This is the incidental parameters problem. [See Neyman and Scott (1948) and Lancaster (2000).] There is, as well, a small sample (small T_i) bias in the estimators. How serious this bias is remains a question in the literature. Two pieces of received wisdom are Hsiao's (1986) results for a binary logit model [with additional results in Abrevaya (1997)] and Heckman and MaCurdy's (1980) results for the probit model. Hsiao found that for $T_i = 2$, the bias in the MLE of $\boldsymbol{\beta}$ is 100 percent, which is extremely pessimistic. Heckman and MaCurdy

found in a Monte Carlo study that in samples of $n = 100$ and $T = 8$, the bias appeared to be on the order of 10 percent, which is substantive, but certainly less severe than Hsiao's results suggest. No other theoretical results have been shown for other models, although in *very* few cases, it can be shown that there is no incidental parameters problem. (The Poisson model mentioned in Chapter 14 is one of these special cases.) The fixed effects approach does have some appeal in that it does not require an assumption of orthogonality of the independent variables and the heterogeneity. An ongoing pursuit in the literature is concerned with the severity of the tradeoff of this virtue against the incidental parameters problem. Some commentary on this issue appears in Arellano (2001). Results of our own investigation appear in Section 15.5.2 and Greene (2004).

17.4.4 A CONDITIONAL FIXED EFFECTS ESTIMATOR

Why does the incidental parameters problem arise here and not in the linear regression model?[29] Recall that estimation in the regression model was based on the deviations from group means, not the original data as it is here. The result we exploited there was that although $f(y_{it} \mid \mathbf{X}_i)$ is a function of α_i, $f(y_{it} \mid \mathbf{X}_i, \bar{y}_i)$ is not a function of α_i, and we used the latter in estimation of $\boldsymbol{\beta}$. In that setting, $\bar{y}_i$ is a **minimal sufficient statistic** for α_i. Sufficient statistics are available for a few distributions that we will examine, but not for the probit model. They are available for the logit model, as we now examine.

A fixed effects binary logit model is

$$\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}) = \frac{e^{\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}}}{1 + e^{\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}}}.$$

The unconditional likelihood for the nT independent observations is

$$L = \prod_i \prod_t (F_{it})^{y_{it}} (1 - F_{it})^{1 - y_{it}}.$$

Chamberlain (1980) [following Rasch (1960) and Andersen (1970)] observed that the **conditional likelihood function**,

$$L^c = \prod_{i=1}^{n} \text{Prob}\left(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iT_i} = y_{iT_i} \,\middle|\, \sum_{t=1}^{T_i} y_{it} \right),$$

is free of the incidental parameters, α_i. The joint likelihood for each set of T_i observations conditioned on the number of ones in the set is

$$\text{Prob}\left(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iT_i} = y_{iT_i} \,\middle|\, \sum_{t=1}^{T_i} y_{it}, \text{data} \right)$$
$$= \frac{\exp\left(\sum_{t=1}^{T_i} y_{it} \mathbf{x}_{it}'\boldsymbol{\beta} \right)}{\sum_{\Sigma_t d_{it} = S_i} \exp\left(\sum_{t=1}^{T_i} d_{it} \mathbf{x}_{it}'\boldsymbol{\beta} \right)}. \qquad \textbf{(17-46)}$$

[29] The incidental parameters problem does show up in ML estimation of the FE linear model, where Neyman and Scott (1948) discovered it, in estimation of σ_ε^2. The MLE of σ_ε^2 is $\boldsymbol{e}'\boldsymbol{e}/nT$, which converges to $[(T - 1)/T]\sigma_\varepsilon^2 < \sigma_\varepsilon^2$.

The function in the denominator is summed over the set of all $\binom{T_i}{S_i}$ different sequences of T_i zeros and ones that have the same sum as $S_i = \sum_{t=1}^{T_i} y_{it}$.[30]

Consider the example of $T_i = 2$. The unconditional likelihood is

$$L = \prod_i \text{Prob}(Y_{i1} = y_{i1})\text{Prob}(Y_{i2} = y_{i2}).$$

For each pair of observations, we have these possibilities:

1. $y_{i1} = 0$ and $y_{i2} = 0$. $\text{Prob}(0, 0 \mid \text{sum} = 0) = 1$.
2. $y_{i1} = 1$ and $y_{i2} = 1$. $\text{Prob}(1, 1 \mid \text{sum} = 2) = 1$.

The ith term in L^c for either of these is just one, so they contribute nothing to the conditional likelihood function.[31] When we take logs, these terms (and these observations) will drop out. But suppose that $y_{i1} = 0$ and $y_{i2} = 1$. Then

$$\textbf{3.}\quad \text{Prob}(0, 1 \mid \text{sum} = 1) = \frac{\text{Prob}(0, 1 \textit{ and } \text{sum} = 1)}{\text{Prob}(\text{sum} = 1)} = \frac{\text{Prob}(0, 1)}{\text{Prob}(0, 1) + \text{Prob}(1, 0)}.$$

Therefore, for this pair of observations, the conditional probability is

$$\frac{\dfrac{1}{1+e^{\alpha_i+\mathbf{x}_{i1}'\boldsymbol{\beta}}}\dfrac{e^{\alpha_i+\mathbf{x}_{i2}'\boldsymbol{\beta}}}{1+e^{\alpha_i+\mathbf{x}_{i2}'\boldsymbol{\beta}}}}{\dfrac{1}{1+e^{\alpha_i+\mathbf{x}_{i1}'\boldsymbol{\beta}}}\dfrac{e^{\alpha_i+\mathbf{x}_{i2}'\boldsymbol{\beta}}}{1+e^{\alpha_i+\mathbf{x}_{i2}'\boldsymbol{\beta}}} + \dfrac{e^{\alpha_i+\mathbf{x}_{i1}'\boldsymbol{\beta}}}{1+e^{\alpha_i+\mathbf{x}_{i1}'\boldsymbol{\beta}}}\dfrac{1}{1+e^{\alpha_i+\mathbf{x}_{i2}'\boldsymbol{\beta}}}} = \frac{e^{\mathbf{x}_{i2}'\boldsymbol{\beta}}}{e^{\mathbf{x}_{i1}'\boldsymbol{\beta}} + e^{\mathbf{x}_{i2}'\boldsymbol{\beta}}}.$$

By conditioning on the sum of the two observations, we have removed the heterogeneity. Therefore, we can construct the conditional likelihood function as the product of these terms for the pairs of observations for which the two observations are (0, 1). Pairs of observations with one and zero are included analogously. The product of the terms such as the preceding, for those observation sets for which the sum is not zero or T_i, constitutes the conditional likelihood. Maximization of the resulting function is straightforward and may be done by conventional methods.

As in the linear regression model, it is of some interest to test whether there is indeed heterogeneity. With homogeneity ($\alpha_i = \alpha$), there is no unusual problem, and the model can be estimated, as usual, as a logit model. It is not possible to test the hypothesis using the likelihood ratio test, however, because the two likelihoods are not comparable. (The conditional likelihood is based on a restricted data set.) None of the usual tests of restrictions can be used because the individual effects are never actually estimated.[32] Hausman's (1978) specification test is a natural one to use here,

[30] The enumeration of all these computations stands to be quite a burden—see Arellano (2000, p. 47) or Baltagi (2005, p. 235). In fact, using a recursion suggested by Krailo and Pike (1984), the computation even with T_i up to 100 is routine.

[31] Recall that in the probit model when we encountered this situation, the individual constant term could not be estimated and the group was removed from the sample. The same effect is at work here.

[32] This produces a difficulty for this estimator that is shared by the semiparametric estimators discussed in the next section. Because the fixed effects are not estimated, it is not possible to compute probabilities or marginal effects with these estimated coefficients, and it is a bit ambiguous what one can do with the results of the computations. The brute force estimator that actually computes the individual effects might be preferable.

however. Under the null hypothesis of homogeneity, both Chamberlain's conditional maximum likelihood estimator (CMLE) and the usual maximum likelihood estimator are consistent, but Chamberlain's is inefficient. (It fails to use the information that $\alpha_i = \alpha$, and it may not use all the data.) Under the alternative hypothesis, the unconditional maximum likelihood estimator is inconsistent,[33] whereas Chamberlain's estimator is consistent and efficient. The Hausman test can be based on the chi-squared statistic

$$\chi^2 = (\hat{\boldsymbol{\beta}}_{\text{CML}} - \hat{\boldsymbol{\beta}}_{\text{ML}})'(\text{Var[CML]} - \text{Var[ML]})^{-1}(\hat{\boldsymbol{\beta}}_{\text{CML}} - \hat{\boldsymbol{\beta}}_{\text{ML}}). \qquad \textbf{(17-47)}$$

The estimated covariance matrices are those computed for the two maximum likelihood estimators. For the unconditional maximum likelihood estimator, the row and column corresponding to the constant term are dropped. A large value will cast doubt on the hypothesis of homogeneity. (There are K degrees of freedom for the test.) It is possible that the covariance matrix for the maximum likelihood estimator will be larger than that for the conditional maximum likelihood estimator. If so, then the difference matrix in brackets is assumed to be a zero matrix, and the chi-squared statistic is therefore zero.

Example 17.11 Binary Choice Models for Panel Data

In Example 17.3, we fit a pooled binary logit model $y = 1(DocVis > 0)$ using the German health care utilization data examined in appendix Table F7.1. The model is

$$\text{Prob}(DocVis_{it} > 0) = \Lambda(\beta_1 + \beta_2 Age_{it} + \beta_3 Income_{it} + \beta_4 Kids_{it} + \beta_5 Education_{it} + \beta_6 Married_{it}).$$

No account of the panel nature of the data set was taken in that exercise. The sample contains a total of 27,326 observations on 7,293 families with T_i dispersed from one to seven. Table 17.8 lists estimates of parameter estimates and estimated standard errors for probit and logit random and fixed effects models. There is a surprising amount of variation across the estimators. The coefficients are in bold to facilitate reading the table. It is generally difficult to compare across the estimators. The three estimators would be expected to produce very different estimates in any of the three specifications—recall, for example, the pooled estimator is inconsistent in either the fixed or random effects cases. The logit results include two fixed effects estimators. The line market "U" is the unconditional (inconsistent) estimator. The one marked "C" is Chamberlain's consistent estimator. Note for all three fixed effects estimators, it is necessary to drop from the sample any groups that have $DocVis_{it}$ equal to zero or one for every period. There were 3,046 such groups, which is about 42 percent of the sample. We also computed the probit random effects model in two ways, first by using the Butler and Moffitt method, then by using maximum simulated likelihood estimation. In this case, the estimators are very similar, as might be expected. The estimated correlation coefficient, ρ, is computed as $\sigma_u^2/(\sigma_\varepsilon^2 + \sigma_u^2)$. For the probit model, $\sigma_\varepsilon^2 = 1$. The MSL estimator computes $s_u = 0.9088376$, from which we obtained ρ. The estimated partial effects for the models are shown in Table 17.9. The average of the fixed effects constant terms is used to obtain a constant term for the fixed effects case. Once again there is a considerable amount of variation across the different estimators. On average, the fixed effects models tend to produce much larger values than the pooled or random effects models.

[33] Hsiao (2003) derives the result explicitly for some particular cases.

TABLE 17.8 Estimated Parameters for Panel Data Binary Choice Models

			Variable					
Model	***Estimate***	***ln L***	***Constant***	***Age***	***Income***	***Kids***	***Education***	***Married***
Logit	β		**0.25112**	**0.020709**	**−0.18592**	**−0.22947**	**−0.045587**	**0.085293**
Pooled	St.Err.	−17673.10	0.091135	0.0012852	0.075064	0.029537	0.005646	0.033286
	Rob.SE[e]		0.12827	0.0017429	0.091546	0.038313	0.008075	0.045314
Logit R.E.	β	−15261.90	**−0.13460**	**0.039267**	**0.021914**	**−0.21598**	**−0.063578**	**0.025071**
$\rho = 0.41607$	St.Err.		0.17764	0.0024659	0.11866	0.047738	0.011322	0.056282
Logit	β	−9458.64		**0.10475**	**−0.060973**	**−0.088407**	**−0.11671**	**−0.057318**
F.E.(U)[a]	St.Err.			0.0072548	0.17829	0.074399	0.066749	0.10609
Logit	β	−6312.57		**.08384**	**−.06521**	**−.07802**	**−.12179**	**−.04897**
F.E.(C)[b]	St.Err.			(.006382)	(.15793)	(.066186)	(.05466)	(.092639)
Probit	β	−17670.94	**0.15500**	**0.012835**	**−0.11643**	**−0.14118**	**−0.028115**	**0.052260**
Pooled	St.Err.		0.056516	0.0007903	0.046329	0.018218	0.003503	0.020462
	Rob.SE[e]		0.079591	0.0010739	0.056543	0.023614	0.005014	0.027904
Probit:RE[c]	β	−16273.96	**0.034113**	**0.020143**	**−0.003176**	**−0.15379**	**−0.033694**	**0.016325**
$\rho = 0.44789$	St.Err.		0.096354	0.0013189	0.066672	0.027043	0.006289	0.031347
Probit:RE[d]	β	−16279.97	**0.033290**	**0.020078**	**−0.002973**	**−0.153579**	**−0.033489**	**0.016826**
$\rho = 0.44799$	St.Err.		0.063229	0.0009013	0.052012	0.020286	0.003931	0.022771
Probit	β	−9453.71		**0.062528**	**−0.034328**	**−0.048270**	**−0.072189**	**−0.032774**
F.E.(U)	St.Err.			0.0043219	0.10745	0.044559	0.040731	0.063627

[a]Unconditional fixed effects estimator
[b]Conditional fixed effects estimator
[c]Butler and Moffitt estimator
[d]Maximum simulated likelihood estimator
[e]Robust, "cluster" corrected standard error

TABLE 17.9 Estimated Partial Effects for Panel Data Binary Choice Models

Model	*Age*	*Income*	*Kids*	*Education*	*Married*
Logit, P[a]	0.0048133	−0.043213	−0.053598	−0.010596	0.019936
Logit: RE,Q[b]	0.0064213	0.0035835	−0.035448	−0.010397	0.0041049
Logit: F,U[c]	0.024871	−0.014477	−0.020991	−0.027711	−0.013609
Logit: F,C[d]	0.0072991	−0.0043387	−0.0066967	−0.0078206	−0.0044842
Probit, P[a]	0.0048374	−0.043883	−0.053414	−0.010597	0.019783
Probit RE.Q[b]	0.0056049	−0.0008836	−0.042792	−0.0093756	0.0045426
Probit:RE,S[e]	0.0071455	−0.0010582	−0.054655	−0.011917	0.0059878
Probit: F,U[c]	0.023958	−0.013152	−0.018495	−0.027659	−0.012557

[a]Pooled estimator
[b]Butler and Moffitt estimator
[c]Unconditional fixed effects estimator
[d]Conditional fixed effects estimator
[e]Maximum simulated likelihood estimator

Example 17.12 Fixed Effects Logit Models: Magazine Prices Revisited

The fixed effects model does have some appeal, but the incidental parameters problem is a significant shortcoming of the unconditional probit and logit estimators. The conditional MLE for the fixed effects logit model is a fairly common approach. A widely cited application of the model is Cecchetti's (1986) analysis of changes in newsstand prices of magazines. Cecchetti's model was

$$\text{Prob}(\textit{Price change in year i of magazine t}) = \Lambda(\alpha_j + \mathbf{x}_{it}'\boldsymbol{\beta}),$$

where the variables in $\mathbf{x}_{it}$ are (1) time since last price change, (2) inflation since last change, (3) previous fixed price change, (4) current inflation, (5) industry sales growth, and (6) sales volatility. The fixed effect in the model is indexed "*j*" rather than "*i*" as it is defined as a three-year interval for magazine *i*. Thus, a magazine that had been on the newstands for nine years would have three constants, not just one. In addition to estimating several specifications of the price change model, Cecchetti used the Hausman test in (17-47) to test for the existence of the common effects. Some of Cecchetti's results appear in Table 17.10.

Willis (2006) argued that Cecchetti's estimates were inconsistent and the Hausman test is invalid because right-hand-side variables (1), (2), and (6) are all functions of lagged dependent variables. This state dependence invalidates the use of the sum of the observations for the group as a sufficient statistic in the Chamberlain estimator and the Hausman tests. He proposes, instead, a method suggested by Heckman and Singer (1984b) to incorporate the unobserved heterogeneity in the *unconditional* likelihood function. The Heckman and Singer model can be formulated as a latent class model (see Sections 14.10 and 17.4.7) in which the classes are defined by different constant terms—the remaining parameters in the model

TABLE 17.10 Models for Magazine Price Changes (standard errors in parentheses)

	Pooled	*Unconditional FE*	*Conditional FE Cecchetti*	*Conditional FE Willis*	*Heckman and Singer*
β_1	−1.10 (0.03)	−0.07 (0.03)	1.12 (3.66)	1.02 (0.28)	−0.09 (0.04)
β_2	6.93 (1.12)	8.83 (1.25)	11.57 (1.68)	19.20 (7.51)	8.23 (1.53)
β_5	−0.36 (0.98)	−1.14 (1.06)	5.85 (1.76)	7.60 (3.46)	−0.13 (1.14)
Constant 1	−1.90 (0.14)				−1.94 (0.20)
Constant 2					−29.15 (1.1e11)
ln *L*	−500.45	−473.18	−82.91	−83.72	−499.65
Sample size	1026	1026		543	1026

are constrained to be equal across classes. Willis fit the Heckman and Singer model with two classes to a restricted version of Cecchetti's model using variables (1), (2), and (5). The results in Table 17.10 show some of the results from Willis's Table I. (Willis reports that he could not reproduce Cecchetti's results—the ones in Cecchetti's second column would be the counterparts—because of some missing values. In fact, Willis's estimates are quite far from Cecchetti's results, so it will be difficult to compare them. Both are reported here.)

The two "mass points" reported by Willis are shown in Table 17.10. He reports that these two values (-1.94 and -29.15) correspond to class probabilities of 0.88 and 0.12, though it is difficult to make the translation based on the reported values. He does note that the change in the log-likelihood in going from one mass point (pooled logit model) to two is marginal, only from -500.45 to -499.65. There is another anomaly in the results that is consistent with this finding. The reported standard error for the second "mass point" is 1.1×10^{11}, or essentially $+\infty$. The finding is consistent with overfitting the latent class model. The results suggest that the better model is a one-class (pooled) model.

17.4.5 MUNDLAK'S APPROACH, VARIABLE ADDITION AND BIAS REDUCTION

Thus far, both the fixed effects (FE) and the random effects (RE) specifications present problems for modeling binary choice with panel data. The MLE of the FE model is inconsistent even when the model is properly specified—this is the incidental parameters problem. (And, like the linear model, the FE probit and logit models do not allow time-invariant regressors.) The random effects specification requires a strong, often unreasonable, assumption that the effects and the regressors are uncorrelated. Of the two, the FE model is the more appealing, though with modern longitudinal data sets with many demographics, the problem of time-invariant variables would seem to be compelling. This would seem to recommend the conditional estimator in Section 17.4.4, save for yet another complication. With no estimates of the constant terms, neither probabilities nor partial effects can be computed with the results. We are left making inferences about ratios of coefficient. Two approaches have been suggested for finding a middle ground: Mundlak's (1978) approach that involves projecting the effects on the group means of the time-varying variables and recent developments such as Fernandez-Val's approach that involves correcting the bias in the FE MLE.

The Mundlak (1978) [and Chamberlain (1984) and Wooldridge, e.g., (2002a)] approach augments (17-44) as follows:

$$
\begin{aligned}
y_{it}^{*} &= \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \\
\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}) &= F(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}) \\
\alpha_i &= \alpha + \bar{\mathbf{x}}_i'\boldsymbol{\delta} + u_i,
\end{aligned}
$$

where we have used $\bar{\mathbf{x}}_i$ generically for the group means of the time varying variables in $\mathbf{x}_{it}$. The reduced form of the model is

$$
\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}) = F(\alpha + \bar{\mathbf{x}}_i'\boldsymbol{\delta} + \mathbf{x}_{it}'\boldsymbol{\beta} + u_i).
$$

(Wooldridge and Chamberlain also suggest using all years of $\mathbf{x}_{it}$ rather than the group means. This raises a problem in unbalanced panels, however. We will ignore this possibility.) The projection of α_i on $\bar{\mathbf{x}}_i$ produces a random effects formulation. As in the linear model (see Section 11.5.6), it also suggests a means of testing for fixed vs. random effects. Since $\boldsymbol{\delta} = \mathbf{0}$ produces the pure random effects model, a joint Wald test of the null hypothesis that $\boldsymbol{\delta}$ equals zero can be used.

TABLE 17.11 Estimated Random Effects Models

	Constant	*Age*	*Income*	*Kids*	*Education*	*Married*
Random	0.03411	0.02014	−0.00318	−0.15379	−0.03369	0.01633
Effects	(0.09635)	(0.00132)	(0.06667)	(0.02704)	(0.00629)	(0.03135)
Augmented	0.37485	0.05035	−0.03057	−0.04202	−0.05449	−0.02645
Model	(0.10501)	(0.00357)	(0.09318)	(0.03751)	(0.03307)	(0.05180)
		−0.03659	−0.35065	−0.22509	0.02387	0.14668
Means		(0.00384)	(0.13984)	(0.05499)	(0.03374)	(0.06607)

Example 17.13 Panel Data Random Effects Estimators

Example 17.11 presents several estimators of panel data estimators for the probit and logit models. Pooled, random effects and fixed effects estimates are given for the probit model

$$\text{Prob}(DocVis_{it} > 0) = \Phi(\beta_1 + \beta_2 Age_{it} + \beta_3 Income_{it} + \beta_4 Kids_{it} + \beta_5 Education_{it} + \beta_6 Married_{it}).$$

We continue that analysis here by considering Mundlak's approach to the common effects model. Table 17.11 presents the random effects model from earlier, and the augmented estimator that contains the group means of the variables, all of which are time varying. The addition of the group means to the regression brings large changes to the estimates of the parameters, which might suggest the appropriateness of the fixed effects model. A formal test is carried by computing a Wald statistic for the null hypothesis that the last five coefficients in the augmented model equal zero. The chi-squared statistic equals 113.282 with five degrees of freedom. The critical value from the chi-squared table for 95 percent significance is 11.07, so the hypothesis that δ equals zero, that is, the hypothesis of the random effects model (restrictions), is rejected. The two log likelihoods are −16273.96 for the REM and −16222.06 for the augmented REM. The LR statistic would be twice the difference, or 103.8. This produces the same conclusion. The FEM appears to be the preferred model.

A series of recent studies has sought to maintain the fixed effects specification while correcting the bias due to the incidental parameters problem. There are two broad approaches. Hahn and Kuersteiner (2004), Hahn and Newey (2005), and Fernandez-Val (2009) have developed an approximate, "large T" result for plim$(\hat{\boldsymbol{\beta}}_{FEMALE} - \boldsymbol{\beta})$ that produces a direct correction to the estimator, itself. Fernandez-Val (2009) develops corrections for the estimated constant terms as well. Arellano and Hahn (2006, 2007) propose a modification of the log-likelihood function with, in turn, different first-order estimation equations, that produces an approximately unbiased estimator of $\boldsymbol{\beta}$. In a similar fashion to the second of these approaches, Carro (2007) modifies the first-order conditions (estimating equations) from the original log-likelihood function, once again to produce an approximately unbiased estimator of $\boldsymbol{\beta}$. (In general, given the overall approach of using a large T approximation, the payoff to these estimators is to reduce the bias of the FEMALE from O$(1/T)$ to O$(1/T^2)$, which is a considerable reduction.) These estimators are not yet in widespread use. The received evidence suggests that in the simple case we are considering here, the incidental parameters problem is a secondary concern when T reaches say 10 or so. For some modern public use data sets, such as the BHPS or GSOEP which are beyond their 15th wave, the incidental parameters problem may not be too severe. However, most of the studies mentioned above are concerned with dynamic models (see Section 17.4.6), where the problem is possible more severe than in the static case. Research in this area is ongoing.

17.4.6 DYNAMIC BINARY CHOICE MODELS

A random or fixed effects model that explicitly allows for lagged effects would be

$$y_{it} = \mathbf{1}(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \gamma y_{i,t-1} + \varepsilon_{it} > 0).$$

Lagged effects, or **persistence**, in a binary choice setting can arise from three sources, serial correlation in ε_{it}, the **heterogeneity**, α_i, or true **state dependence** through the term $\gamma y_{i,t-1}$. Chiappori (1998) [and see Arellano (2001)] suggests an application to the French automobile insurance market in which the incentives built into the pricing system are such that having an accident in one period should lower the probability of having one in the next (state dependence), but, some drivers remain more likely to have accidents than others in every period, which would reflect the heterogeneity instead. State dependence is likely to be particularly important in the typical panel which has only a few observations for each individual. Heckman (1981a) examined this issue at length. Among his findings were that the somewhat muted small sample bias in fixed effects models with $T = 8$ was made much worse when there was state dependence. A related problem is that with a relatively short panel, the **initial conditions**, y_{i0}, have a crucial impact on the entire path of outcomes. Modeling dynamic effects and initial conditions in binary choice models is more complex than in the linear model, and by comparison there are relatively fewer firm results in the applied literature.[34]

The correlation between α_i and $y_{i,t-1}$ in the dynamic binary choice model makes $y_{i,t-1}$ endogenous. Thus, the estimators we have examined thus far will not be consistent. Two familiar alternative approaches that have appeared in recent applications are due to Heckman (1981) and Wooldridge (2005), both of which build on the random effects specification. Heckman's approach provides a separate equation for the initial condition,

$$\text{Prob}(y_{i1} = 1 \,|\, \mathbf{x}_{i1}, \mathbf{z}_i, \alpha_i) = \Phi(\mathbf{x}_{i1}'\boldsymbol{\delta} + \mathbf{z}_i'\boldsymbol{\tau} + \theta\alpha_i)$$

$$\text{Prob}(y_{it} = 1 \,|\, \mathbf{x}_{it}, y_{i,t-1}, \alpha_i) = \Phi(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \alpha_i), t = 2, \ldots, T_i,$$

where $\mathbf{z}_i$ is a set of "instruments" observed at the first period that are not contained in $\mathbf{x}_{it}$. The conditional log-likelihood is

$$\ln L \,|\, \alpha = \sum_{i=1}^{n} \ln \left\{ \Phi\left[(2y_{i1} - 1)(\mathbf{x}_{i1}'\boldsymbol{\delta} + \mathbf{z}_i'\boldsymbol{\tau} + \theta\alpha_i)\right] \prod_{t=2}^{T_i} \Phi\left[(2y_{it} - 1)(\mathbf{x}_{i1}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \alpha_i)\right] \right\}$$

$$= \sum_{i=1}^{n} \ln L_i \,|\, \alpha_i.$$

We now adopt the random effects approach and further assume that α_i is normally distributed with mean zero and variance σ_α^2. The random effects log-likelihood function can be maximized with respect to $(\boldsymbol{\delta}, \boldsymbol{\tau}, \theta, \boldsymbol{\beta}, \gamma, \sigma_\alpha)$ using either the Butler and Moffitt

[34] A survey of some of these results is given by Hsiao (2003). Most of Hsiao (2003) is devoted to the linear regression model. A number of studies specifically focused on discrete choice models and panel data have appeared recently, including Beck, Epstein, Jackman and O'Halloran (2001), Arellano (2001) and Greene (2001). Vella and Verbeek (1998) provide an application to the joint determination of wages and union membership. Other important references are aguirregabiria and Mira (2010), Carro (2007), and Fernandeg–Val (2009). Stewart (2006) and Arulampalam and Stewart (2007) provide several results for practitioners.

quadrature method or the maximum simulated likelihood method described in Section 17.4.2. Stewart and Arulampalam (2007) suggest a useful shortcut for formulating the Heckman model. Let $D_{it} = 1$ in period 1 and 0 in every other period and let $C_{it} = 1 - D_{it}$. Then, the two parts may be combined in

$$\ln L \mid \boldsymbol{\alpha} = \sum_{i=1}^{n} \ln \prod_{t=1}^{T_i} \left\{ \Phi \left[(2y_{it} - 1) \left\langle C_{it}(\mathbf{x}_{i1}'\boldsymbol{\beta} + \gamma y_{i,t-1}) + D_{it}(\mathbf{x}_{it}'\boldsymbol{\delta} + \mathbf{z}_i'\boldsymbol{\tau}) + (1 + \lambda D_{it})\alpha_i \right\rangle \right] \right\}.$$

In this form, the model can be viewed as a random parameters (random constant term) model in which there is heteroscedasticity in the random part of the constant term.

Wooldridge's approach builds on the Mundlak device of the previous section. Starting from the same point, he suggests a model for the random effect conditioned on the initial value. Thus,

$$\alpha_i \mid y_{i1}, \mathbf{z}_i \sim N\left[\alpha_0 + \eta y_{i1} + \mathbf{z}_i'\boldsymbol{\tau}, \sigma_\alpha^2\right].$$

Assembling the parts, Wooldridge's model is a bit simpler than Heckman's;

$$\begin{aligned} &\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}, y_{i1}, u_i) \\ &\quad = \Phi[(2y_{it} - 1)(\alpha_0 + \mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \eta y_{i1} + \mathbf{z}_i'\boldsymbol{\tau} + u_i)], t = 2, \ldots, T_i. \end{aligned}$$

Much of the contemporary literature has focused on methods of avoiding the strong parametric assumptions of the probit and logit models. Manski (1987) and Honore and Kyriazidou (2000) show that Manski's (1986) maximum score estimator can be applied to the differences of unequal pairs of observations in a two-period panel with fixed effects. However, the limitations of the maximum score estimator have motivated research on other approaches. An extension of lagged effects to a parametric model is Chamberlain (1985), Jones and Landwehr (1988), and Magnac (1997), who added state dependence to Chamberlain's fixed effects logit estimator. Unfortunately, once the identification issues are settled, the model is only operational if there are no other exogenous variables in it, which limits its usefulness for practical application. Lewbel (2000) has extended his fixed effects estimator to dynamic models as well.

Dong and Lewbel (2010) have extended Lewbel's "special regressor" method to dynamic binary choice models and have devised an estimator based on an IV linear regression. Honore and Kyriazidou (2000) have combined the logic of the **conditional logit model** and Manski's maximum score estimator. They specify

$$\begin{aligned} &\text{Prob}(y_{i0} = 1 \mid \mathbf{x}_i, \alpha_i) = p_0(\mathbf{x}_i, \alpha_i) \quad \text{where } \mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}), \\ &\text{Prob}(y_{it} = 1 \mid \mathbf{x}_i, \alpha_i, y_{i0}, y_{i1}, \ldots, y_{i,t-1}) = F(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \gamma y_{i,t-1}) \quad t = 1, \ldots, T. \end{aligned}$$

The analysis assumes a single regressor and focuses on the case of $T = 3$. The resulting estimator resembles Chamberlain's but relies on observations for which $\mathbf{x}_{it} = \mathbf{x}_{i,t-1}$, which rules out direct time effects as well as, for practical purposes, any continuous variable. The restriction to a single regressor limits the generality of the technique as well. The need for observations with equal values of $\mathbf{x}_{it}$ is a considerable restriction, and the authors propose a kernel density estimator for the difference, $\mathbf{x}_{it} - \mathbf{x}_{i,t-1}$, instead which does relax that restriction a bit. The end result is an estimator that converges (they conjecture) but to a nonnormal distribution and at a rate slower than $n^{-1/3}$.

Semiparametric estimators for dynamic models at this point in the development are still primarily of theoretical interest. Models that extend the parametric formulations to

include state dependence have a much longer history, including Heckman (1978, 1981a, 1981b), Heckman and MaCurdy (1980), Jakubson (1988), Keane (1993), and Beck et al. (2001) to name a few.[35] In general, even without heterogeneity, dynamic models ultimately involve modeling the joint outcome $(y_{i0}, \ldots, y_{iT})$, which necessitates some treatment involving multivariate integration. Example 17.14 describes an application. Stewart (2006) provides another.

Example 17.14 An Intertemporal Labor Force Participation Equation

Hyslop (1999) presents a model of the labor force participation of married women. The focus of the study is the high degree of persistence in the participation decision. Data used in the study were the years 1979–1985 of the Panel Study of Income Dynamics. A sample of 1,812 continuously married couples were studied. Exogenous variables that appeared in the model were measures of permanent and transitory income and fertility captured in yearly counts of the number of children from 0–2, 3–5, and 6–17 years old. Hyslop's formulation, in general terms, is

$$\text{(initial condition) } y_{i0} = 1(\mathbf{x}_{i0}'\boldsymbol{\beta}_0 + v_{i0} > 0),$$

$$\text{(dynamic model) } y_{it} = 1(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \alpha_i + v_{it} > 0)$$

$$\text{(heterogeneity correlated with participation) } \alpha_i = \mathbf{z}_i'\boldsymbol{\delta} + \eta_i,$$

(stochastic specification)

$$\eta_i \mid \mathbf{X}_i \sim \mathrm{N}\left[0, \sigma_\eta^2\right],$$

$$v_{i0} \mid \mathbf{X}_i \sim \mathrm{N}\left[0, \sigma_0^2\right],$$

$$w_{it} \mid \mathbf{X}_i \sim \mathrm{N}\left[0, \sigma_w^2\right],$$

$$v_{it} = \rho v_{i,t-1} + w_{it}, \sigma_\eta^2 + \sigma_w^2 = 1.$$

$$\mathrm{Corr}[v_{i0}, v_{it}] = \rho^t, \quad t = 1, \ldots, T-1.$$

The presence of the autocorrelation and state dependence in the model invalidate the simple maximum likelihood procedures we examined earlier. The appropriate likelihood function is constructed by formulating the probabilities as

$$\mathrm{Prob}(y_{i0}, y_{i1}, \ldots) = \mathrm{Prob}(y_{i0}) \times \mathrm{Prob}(y_{i1} \mid y_{i0}) \times \cdots \times \mathrm{Prob}(y_{iT} \mid y_{i,T-1}).$$

This still involves a $T = 7$ order normal integration, which is approximated in the study using a simulator similar to the GHK simulator discussed in 15.6.2.b. Among Hyslop's results are a comparison of the model fit by the simulator for the multivariate normal probabilities with the same model fit using the maximum simulated likelihood technique described in Section 15.6.

17.4.7 A SEMIPARAMETRIC MODEL FOR INDIVIDUAL HETEROGENEITY

The panel data analysis considered thus far has focused on modeling heterogeneity with the fixed and random effects specifications. Both assume that the heterogeneity is continuously distributed among individuals. The random effects model is fully parametric, requiring a full specification of the likelihood for estimation. The fixed effects model

[35] Beck et al. (2001) is a bit different from the others mentioned in that in their study of "state failure," they observe a large sample of countries (147) observed over a fairly large number of years, 40. As such, they are able to formulate their models in a way that makes the asymptotics with respect to T appropriate. They can analyze the data essentially in a time-series framework. Sepanski (2000) is another application that combines state dependence and the random coefficient specification of Akin, Guilkey, and Sickles (1979).

is essentially semiparametric. It requires no specific distributional assumption, however, it does require that the realizations of the latent heterogeneity be treated as parameters, either estimated in the unconditional fixed effects estimator or conditioned out of the likelihood function when possible. As noted in the preceding example, Heckman and Singer's (1984b) model provides a less stringent model specification based on a discrete distribution of the latent heterogeneity. A straightforward method of implementing their model is to cast it as a latent class model in which the classes are distinguished by different constant terms and the associated probabilities. The class probabilities are treated as parameters to be estimated with the model parameters.

Example 17.15 Semiparametric Models of Heterogeneity

We have extended the random effects and fixed effects logit models in Example 17.11 by fitting the Heckman and Singer (1984b) model. Table 17.12 shows the specification search and the results under different specifications. The first column of results shows the estimated fixed effects model from Example 17.11. The conditional estimates are shown in parentheses. Of the 7,293 groups in the sample, 3,056 are not used in estimation of the fixed effects models because the sum of $Doctor_{it}$ is either 0 or T_i for the group. The mean and standard deviation of the estimated underlying heterogeneity distribution are computed using the estimates of α_i for the remaining 4,237 groups. The remaining five columns in the table show the results for different numbers of latent classes in the Heckman and Singer model. The listed constant terms are the "mass points" of the underlying distributions. The associated class probabilities are shown in parentheses under them. The mean and standard deviation are derived from the

TABLE 17.12 Estimated Heterogeneity Models

		Number of Classes				
	Fixed Effect	***1***	***2***	***3***	***4***	***5***
β_1	0.10475	0.020708	0.030325	0.033684	0.034083	0.034159
	(0.084760)					
β_2	−0.060973	−0.18592	0.025550	−0.0058013	−0.0063516	−0.013627
	(−0.050383)					
β_3	−0.088407	−0.22947	−0.24708	−0.26388	−0.26590	−0.26626
	(−0.077764)					
β_4	−0.11671	−0.045588	−0.050924	−0.058022	−0.059751	−0.059176
	(−0.090816)					
β_5	−0.057318	0.085293	0.042974	0.037944	0.029227	0.030699
	(−0.52072)					
α_1	−2.62334	0.25111	0.91764	1.71669	1.94536	2.76670
		(1.00000)	(0.62681)	(0.34838)	(0.29309)	(0.11633)
α_2			−1.47800	−2.23491	−1.76371	1.18323
			(0.37319)	(0.18412)	(0.21714)	(0.26468)
α_3				−0.28133	−0.036739	−1.96750
				(0.46749)	(0.46341)	(0.19573)
α_4					−4.03970	−0.25588
					(0.026360)	(0.40930)
α_5						−6.48191
						(0.013960)
Mean	−2.62334	0.00000	0.023613	0.055059	0.063685	0.054705
Std. Dev.	3.13415	0.00000	1.158655	1.40723	1.48707	1.62143
ln L	−9458.638	−17673.10	−16353.14	−16278.56	−16276.07	−16275.85
	(−6299.02)					
AIC	1.00349	1.29394	1.19748	1.19217	1.19213	1.19226

2- to 5-point discrete distributions shown. It is noteworthy that the mean of the distribution is relatively stable, but the standard deviation rises monotonically. The search for the best model would be based on the AIC. As noted in Section 14.10, using a likelihood ratio test in this context is dubious, as the number of degrees of freedom is ambiguous. Based on the AIC, the four-class model is the preferred specification.

17.4.8 MODELING PARAMETER HETEROGENEITY

In Section 11.11, we examined specifications that extend the underlying heterogeneity to all the parameters of the model. We have considered two approaches. The random parameters, or mixed models discussed in Chapter 15 allow parameters to be distributed continuously across individuals. The latent class model in Section 16.10 specifies a discrete distribution instead. (The Heckman and Singer model in the previous section applies this method to the constant term.) Most of the focus to this point, save for Example 16.16, has been on linear models.

The random effects model can be cast as a model with a random constant term;

$$y_{it}^* = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \ldots, n, \; t = 1, \ldots, T_i,$$
$$y_{it} = 1 \quad \text{if } y_{it}^* > 0, \text{ and 0 otherwise,}$$

where $\alpha_i = \alpha + \sigma_u u_i$. This is simply a reinterpretation of the model we just analyzed. We might, however, now extend this formulation to the full parameter vector. The resulting structure is

$$y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it}, \quad i = 1, \ldots, n, \; t = 1, \ldots, T_i,$$
$$y_{it} = 1 \quad \text{if } y_{it}^* > 0, \text{ and 0 otherwise,}$$

where $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Gamma}\mathbf{u}_i$ where $\boldsymbol{\Gamma}$ is a nonnegative definite diagonal matrix—some of its diagonal elements could be zero for nonrandom parameters. The method of estimation is **maximum simulated likelihood**. The simulated log-likelihood is now

$$\ln L_{Simulated} = \sum_{i=1}^{n} \ln \left\{ \frac{1}{R} \sum_{r=1}^{R} \left[\prod_{t=1}^{T_i} F[q_{it}(\mathbf{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{\Gamma}\mathbf{u}_{ir}))] \right] \right\}.$$

The simulation now involves R draws from the multivariate distribution of $\mathbf{u}$. Because the draws are uncorrelated—$\boldsymbol{\Gamma}$ is diagonal—this is essentially the same estimation problem as the random effects model considered previously. This model is estimated in Example 17.16. Example 17.16 also presents a similar model that assumes that the distribution of $\boldsymbol{\beta}_i$ is discrete rather than continuous.

Example 17.16 Parameter Heterogeneity in a Binary Choice Model

We have extended the logit model for doctor visits from Example 17.15 to allow the parameters to vary randomly across individuals. The random parameters logit model is

$$\text{Prob}(Doctor_{it} = 1) = \Lambda(\beta_{1i} + \beta_{2i}\, Age_{it} + \beta_{3i}\, Income_{it} + \beta_{4i}\, Kids_{it} + \beta_{5i}\, Educ_{it} + \beta_{6i}\, Married_{it}),$$

where the two models for the parameter variation we have employed are:

$$\text{Continuous:} \quad \beta_{ki} = \beta_k + \sigma_k u_{ki}, \, u_{ki} \sim \text{N}[0, 1], k = 1, \ldots, 6, \text{Cov}[u_{ki}, u_{mi}] = 0,$$
$$\text{Discrete:} \quad \beta_{ki} = \beta_k^1 \text{ with probability } \pi_1$$
$$\beta_k^2 \text{ with probability } \pi_2$$
$$\beta_k^3 \text{ with probability } \pi_3.$$

TABLE 17.13 Estimated Heterogeneous Parameter Models

	Pooled	*Random Parameters*		*Latent Class*		
Variable	*Estimate: β*	*Estimate: β*	*Estimate: σ*	*Estimate: β*	*Estimate: β*	*Estimate: β*
Constant	0.25111	−0.034964	0.81651	0.96605	−0.18579	−1.52595
	(0.091135)	(0.075533)	(0.016542)	(0.43757)	(0.23907)	(0.43498)
Age	0.020709	0.026306	0.025330	0.049058	0.032248	0.019981
	(0.0012852)	(0.0011038)	(0.0004226)	(0.0069455)	(0.0031462)	(0.0062550)
Income	−0.18592	−0.0043649	0.10737	−0.27917	−0.068633	0.45487
	(0.075064)	(0.062445)	(0.038276)	(0.37149)	(0.16748)	(0.31153)
Kids	−0.22947	−0.17461	0.55520	−0.28385	−0.28336	−0.11708
	(0.029537)	(0.024522)	(0.023866)	(0.14279)	(0.066404)	(0.12363)
Education	−0.045588	−0.040510	0.037915	−0.025301	−0.057335	−0.09385
	(0.0056465)	(0.0047520)	(0.0013416)	(0.027768)	(0.012465)	(0.027965)
Married	0.085293	0.014618	0.070696	−0.10875	0.025331	0.23571
	(0.033286)	(0.027417)	(0.017362)	(0.17228)	(0.075929)	(0.14369)
Class	1.00000	1.00000		0.34833	0.46181	0.18986
Prob.	(0.00000)	(0.00000)		(0.038495)	(0.028062)	(0.022335)
ln L	−17673.10	−16271.72			−16265.59	

We have chosen a three-class latent class model for the illustration. In an application, one might undertake a systematic search, such as in Example 17.15, to find a preferred specification. Table 17.13 presents the fixed parameter (pooled) logit model and the two random parameters versions. (There are infinite variations on these specifications that one might explore—See Chapter 15 for discussion—we have shown only the simplest to illustrate the models.[36]

Figure 17.3 shows the implied distribution for the coefficient on age. For the continuous distribution, we have simply plotted the normal density. For the discrete distribution, we first obtained the mean (0.0358) and standard deviation (0.0107). Notice that the distribution is tighter than the estimated continuous normal (mean, 0.026, standard deviation, 0.0253). To suggest the variation of the parameter (purely for purpose of the display, because the distribution is discrete), we placed the mass of the center interval, 0.462, between the midpoints of the intervals between the center mass point and the two extremes. With a width of 0.0145 the density is $0.461 / 0.0145 = 31.8$. We used the same interval widths for the outer segments. This range of variation covers about five standard deviations of the distribution.

17.4.9 NONRESPONSE, ATTRITION AND INVERSE PROBABILITY WEIGHTING

Missing observations is a common problem in the analysis of panel data. Nicoletti and Peracchi (2005) suggest several reasons that, for example, panels become unbalanced:

- Demographic events such as death
- Movement out of the scope of the survey, such as institutionalization or emigration

[36] We have arrived (once again) at a point where the question of replicability arises. Nonreplicability is an ongoing challenge in empirical work in economics. (See, e.g., Example 17.12.) The problem is particularly acute in analyses that involve simulation such as Monte Carlo studies and random parameter models. In the interest of replicability, we note that the random parameter estimates in Table 17.14 were computed with NLOGIT [Econometric Software (2007)] and are based on 50 Halton draws. We used the first six sequences (prime numbers 2, 3, 5, 7, 11, 13) and discarded the first 10 draws in each sequence.

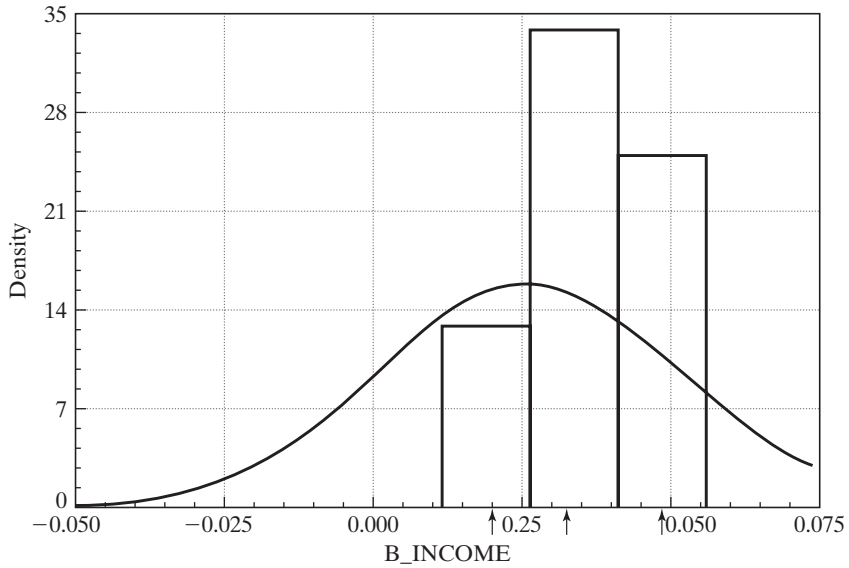

FIGURE 17.3 Distributions of Income Coefficient.

- Refusal to respond at subsequent waves
- Absence of the person at the address
- Other types of noncontact

The GSOEP that we (from Riphahn, Wambach, and Million (2003)) have used in many examples in this text is one such data set. Jones, Koolman, and Rice (2006) (JKR) list several other applications, including the British Household Panel Survey (BHPS), the European Community Household Panel (ECHP), and the Panel Study of Income Dynamics (PSID).

If observations are missing completely at random (MCAR), then the problem of nonresponse can be ignored, though for estimation of dynamic models, either the analysis will have to be restricted to observations with uninterrupted sequences of observations, or some very strong assumptions and interpolation methods will have to be employed to fill the gaps. (See Section 4.7.4 for discussion of the terminology and issues in handling missing data.) The problem for estimation arises when observations are missing for reasons that are related to the outcome variable of interest. **Nonresponse bias** and a related problem, **attrition bias** (individuals leave permanently during the study) result when conventional estimators, such as least squares or the probit maximum likelihood estimator being used here, are applied to samples in which observations are present or absent from the sample for reasons related to the outcome variable. It is a form of **sample selection bias**, that we will examine further in Chapter 19.

Verbeek and Nijman (1992) have suggested a test for endogeneity of the sample response pattern. (We will adopt JKR's notation and terminology for this.) Let h denote the outcome of interest and $\mathbf{x}$ denote the relevant set of covariates. Let R denote the pattern of response. If nonresponse is (completely) random, then $E[h \mid \mathbf{x}, R] = E[h \mid \mathbf{x}]$. This suggests a variable addition test (neglecting other panel data effects); a pooled

model that contains R in addition to $\mathbf{x}$ can provide the means for a simple test of endogeneity. JKR (and Verbeek and Nijman) suggest using the number of waves at which the individual is present as the measure of R. Thus, adding R to the pooled model, we can use a simple t test for the hypothesis.

Devising an estimator given that (non)response is nonignorable requires a more detailed understanding of the process generating the response pattern. The crucial issue is whether the sample selection is based "on unobservables" or "on observables." **Selection on unobservables** results when, after conditioning on the relevant variables, $\mathbf{x}$ and other information, $\mathbf{z}$, the sampling mechanism is still nonrandom with respect to the disturbances in the models. Selection on unobservables is at the heart of the sample selectivity methodology pioneered by Heckman (1979) that we will study in Chapter 19. (Some applications of the role of unobservables in biased estimation are discussed in Chapter 8, where we examine sources of endogeneity in regression models.) If selection is on observables and then conditioned on an appropriate specification involving the observable information, $(\mathbf{x},\mathbf{z})$, a consistent estimator of the model parameters will be available by "purging" the estimator of the endogeneity of the sampling mechanism.

JKR adopt an **inverse probability weighted (IPW)** estimator devised by Robins, Rotnitsky and Zhao (1995), Fitzgerald, Gottshalk, and Moffitt (1998), Moffitt, Fitzgerald and Gottshalk (1999), and Wooldridge (2002). The estimator is based on the general MCAR assumption that $P(R = 1 \mid h, \mathbf{x}, \mathbf{z}) = P(R = 1 \mid \mathbf{x}, z)$. That is, the observable covariates convey all the information that determines the response pattern—the probability of nonresponse does not vary systematically with the outcome variable once the exogenous information is accounted for. Implementing this idea in an estimator would require that $\mathbf{x}$ and $\mathbf{z}$ be observable when $R = 0$, that is, the exogenous data be available for the nonresponders. This will typically not be the case; in an unbalanced panel, the entire observation is missing. Wooldridge (2002) proposed a somewhat stronger assumption that makes estimation feasible: $P(R = 1 \mid h, \mathbf{x}, \mathbf{z}) = P(R = 1 \mid \mathbf{z})$ where $\mathbf{z}$ is a set of covariates available at wave 1 (entry to the study). To compute Wooldridge's IPW estimator, we will begin with the sample of all individuals who are present at wave 1 of the study. (In our Example 17.17, based on the GSOEP data, not all individuals are present at the first wave.) At wave 1, $(\mathbf{x}_{i1}, \mathbf{z}_{i1})$ are observed for all individuals to be studied; $\mathbf{z}_{i1}$ contains information on observables that are not included in the outcome equation and that predict the response pattern at subsequent waves, including the response variable at the first wave. At wave 1, then, $P(R_{i1} = 1 \mid \mathbf{x}_{i1}, \mathbf{z}_{i1}) = 1$. Wooldridge suggests using a probit model for $P(R_{it} = 1 \mid \mathbf{x}_{i1}, \mathbf{z}_{i1})$, $t = 2, \ldots, T$ for the remaining waves to obtain predicted probabilities of response, $\hat{p}_{it}$. The IPW estimator then maximizes the weighted log likelihood

$$\ln L_{IPW} = \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{R_{it}}{\hat{p}_{it}} \ln L_{it}.$$

Inference based on the weighted log-likelihood function can proceed as in Section 17.3. A remaining detail concerns whether the use of the predicted probabilities in the weighted log-likelihood function makes it necessary to correct the standard errors for two-step estimation. The case here is not an application of the two-step estimators we considered in Section 14.7, since the first step is not used to produce an estimated parameter vector in the second. Wooldridge (2002) shows that the standard errors computed

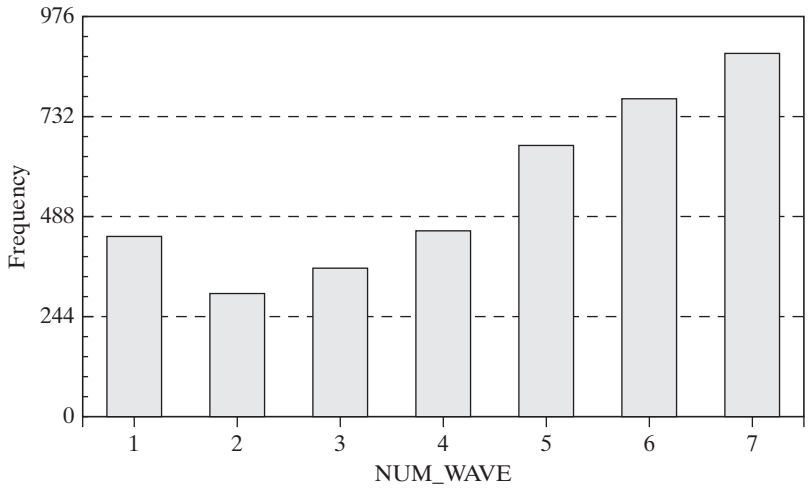

FIGURE 17.4 Number of Waves Responded for Those Present at Wave

without the adjustment are "conservative" in that they are larger than they would be with the adjustment.

Example 17.17 Nonresponse in the GSOEP Sample

Of the 7,293 individuals in the GSOEP data that we have used in several earlier examples, 3,874 were present at wave 1 (1984) of the sample. The pattern of the number of waves present by these 3,874 is shown in Figure 17.4. The waves are 1984–1988, 1991, and 1994. A dynamic model would be based on the 1,600 of those present at wave 1 who were also present for the next four waves. There is a substantial amount of nonresponse in these data. Not all individuals exit the sample with the first nonresponse, however, so the resulting panel remains unbalanced. The impression suggested by Figure 17.4 could be a bit misleading—the nonresponse pattern is quite different from simple attrition. For example, of the 3,874 individuals who responded at wave 1, 364 did not respond at wave 2 but returned to the sample at wave 3.

To employ the Verbeek and Nijman test, we used the entire sample of 27,326 household years of data. The pooled probit model for DocVis $>$ 0 produced the results at the left in Table 17.14. A *t* (Wald) test of the hypothesis that the coefficient on number of waves present is zero is strongly rejected, so we proceed to the inverse probability weighted estimator. For computing the inverse probability weights, we used the following specification:

x_{i1} = *constant, age, income, educ, kids, married*

z_{i1} = *female, handicapped dummy, percentage handicapped, university, working, blue collar, white collar, public servant,* y_{i1}

y_{i1} = *Doctor Visits* $>$ 0 *in period* 1.

This first-year data vector is used as the observed explanatory variables in probit models for waves 2–7 for the 3,874 individuals who were present at wave 1. There are 3,874 observations for each of these probit models, since all were observed at wave 1. Fitted probabilities for R_{it} are computed for waves 2–7, while $R_{i1} = 1$. The sample means of these probabilities which equals the proportion of the 3,874 who responded at each wave are 1.000, 0.730, 0.672, 0.626, 0.682, 0.568, and 0.386, respectively. Table 17.14 presents the estimated models for several specifications In each case, it appears that the weighting brings some moderate changes in the parameters and, uniformly, reductions in the standard errors.

TABLE 17.14 Inverse Probability Weighted Estimators

		Pooled Model		*Random Effects—Mundlak*		*Fixed Effects*	
Variable	***Endog. Test***	***Unwtd.***	***IPW***	***Unwtd.***	***IPW***	***Unwtd.***	***IPW***
Constant	0.26411	0.03369	−0.02373	0.09838	0.13237		
	(0.05893)	(0.07684)	(0.06385)	(0.16081)	(0.17019)		
Age	0.01369	0.01667	0.01831	0.05141	0.05656	0.06210	0.06841
	(0.00080)	(0.00107)	(0.00088)	(0.00422)	(0.00388)	(0.00506)	(0.00465)
Income	−0.12446	−0.17097	−0.22263	0.05794	0.01699	0.07880	0.03603
	(0.04636)	(0.05981)	(0.04801)	(0.11256)	(0.10580)	(0.12891)	(0.12193)
Education	−0.02925	−0.03614	−0.03513	−0.06456	−0.07058	−0.07752	−0.08574
	(0.00351)	(0.00449)	(0.00365)	(0.06104)	(0.05792)	(0.06582)	(0.06149)
Kids	−0.13130	−0.13077	−0.13277	−0.04961	−0.03427	−0.05776	−0.03546
	(0.01828)	(0.02303)	(0.01950)	(0.04500)	(0.04356)	(0.05296)	(0.05166)
Married	0.06759	0.06237	0.07015	−0.06582	−0.09235	−0.07939	−0.11283
	(0.02060)	(0.02616)	(0.02097)	(0.06596)	(0.06330)	(0.08146)	(0.07838)
Mean Age				−0.03056	−0.03401		
				(0.00479)	(0.00455)		
Mean Income				−0.66388	−0.78077		
				(0.18646)	(0.18866)		
Mean Education				0.02656	0.02899		
				(0.06160)	(0.05848)		
Mean Kids				−0.17524	−0.20615		
				(0.07266)	(0.07464)		
Mean Married				0.22346	0.25763		
				(0.08719)	(0.08433)		
Number of Waves	−0.02977						
	(0.00450)						
ρ				0.46538	0.48616		

17.5 BIVARIATE AND MULTIVARIATE PROBIT MODELS

In Chapter 10, we analyzed a number of different multiple-equation extensions of the classical and generalized regression model. A natural extension of the probit model would be to allow more than one equation, with correlated disturbances, in the same spirit as the seemingly unrelated regressions model. The general specification for a two-equation model would be

$$\begin{aligned} y_1^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \quad y_1 = 1 \text{ if } y_1^* > 0, 0 \quad \text{otherwise,} \\ y_2^* &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2, \quad y_2 = 1 \text{ if } y_2^* > 0, 0 \quad \text{otherwise,} \\ &\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} | \mathbf{x}_1, \mathbf{x}_2 \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]. \end{aligned} \tag{17-48}$$

This bivariate probit model is interesting in its own right for modeling the joint determination of two variables, such as doctor and hospital visits in the next example. It also provides the framework for modeling in two common applications. In many cases, a treatment effect, or endogenous influence, takes place in a binary choice context. The bivariate probit model provides a specification for analyzing a case in which a probit

model contains an endogenous binary variable in one of the equations. In Example 17.21, we will extend (17-48) to

$$\begin{aligned} &W^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \quad W = 1 \text{ if } W^* > 0, 0 \text{ otherwise,} \\ &y^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \gamma W + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, 0 \text{ otherwise,} \\ &\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} | \mathbf{x}_1, \mathbf{x}_2 \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]. \end{aligned} \tag{17-49}$$

This model extends the case in Section 17.3.5, where W^*, rather than W, appears on the right-hand side of the second equation. In the example, W denotes whether a liberal arts college supports a women's studies program on the campus while y is a binary indicator of whether the economics department provides a gender economics course. A second common application, in which the first equation is an endogenous sampling rule, is another variant of the bivariate probit model:

$$\begin{aligned} &S^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \quad \text{S} = 1 \text{ if } S^* > 0, 0 \quad \text{otherwise,} \\ &y^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, 0 \quad \text{otherwise,} \\ &\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} | \mathbf{x}_1, \mathbf{x}_2 \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right], \\ &(y,\ \mathbf{x}_2) \text{ observed only when } S = 1. \end{aligned} \tag{17-50}$$

In Example 17.22, we will study an application in which S is the result of a credit card application (or any sort of loan application) while y_2 is a binary indicator for whether the individual defaults on the credit account (loan). This is a form of endogenous sampling (in this instance, sampling on unobservables) that has some commonality with the attrition problem that we encountered in Section 17.4.9.

At the end of this section, we will extend (17-48) to more than two equations. This will allow direct treatment of multiple binary outcomes. It will also allow a more general panel data model for T periods than is provided by the random effects specification.

17.5.1 MAXIMUM LIKELIHOOD ESTIMATION

The bivariate normal cdf is

$$\text{Prob}(X_1 < x_1, X_2 < x_2) = \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} \phi_2(z_1, z_2, \rho)\, dz_1 dz_2,$$

which we denote $\Phi_2(x_1, x_2, \rho)$. The density is[37]

$$\phi_2(x_1, x_2, \rho) = \frac{e^{-(1/2)(x_1^2 + x_2^2 - 2\rho x_1 x_2)/(1-\rho^2)}}{2\pi(1-\rho^2)^{1/2}}.$$

To construct the log-likelihood, let $q_{i1} = 2y_{i1} - 1$ and $q_{i2} = 2y_{i2} - 1$. Thus, $q_{ij} = 1$ if $y_{ij} = 1$ and -1 if $y_{ij} = 0$ for $j = 1$ and 2. Now let

$$z_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}_j \quad \text{and} \quad w_{ij} = q_{ij} z_{ij}, \quad j = 1, 2,$$

[37] See Section B.9.

and

$$\rho_{i^*} = q_{i1}q_{i2}\rho.$$

Note the notational convention. The subscript 2 is used to indicate the bivariate normal distribution in the density ϕ_2 and cdf Φ_2. In all other cases, the subscript 2 indicates the variables in the second equation. As before, $\phi(.)$ and $\Phi(.)$ without subscripts denote the univariate standard normal density and cdf.

The probabilities that enter the likelihood function are

$$\text{Prob}(Y_1 = y_{i1}, Y_2 = y_{i2} \mid \mathbf{x}_1, \mathbf{x}_2) = \Phi_2(w_{i1}, w_{i2}, \rho_{i^*}),$$

which accounts for all the necessary sign changes needed to compute probabilities for y's equal to zero and one. Thus,[38]

$$\ln L = \sum_{i=1}^{n} \ln \Phi_2(w_{i1}, w_{i2}, \rho_{i^*}).$$

The derivatives of the log-likelihood then reduce to

$$\begin{aligned} \frac{\partial \ln L}{\partial \boldsymbol{\beta}_j} &= \sum_{i=1}^{n} \left(\frac{q_{ij} g_{ij}}{\Phi_2} \right) \mathbf{x}_{ij}, \quad j = 1, 2, \\ \frac{\partial \ln L}{\partial \rho} &= \sum_{i=1}^{n} \frac{q_{i1}q_{i2}\phi_2}{\Phi_2}, \end{aligned} \tag{17-51}$$

where

$$g_{i1} = \phi(w_{i1})\Phi\left[\frac{w_{i2} - \rho_{i^*} w_{i1}}{\sqrt{1 - \rho_{i^*}^2}} \right] \tag{17-52}$$

and the subscripts 1 and 2 in g_{i1} are reversed to obtain g_{i2}. Before considering the Hessian, it is useful to note what becomes of the preceding if $\rho = 0$. For $\partial \ln L/\partial \boldsymbol{\beta}_1$, if $\rho = \rho_{i^*} = 0$, then g_{i1} reduces to $\phi(w_{i1})\Phi(w_{i2})$, ϕ_2 is $\phi(w_{i1})\phi(w_{i2})$, and Φ_2 is $\Phi(w_{i1})\Phi(w_{i2})$. Inserting these results in (17-51) with q_{i1} and q_{i2} produces (17-21). Because both functions in $\partial \ln L/\partial\rho$ factor into the product of the univariate functions, $\partial \ln L/\partial\rho$ reduces to $\sum_{i=1}^{n} \lambda_{i1}\lambda_{i2}$, where λ_{ij}, $j = 1, 2$, is defined in (17-20). (This result will reappear in the LM statistic shown later.)

The maximum likelihood estimates are obtained by simultaneously setting the three derivatives to zero. The second derivatives are relatively straightforward but tedious. Some simplifications are useful. Let

$$\begin{aligned} \delta_i &= \frac{1}{\sqrt{1 - \rho_{i^*}^2}}, \\ v_{i1} &= \delta_i(w_{i2} - \rho_{i^*} w_{i1}), \quad \text{so } g_{i1} = \phi(w_{i1})\Phi(v_{i1}), \\ v_{i2} &= \delta_i(w_{i1} - \rho_{i^*} w_{i2}), \quad \text{so } g_{i2} = \phi(w_{i2})\Phi(v_{i2}). \end{aligned}$$

By multiplying it out, you can show that

$$\delta_i \phi(w_{i1})\phi(v_{i1}) = \delta_i \phi(w_{i2})\phi(v_{i2}) = \phi_2.$$

[38] To avoid further ambiguity, and for convenience, the observation subscript will be omitted from $\Phi_2 = \Phi_2(w_{i1}, w_{i2}, \rho_{i^*})$ and from $\phi_2 = \phi_2(w_{i1}, w_{i2}, \rho_{i^*})$.

Then

$$\begin{aligned}
\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1'} &= \sum_{i=1}^{n} \mathbf{x}_{i1}\mathbf{x}_{i1}' \left[\frac{-w_{i1}g_{i1}}{\Phi_2} - \frac{\rho_{i*}\phi_2}{\Phi_2} - \frac{g_{i1}^2}{\Phi_2^2} \right], \\
\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_2'} &= \sum_{i=1}^{n} q_{i1}q_{i2}\mathbf{x}_{i1}\mathbf{x}_{i2}' \left[\frac{\phi_2}{\Phi_2} - \frac{g_{i1}g_{i2}}{\Phi_2^2} \right], \\
\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}_1 \partial \rho} &= \sum_{i=1}^{n} q_{i2}\mathbf{x}_{i1} \frac{\phi_2}{\Phi_2} \left[\rho_{i*}\delta_i v_{i1} - w_{i1} - \frac{g_{i1}}{\Phi_2} \right], \\
\frac{\partial^2 \log L}{\partial \rho^2} &= \sum_{i=1}^{n} \frac{\phi_2}{\Phi_2} \left[\delta_i^2 \rho_{i*}(1 - \mathbf{w}_i'\mathbf{R}_i^{-1}\mathbf{w}_i) + \delta_i^2 w_{i1}w_{i2} - \frac{\phi_2}{\Phi_2} \right],
\end{aligned} \tag{17-53}$$

where $\mathbf{w}_i'\mathbf{R}_i^{-1}\mathbf{w}_i = \delta_i^2(w_{i1}^2 + w_{i2}^2 - 2\rho_{i*}w_{i1}w_{i2})$. (For $\boldsymbol{\beta}_2$, change the subscripts in $\partial^2 \ln L/\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1'$ and $\partial^2 \ln L/\partial\boldsymbol{\beta}_1\partial\rho$ accordingly.) The complexity of the second derivatives for this model makes it an excellent candidate for the Berndt et al. estimator of the variance matrix of the maximum likelihood estimator.

Example 17.18 Tetrachoric Correlation

Returning once again to the health care application of Examples 17.4 and several others, we now consider a second binary variable,

$$Hospital_{it} = 1 \text{ if } HospVis_{it} > 0 \text{ and 0 otherwise.}$$

Our previous analyses have focused on

$$Doctor_{it} = 1 \text{ if } DocVis_{it} > 0 \text{ and 0 otherwise.}$$

A simple bivariate frequency count for these two variables is

	Hospital		
Doctor	**0**	**1**	**Total**
0	9,715	420	10,135
1	15,216	1,975	17,191
Total	24,931	2,395	27,326

Looking at the very large value in the lower-left cell, one might surmise that these two binary variables (and the underlying phenomena that they represent) are negatively correlated. The usual Pearson, product moment correlation would be inappropriate as a measure of this correlation since it is used for continuous variables. Consider, instead, a bivariate probit "model,"

$$\begin{aligned}
H_{it}^* &= \mu_1 + \varepsilon_{1,it}, \quad Hospital_{it} = 1(H_{it}^* > 0), \\
D_{it}^* &= \mu_2 + \varepsilon_{2,it}, \quad Doctor_{it} = 1(D_{it}^* > 0),
\end{aligned}$$

where $(\varepsilon_1, \varepsilon_2)$ have a bivariate normal distribution with means (0, 0), variances (1, 1) and correlation ρ. This is the model in (17-48) without independent variables. In this representation, the **tetrachoric correlation,** which is a correlation measure for a pair of binary variables, is precisely the ρ in this model—it is the correlation that would be measured between the underlying continuous variables if they could be observed. This suggests an interpretation of the correlation coefficient in a bivariate probit model—as the conditional tetrachoric correlation. It also suggests a method of easily estimating the tetrachoric correlation coefficient using a program that is built into nearly all commercial software packages.

Applied to the hospital/doctor data defined earlier, we obtained an estimate of ρ of 0.31106, with an estimated asymptotic standard error of 0.01357. Apparently, our earlier intuition was incorrect.

17.5.2 TESTING FOR ZERO CORRELATION

The Lagrange multiplier statistic is a convenient device for testing for the absence of correlation in this model. Under the null hypothesis that ρ equals zero, the model consists of independent probit equations, which can be estimated separately. Moreover, in the multivariate model, all the bivariate (or multivariate) densities and probabilities factor into the products of the marginals if the correlations are zero, which makes construction of the test statistic a simple matter of manipulating the results of the independent probits. The Lagrange multiplier statistic for testing H_0: $\rho = 0$ in a bivariate probit model is[39]

$$\text{LM} = \frac{\left[\sum_{i=1}^{n} q_{i1}q_{i2}\dfrac{\phi(w_{i1})\phi(w_{i2})}{\Phi(w_{i1})\Phi(w_{i2})}\right]^2}{\sum_{i=1}^{n}\dfrac{[\phi(w_{i1})\phi(w_{i2})]^2}{\Phi(w_{i1})\Phi(-w_{i1})\Phi(w_{i2})\Phi(-w_{i2})}}.$$

As usual, the advantage of the LM statistic is that it obviates computing the bivariate probit model. But the full unrestricted model is now fairly common in commercial software, so that advantage is minor. The likelihood ratio or Wald test can often be used with equal ease. To carry out the likelihood ratio test, we note first that if ρ equals zero, then the bivariate probit model becomes two independent univariate probits models. The log-likelihood in that case would simply be the sum of the two separate log-likelihoods. The test statistic would be

$$\lambda_{\text{LR}} = 2[\ln\ L_{\text{BIVARIATE}} - (\ln L_1 + \ln L_2)].$$

This would converge to a chi-squared variable with one degree of freedom. The Wald test is carried out by referring

$$\lambda_{\text{WALD}} = \left[\hat{\rho}_{MLE}/\sqrt{\text{Est. Asy. Var}[\hat{\rho}_{MLE}]}\right]^2$$

to the chi-squared distribution with one degree of freedom. For 95 percent significance, the critical value is 3.84 (or one can refer the positive square root to the standard normal critical value of 1.96). Example 17.19 demonstrates.

17.5.3 PARTIAL EFFECTS

There are several "marginal effects" one might want to evaluate in a bivariate probit model.[40] A natural first step would be the derivatives of $\text{Prob}[y_1 = 1, y_2 = 1 \mid \mathbf{x}_1, \mathbf{x}_2]$. These can be deduced from (17-49) by multiplying by Φ_2, removing the sign carrier, q_{ij} and differentiating with respect to $\mathbf{x}_j$ rather than $\boldsymbol{\beta}_j$. The result is

$$\frac{\partial \Phi_2(\mathbf{x}_1'\boldsymbol{\beta}_1, \mathbf{x}_2'\boldsymbol{\beta}_2, \rho)}{\partial \mathbf{x}_1} = \phi(\mathbf{x}_1'\boldsymbol{\beta}_1)\Phi\left(\frac{\mathbf{x}_2'\boldsymbol{\beta}_2 - \rho\mathbf{x}_1'\boldsymbol{\beta}_1}{\sqrt{1-\rho^2}}\right)\boldsymbol{\beta}_1.$$

Note, however, the bivariate probability, albeit possibly of interest in its own right, is not a conditional mean function. As such, the preceding does not correspond to a regression coefficient or a slope of a conditional expectation.

[39] This is derived in Kiefer (1982).

[40] See Greene (1996b) and Christofides et al. (1997, 2000).

For convenience in evaluating the conditional mean and its partial effects, we will define a vector $\mathbf{x} = \mathbf{x}_1 \cup \mathbf{x}_2$ and let $\mathbf{x}_1'\boldsymbol{\beta}_1 = \mathbf{x}'\boldsymbol{\gamma}_1$. Thus, $\boldsymbol{\gamma}_1$ contains all the nonzero elements of $\boldsymbol{\beta}_1$ and possibly some zeros in the positions of variables in $\mathbf{x}$ that appear only in the other equation; $\boldsymbol{\gamma}_2$ is defined likewise. The bivariate probability is

$$\text{Prob}[y_1 = 1, y_2 = 1 \mid \mathbf{x}] = \Phi_2[\mathbf{x}'\boldsymbol{\gamma}_1, \mathbf{x}'\boldsymbol{\gamma}_2, \rho].$$

Signs are changed appropriately if the probability of the zero outcome is desired in either case. (See 17-48.) The marginal effects of changes in $\mathbf{x}$ on this probability are given by

$$\frac{\partial \Phi_2}{\partial \mathbf{x}} = g_1\boldsymbol{\gamma}_1 + g_2\boldsymbol{\gamma}_2,$$

where g_1 and g_2 are defined in (17-50). The familiar univariate cases will arise if $\rho = 0$, and effects specific to one equation or the other will be produced by zeros in the corresponding position in one or the other parameter vector. There are also some conditional mean functions to consider. The unconditional mean functions are given by the univariate probabilities:

$$E[y_j \mid \mathbf{x}] = \Phi(\mathbf{x}'\boldsymbol{\gamma}_j), \quad j = 1, 2,$$

so the analysis of (17-9) and (17-10) applies. One pair of conditional mean functions that might be of interest are

$$E[y_1 \mid y_2 = 1, \mathbf{x}] = \text{Prob}[y_1 = 1 \mid y_2 = 1, \mathbf{x}] = \frac{\text{Prob}[y_1 = 1, y_2 = 1 \mid \mathbf{x}]}{\text{Prob}[y_2 = 1 \mid \mathbf{x}]}$$

$$= \frac{\Phi_2(\mathbf{x}'\boldsymbol{\gamma}_1, \mathbf{x}'\boldsymbol{\gamma}_2, \rho)}{\Phi(\mathbf{x}'\boldsymbol{\gamma}_2)}$$

and similarly for $E[y_2 \mid y_1 = 1, \mathbf{x}]$. The marginal effects for this function are given by

$$\frac{\partial E[y_1 \mid y_2 = 1, \mathbf{x}]}{\partial \mathbf{x}} = \left(\frac{1}{\Phi(\mathbf{x}'\boldsymbol{\gamma}_2)}\right)\left[g_1\boldsymbol{\gamma}_1 + \left(g_2 - \Phi_2\frac{\phi(\mathbf{x}'\boldsymbol{\gamma}_2)}{\Phi(\mathbf{x}'\boldsymbol{\gamma}_2)}\right)\boldsymbol{\gamma}_2\right].$$

Finally, one might construct the nonlinear conditional mean function

$$E[y_1 \mid y_2, \mathbf{x}] = \frac{\Phi_2[\mathbf{x}'\boldsymbol{\gamma}_1, (2y_2 - 1)\mathbf{x}'\boldsymbol{\gamma}_2, (2y_2 - 1)\rho]}{\Phi[(2y_2 - 1)\mathbf{x}'\boldsymbol{\gamma}_2]}.$$

The derivatives of this function are the same as those presented earlier, with sign changes in several places if $y_2 = 0$ is the argument.

Example 17.19 Bivariate Probit Model for Health Care Utilization

We have extended the bivariate probit model of the previous example by specifying a set of independent variables,

$$\mathbf{x}_i = \textit{Constant}, \textit{Female}_i, \textit{Age}_{it}, \textit{Income}_{it}, \textit{Kids}_{it}, \textit{Education}_{it}, \textit{Married}_{it}.$$

We have specified that the same exogenous variables appear in both equations. (There is no requirement that different variables appear in the equations, nor that a variable be excluded from each equation.) The correct analogy here is to the seemingly unrelated regressions model, not to the linear simultaneous equations model. Unlike the SUR model of Chapter 10, it is not the case here that having the same variables in the two equations implies that the model can be fit equation by equation, one equation at a time. That result only applies to the estimation of sets of linear regression equations.

TABLE 17.15 Estimated Bivariate Probit Model[a]

	Doctor					*Hospital*	
	Model Estimates		*Partial Effects*			*Model Estimates*	
Variable	*Univariate*	*Bivariate*	*Direct*	*Indirect*	*Total*	*Univariate*	*Bivariate*
Constant	−0.1243 (0.05815)	−0.1243 (0.05814)				−1.3328 (0.08320)	−1.3385 (0.07957)
Female	0.3559 (0.01602)	0.3551 (0.01604)	0.09650 (0.004957)	−0.00724 (0.001515)	0.08926 (0.005127)	0.1023 (0.02195)	0.1050 (0.02174)
Age	0.01189 (0.0007957)	0.01188 (0.000802)	0.003227 (0.000231)	−0.00032 (0.000073)	0.002909 (0.000238)	0.004605 (0.001082)	0.00461 (0.001058)
Income	−0.1324 (0.04655)	−0.1337 (0.04628)	−0.03632 (0.01260)	−0.003064 (0.004105)	−0.03939 (0.01254)	0.03739 (0.06329)	0.04441 (0.05946)
Kids	−0.1521 (0.01833)	−0.1523 (0.01825)	−0.04140 (0.005053)	0.001047 (0.001773)	−0.04036 (0.005168)	−0.01714 (0.02562)	−0.01517 (0.02570)
Education	−0.01497 (0.003575)	−0.01484 (0.003575)	−0.004033 (0.000977)	0.001512 (0.00035)	−0.002521 (0.0010)	−0.02196 (0.005215)	−0.02191 (0.005110)
Married	0.07352 (0.02064)	0.07351 (0.02063)	0.01998 (0.005626)	0.003303 (0.001917)	0.02328 (0.005735)	−0.04824 (0.02788)	−0.04789 (0.02777)

[a] Estimated correlation coefficient = 0.2981 (0.0139).

Table 17.15 contains the estimates of the parameters of the univariate and bivariate probit models. The tests of the null hypothesis of zero correlation strongly reject the hypothesis that ρ equals zero. The t statistic for ρ based on the full model is $0.2981 / 0.0139 = 21.446$, which is much larger than the critical value of 1.96. For the likelihood ratio test, we compute

$$\lambda_{LR} = 2\{-25285.07 - [-17422.72 - (-8073.604)]\} = 422.508.$$

Once again, the hypothesis is rejected. (The Wald statistic is $21.446^2 = 459.957$.) The LM statistic is 383.953. The coefficient estimates agree with expectations. The income coefficient is statistically significant in the doctor equation, but not in the hospital equation, suggesting, perhaps, that physican visits are at least to some extent discretionary while hospital visits occur on an emergency basis that would be much less tied to income. The table also contains the decomposition of the partial effects for $E[y_1 \mid y_2 = 1]$. The direct effect is $[g_1/\Phi(\mathbf{x}'\gamma_2)]\boldsymbol{\gamma}_1$ in the definition given earlier. The mean estimate of $E[y_1 \mid y_2 = 1]$ is 0.821285. In the table in Example 17.8, this would correspond to the raw proportion $P(D = 1, H = 1) / P(H = 1) = (1975 / 27326) / 2395 / 27326) = 0.8246$.

17.5.4 A PANEL DATA MODEL FOR BIVARIATE BINARY RESPONSE

Extending multiple equation models to accommodate unobserved common effects in panel data settings is straightforward in theory, but complicated in practice. For the bivariate probit case, for example, the natural extension of (17-48) would be

$$y^*_{1,it} = \mathbf{x}'_{1,it}\boldsymbol{\beta}_1 + \varepsilon_{1,it} + \alpha_{1,i}, \quad y_{1,it} = 1 \text{ if } y^*_{1,it} > 0, 0 \text{ otherwise},$$
$$y^*_{2,it} = \mathbf{x}'_{2,it}\boldsymbol{\beta}_2 + \varepsilon_{2,it} + \alpha_{2,i}, \quad y_{2,it} = 1 \text{ if } y^*_{2,it} > 0, 0 \text{ otherwise},$$
$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} | \mathbf{x}_1, \mathbf{x}_2 \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right].$$

The complication will be in how to treat (α_1, α_2). A fixed effects treatment will require estimation of two full sets of dummy coefficients, will likely encounter the incidental parameters problem in double measure, and will be complicated in practical terms.

As in all earlier cases, the fixed effects case also preempts any specification involving time-invariant variables. It is also unclear in a fixed effects model, how any correlation between α_1 and α_2 would be handled. It should be noted that strictly from a consistency standpoint, these considerations are moot. The two equations can be estimated separately, only with some loss of efficiency. The analogous situation would be the seemingly unrelated regressions model in Chapter 10. A random effects treatment (perhaps accommodated with Mundlak's approach of adding the group means to the equations as in Section 17.4.5) offers greater promise. If $(\alpha_1, \alpha_2) = (u_1, u_2)$ are normally distributed random effects, with

$$\begin{pmatrix} u_{1,i} \\ u_{2,i} \end{pmatrix} | \mathbf{X}_{1,i}, \mathbf{X}_{2,i} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right],$$

then the unconditional log likelihood for the bivariate probit model,

$$\ln L = \sum_{i=1}^{n} \ln \int_{u_1,u_2} \prod_{t=1}^{T_i} \Phi_2(w_{1,it} \mid u_{1,i}, w_{2,it} \mid u_{2,i}, \rho_{it}^*) \, f(u_{1,i}, u_{2,i}) du_{1,i} \, du_{2,i},$$

can be maximized using simulation or quadrature as we have done in previous applications. A possible variation on this specification would specify that the same common effect enter both equations. In that instance, the integration would only be over a single dimension. In this case, there would only be a single new parameter to estimate, σ^2, the variance of the common random effect while ρ would equal one. A refinement on this form of the model would allow the scaling to be different in the two equations by placing u_i in the first equation and θu_i in the second. This would introduce the additional scaling parameter, but ρ would still equal one. This is the formulation of a common random effect used in Heckman's formulation of the dynamic panel probit model in the Section 17.4.6.

Example 17.20 Bivariate Random Effects Model for Doctor and Hospital Visits

We will extend the pooled bivariate probit model presented in Example 17.19 by allowing a general random effects formulation, with free correlation between the time-varying components $(\varepsilon_1, \varepsilon_2)$ and between the time-invariant effects, (u_1, u_2). We used simulation to fit the model. Table 17.16 presents the pooled and random effects estimates. The log-likelihood functions for the pooled and random effects models are −25285.07 and −23769.67, respectively. Two times the difference is 3030.76. This would be a chi squared with three degrees of freedom (for the three free elements in the covariance matrix of u_1 and u_2). The 95 percent critical value is 7.81, so the pooling hypothesis would be rejected. The change in the correlation coefficient from .2981 to .1501 suggests that we have decomposed the disturbance in the model into a time-varying part and a time-invariant part. The latter seems to be the smaller of the two. Although the time-invariant elements are more highly correlated, their variances are only $0.2233^2 = 0.0499$ and $0.6338^2 = 0.4017$ compared to 1.0 for both ε_1 and ε_2.

17.5.5 ENDOGENOUS BINARY VARIABLE A RECURSIVE BIVARIATE PROBIT MODEL

Section 17.3.5 examines a case in which there is an endogenous variable in a binary choice (probit) model. The model is

$$W^* = \mathbf{x}_1'\beta_1 + \varepsilon_1,$$
$$y^* = \mathbf{x}_2'\beta_2 + \gamma W^* + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, 0 \text{ otherwise},$$
$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} | \mathbf{x}_1, \mathbf{x}_2 \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right].$$

TABLE 17.16 Estimated Random Effects Bivariate Probit Model

	Doctor		*Hospital*	
	Pooled	***Random Effects***	***Pooled***	***Random Effects***
Constant	−0.1243	−0.2976	−1.3385	−1.5855
	(0.05814)	(0.09650)	(0.07957)	(0.10853)
Female	0.3551	0.4548	0.1050	0.1280
	(0.01604)	(0.02857)	(0.02174)	(0.02954)
Age	0.01188	0.01983	0.00461	0.00496
	(0.000802)	(0.00130)	(0.001058)	(0.00139)
Income	−0.1337	−0.01059	0.04441	0.13358
	(0.04628)	(0.06488)	(0.05946)	(0.07728)
Kids	−0.1523	−0.1544	−0.01517	0.02155
	(0.01825)	(0.02692)	(0.02570)	(0.03211)
Education	−0.01484	−0.02573	−0.02191	−0.02444
	(0.003575)	(0.00612)	(0.005110)	(0.00675)
Married	0.07351	0.02876	−0.04789	−0.10504
	(0.02063)	(0.03167)	(0.02777)	(0.03547)
Corr($\varepsilon_1, \varepsilon_2$)	0.2981	0.1501	0.2981	0.1501
Corr(u_1, u_2)	0.0000	0.5382	0.0000	0.5382
Std. Dev. u	0.0000	0.2233	0.0000	0.6338
Std. Dev. ε	1.0000	1.0000	1.0000	1.0000

The application examined there involved a labor force participation model that was conditioned on an endogeous variable, the spouse's hours of work. In many cases, the endogenous variable in the equation is also binary. In the application we will examine next, the presence of a gender economics course in the economics curriculum at liberal arts colleges is conditioned on whether or not there is a women's studies program on the campus. The model in this case becomes

$$W^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \quad W = 1 \text{ if } W^* > 0, 0 \text{ otherwise},$$
$$y^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \gamma W + \varepsilon_2, \quad y = 1 \text{ if } y^* > 0, 0 \text{ otherwise},$$
$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} | \mathbf{x}_1, \mathbf{x}_2 \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right].$$

This model illustrates a number of interesting aspects of the bivariate probit model. Note that this model is qualitatively different from the bivariate probit model in (17-48); the first dependent variable, W, appears on the right-hand side of the second equation.[41] This model is a **recursive**, simultaneous-equations model. Surprisingly, the endogenous nature of one of the variables on the right-hand side of the second equation can be ignored in formulating the log-likelihood. [The model appears in Maddala (1983, p. 123).] We can establish this fact with the following (admittedly trivial) argument: The term that enters the log-likelihood is $P(y = 1, W = 1) = P(y = 1 \mid W = 1)P(W = 1)$. Given the model as stated, the marginal probability for W is just $\boldsymbol{\Phi}(\mathbf{x}_1'\boldsymbol{\beta}_1)$, whereas the conditional probability is $\boldsymbol{\Phi}_2(\cdots)/\boldsymbol{\Phi}(\mathbf{x}_1'\boldsymbol{\beta}_1)$. The product returns the bivariate normal probability

[41] Eisenberg and Rowe (2006) is another application of this model. In their study, they analyzed the joint (recursive) effect of W = veteran status on y, smoking behavior. The estimator they used was two-stage least squares and GMM.

we had earlier. The other three terms in the log-likelihood are derived similarly, which produces (Maddala's results with some sign changes):

$$P(y = 1,\ W = 1) = \Phi(\mathbf{x}_2'\boldsymbol{\beta}_2 + \gamma, \mathbf{x}_1'\boldsymbol{\beta}_1, \rho),$$
$$P(y = 1,\ W = 0) = \Phi(\mathbf{x}_2'\boldsymbol{\beta}_2, -\mathbf{x}_1'\boldsymbol{\beta}_1, -\rho),$$
$$P(y = 0,\ W = 1) = \Phi[-(\mathbf{x}_2'\boldsymbol{\beta}_2 + \boldsymbol{\gamma}), \mathbf{x}_1'\boldsymbol{\beta}_1, -\rho),$$
$$P(y = 0,\ W = 0) = \Phi(-\mathbf{x}_2'\boldsymbol{\beta}_2, -\mathbf{x}_1'\boldsymbol{\beta}_1, \rho).$$

These terms are exactly those of (17-48) that we obtain just by carrying W in the second equation with no special attention to its endogenous nature. We can ignore the simultaneity in this model and we cannot in the linear regression model because, in this instance, we are maximizing the log-likelihood, whereas in the linear regression case, we are manipulating certain sample moments that do not converge to the necessary population parameters in the presence of simultaneity.

Example 17.21 Gender Economics Courses at Liberal Arts Colleges

Burnett (1997) proposed the following bivariate probit model for the presence of a gender economics course in the curriculum of a liberal arts college:

$$\text{Prob}[G = 1, W = 1 \mid \mathbf{x}_G, \mathbf{x}_W] = \Phi_2(\mathbf{x}_G'\boldsymbol{\beta}_G + \boldsymbol{\gamma} W, \mathbf{x}_W'\boldsymbol{\beta}_W, \rho).$$

The dependent variables in the model are

G = presence of a gender economics course
W = presence of a women's studies program on the campus.

The independent variables in the model are

z_1 = constant term
z_2 = academic reputation of the college, coded 1 (best), 2, ... to 141
z_3 = size of the full-time economics faculty, a count
z_4 = percentage of the economics faculty that are women, proportion (0 to 1)
z_5 = religious affiliation of the college, 0 = no, 1 = yes
z_6 = percentage of the college faculty that are women, proportion (0 to 1)
z_7–z_{10} = regional dummy variables, South, Midwest, Northeast, West

The regressor vectors are

$$\mathbf{x}_G = z_1, z_2, z_3, z_4, z_5 \quad \text{(gender economics course equation)},$$
$$\mathbf{x}_W = z_2, z_5, z_6, z_7 - z_{10} \text{ (women's studies program equation)}.$$

Maximum likelihood estimates of the parameters of Burnett's model were computed by Greene (1998) using her sample of 132 liberal arts colleges; 31 of the schools offer gender economics, 58 have women's studies, and 29 have both. (See Appendix Table F17.1.) The estimated parameters are given in Table 17.17. Both bivariate probit and the single-equation estimates are given. The estimate of ρ is only 0.1359, with a standard error of 1.2359. The Wald statistic for the test of the hypothesis that ρ equals zero is $(0.1359/1.2539)^2 = 0.011753$. For a single restriction, the critical value from the chi-squared table is 3.84, so the hypothesis cannot be rejected. The likelihood ratio statistic for the same hypothesis is $2[-85.6317 - (-85.6458)] = 0.0282$, which leads to the same conclusion. The Lagrange multiplier statistic is 0.003807, which is consistent. This result might seem counterintuitive, given the setting. Surely "gender economics" and "women's studies" are highly correlated, but this finding does not contradict that proposition. The correlation coefficient measures the correlation between the disturbances in the equations, the omitted factors. That is, ρ measures (roughly) the correlation between the outcomes after the influence of the included factors is accounted

TABLE 17.17 Estimates of a Recursive Simultaneous Bivariate Probit Model (estimated standard errors in parentheses)

	Single Equation		*Bivariate Probit*	
Variable	*Coefficient*	*Standard Error*	*Coefficient*	*Standard Error*
Gender Economics Equation				
Constant	−1.4176	(0.8768)	−1.1911	(2.2155)
AcRep	−0.01143	(0.003610)	−0.01233	(0.007937)
WomStud	1.1095	(0.4699)	0.8835	(2.2603)
EconFac	0.06730	(0.05687)	0.06769	(0.06952)
PctWecon	2.5391	(0.8997)	2.5636	(1.0144)
Relig	−0.3482	(0.4212)	−0.3741	(0.5264)
Women's Studies Equation				
AcRep	−0.01957	(0.004117)	−0.01939	(0.005704)
PctWfac	1.9429	(0.9001)	1.8914	(0.8714)
Relig	−0.4494	(0.3072)	−0.4584	(0.3403)
South	1.3597	(0.5948)	1.3471	(0.6897)
West	2.3386	(0.6449)	2.3376	(0.8611)
North	1.8867	(0.5927)	1.9009	(0.8495)
Midwest	1.8248	(0.6595)	1.8070	(0.8952)
ρ	0.0000	(0.0000)	0.1359	(1.2539)
ln L	−85.6458		−85.6317	

for. Thus, the value 0.1359 measures the effect after the influence of women's studies is already accounted for. As discussed in the next paragraph, the proposition turns out to be right. The single most important determinant (at least within this model) of whether a gender economics course will be offered is indeed whether the college offers a women's studies program.

The marginal effects in this model are fairly involved, and as before, we can consider several different types. Consider, for example, z_2, academic reputation. There is a direct effect produced by its presence in the gender economics course equation. But there is also an indirect effect. Academic reputation enters the women's studies equation and, therefore, influences the probability that W equals one. Because W appears in the gender economics course equation, this effect is transmitted back to y. The total effect of academic reputation and, likewise, religious affiliation is the sum of these two parts. Consider first the gender economics variable, y. The conditional mean is

$$\begin{aligned} E[G \mid \mathbf{x}_G, \mathbf{x}_W] &= \text{Prob}[W = 1]E[G \mid W = 1, \mathbf{x}_G, \mathbf{x}_W] \\ &\quad + \text{Prob}[W = 0]E[G \mid W = 0, \mathbf{x}_G, \mathbf{x}_W] \\ &= \Phi_2(\mathbf{x}_G'\boldsymbol{\beta}_G + \gamma, \mathbf{x}_W'\boldsymbol{\beta}_W, \rho) + \Phi_2(\mathbf{x}_G'\boldsymbol{\beta}_G, -\mathbf{x}_W'\boldsymbol{\beta}_W, -\rho). \end{aligned}$$

Derivatives can be computed using our earlier results. We are also interested in the effect of religious affiliation. Because this variable is binary, simply differentiating the conditional mean function may not produce an accurate result. Instead, we would compute the conditional mean function with this variable set to one and then zero, and take the difference. Finally, what is the effect of the presence of a women's studies program on the probability that the college will offer a gender economics course? To compute this effect, we would compute

$$\text{Prob}[G = 1 \mid W = 1, \mathbf{x}_G, \mathbf{x}_W] - \text{Prob}[G = 1 \mid W = 0, \mathbf{x}_G, \mathbf{x}_W].$$

TABLE 17.18 Marginal Effects in Gender Economics Model

	Direct	*Indirect*	*Total*	*(Std. Error)*	*(Type of Variable, Mean)*
Gender Economics Equation					
AcRep	−0.002022	−0.001453	−0.003476	(0.001126)	(Continuous, 119.242)
PctWecon	+0.4491		+0.4491	(0.1568)	(Continuous, 0.24787)
EconFac	+0.01190		+0.1190	(0.01292)	(Continuous, 6.74242)
Relig	−0.06327	−0.02306	−0.08632	(0.08220)	(Binary, 0.57576)
WomStud	+0.1863		+0.1863	(0.0868)	(Endogenous, 0.43939)
PctWfac		+0.14434	+0.14434	(0.09051)	(Continuous, 0.35772)
Women's Studies Equation					
AcRep	−0.00780		−0.00780	(0.001654)	(Continuous, 119.242)
PctWfac	+0.77489		+0.77489	(0.3591)	(Continuous, 0.35772)
Relig	−0.17777		−0.17777	(0.11946)	(Binary, 0.57576)

In all cases, standard errors for the estimated marginal effects can be computed using the delta method or the method of Krinsky and Robb.

Table 17.18 presents the estimates of the marginal effects and some descriptive statistics for the data. The calculations were simplified slightly by using the restricted model with $\rho = 0$. Computations of the marginal effects still require the preceding decomposition, but they are simplified by the result that if ρ equals zero, then the bivariate probabilities factor into the products of the marginals. Numerically, the strongest effect appears to be exerted by the representation of women on the faculty; its coefficient of +0.4491 is by far the largest. This variable, however, cannot change by a full unit because it is a proportion. An increase of 1 percent in the presence of women on the faculty raises the probability by only +0.004, which is comparable in scale to the effect of academic reputation. The effect of women on the faculty is likewise fairly small, only 0.0013 per 1 percent change. As might have been expected, the single most important influence is the presence of a women's studies program, which increases the likelihood of a gender economics course by a full 0.1863. Of course, the raw data would have anticipated this result; of the 31 schools that offer a gender economics course, 29 also have a women's studies program and only two do not. Note finally that the effect of religious affiliation (whatever it is) is mostly direct.

17.5.6 ENDOGENOUS SAMPLING IN A BINARY CHOICE MODEL

We have encountered several instances of nonrandom sampling in the binary choice setting. In Section 17.3.6, we examined an application in credit scoring in which the balance in the sample of responses of the outcome variable, $C = 1$ for acceptance of an application and $C = 0$ for rejection, is different from the known proportions in the population. The sample was specifically skewed in favor of observations with $C = 1$ to enrich the data set. A second type of nonrandom sampling arose in the analysis of nonresponse/attrition in the GSOEP in Example 17.17. The data suggest that the observed sample is not random with respect to individuals' presence in the sample at different waves of the panel. The first of these represents selection specifically on an observable outcome—the observed dependent variable. We constructed a model for the second of these that relied on an assumption of selection on a set of certain observables—the variables that entered the probability weights. We will now examine a third form of nonrandom sample selection, based crucially on the unobservables in the two equations of a bivariate probit model.

We return to the banking application of Example 17.9. In that application, we examined a binary choice model,

$$\begin{aligned}\text{Prob}(\textit{Cardholder} = 1) &= \text{Prob}(C = 1 \mid \mathbf{x}) \\ &= \Phi(\beta_1 + \beta_2\, \textit{Age} + \beta_3\, \textit{Income} + \beta_4\, \textit{OwnRent} \\ &\quad + \beta_5\, \textit{Months at Current Address} \\ &\quad + \beta_6\, \textit{Self-Employed} \\ &\quad + \beta_7\, \textit{Number of Major Derogatory Reports} \\ &\quad + \beta_8\, \textit{Number of Minor Derogatory Reports}).\end{aligned}$$

From the point of view of the lender, cardholder status is not the interesting outcome in the credit history, default is. The more interesting equation describes Prob($\textit{Default} = 1 \mid \mathbf{z}, C = 1$). The natural approach, then, would be to construct a binary choice model for the interesting default variable using the historical data for a sample of cardholders. The problem with the approach is that the sample is not randomly drawn—applicants are screened with an eye specifically toward whether or not they seem likely to default. In this application, and in general, there are three economic agents, the credit scorer (e.g., Fair Isaacs), the lender, and the borrower. Each of them has latent characteristics in the equations that determine their behavior. It is these latent characteristics that drive, in part, the application/scoring process and, ultimately, the consumer behavior.

A model that can accommodate these features is (17-50),

$$\begin{aligned}&S^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \quad S = 1 \text{ if } \mathrm{S}^* > 0, 0 \text{ otherwise}, \\ &y^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2, \quad y = 1 \text{ if } \mathrm{y}^* > 0, 0 \text{ otherwise}, \\ &\begin{pmatrix}\varepsilon_1 \\ \varepsilon_2\end{pmatrix} \mid \mathbf{x}_1, \mathbf{x}_2 \sim N\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix}\right], \\ &(y,\ x_2) \text{ observed only when } S = 1,\end{aligned}$$

which contains an observation rule, $S = 1$, and a behavioral outcome, $y = 0$ or 1. The endogeneity of the sampling rule implies that

$$\text{Prob}(y = 1 \mid S = 1, \mathbf{x}_2) \neq \Phi(\mathbf{x}_2'\boldsymbol{\beta}).$$

From properties of the bivariate normal distribution, the appropriate probability is

$$\text{Prob}(y = 1 \mid S = 1, \mathbf{x}_1, \mathbf{x}_2) = \Phi\left[\frac{\mathbf{x}_2'\boldsymbol{\beta}_2 + \rho\mathbf{x}_1'\boldsymbol{\beta}_1}{\sqrt{1 - \rho^2}}\right].$$

If ρ is not zero, then in using the simple univariate probit model, we are omitting from our model any variables that are in $\mathbf{x}_1$ but not in $\mathbf{x}_2$, and in any case, the estimator is inconsistent by a factor $(1 - \rho^2)^{-1/2}$. To underscore the source of the bias, if ρ equals zero, the conditional probability returns to the model that would be estimated with the selected sample. Thus, the bias arises because of the correlation of (i.e., the selection on) the unobservables, ε_1 and ε_2. This model was employed by Wynand and van Praag (1981) in the first application of Heckman's (1979) sample selection model in a nonlinear

setting, to insurance purchases, by Boyes, Hoffman, and Lowe (1989) in a study of bank lending, by Greene (1992) to the credit card application begun in Example 17.9 and continued in Example 17.22, and hundreds of applications since. [Some discussion appears in Maddala (1983) as well.]

Given that the forms of the probabilities are known, the appropriate log-likelihood function for estimation of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and ρ is easily obtained. The log-likelihood must be constructed for the joint or the marginal probabilities, not the conditional ones. For the "selected observations," that is, $(y = 0, S = 1)$ or $(y = 1, S = 1)$, the relevant probability is simply

$$\text{Prob}(y = 0 \text{ or } 1 \mid S = 1) \times \text{Prob}(S = 1) = \Phi_2[(2y - 1)\mathbf{x}_2'\boldsymbol{\beta}_2, \mathbf{x}_1'\boldsymbol{\beta}_1, (2y - 1)\rho]$$

For the observations with $S = 0$, the probability that enters the likelihood function is simply $\text{Prob}(S = 0 \mid \mathbf{x}_1) = \Phi(-\mathbf{x}_1'\boldsymbol{\beta}_1)$. Estimation is then based on a simpler form of the bivariate probit log-likelihood that we examined in Section 17.5.1. Partial effects and postestimation analysis would follow the analysis for the bivariate probit model. The desired partial effects would differ by the application, whether one desires the partial effects from the conditional, joint, or marginal probability would vary. The necessary results are in Section 17.5.3.

Example 17.22 Cardholder Status and Default Behavior

In Example 17.9, we estimated a logit model for cardholder status,

$$\begin{aligned}\text{Prob}(\textit{Cardholder} = 1) &= \text{Prob}(C = 1 \mid \mathbf{x}) \\ &= \Phi(\beta_1 + \beta_2 \textit{Age} + \beta_3 \textit{Income} + \beta_4 \textit{OwnRent} \\ &\quad + \beta_5 \textit{ Current Address} + \beta_6 \textit{SelfEmployed} \\ &\quad + \beta_7 \textit{ Major Derogatory Reports} \\ &\quad + \beta_8 \textit{ Minor Derogatory Reports}),\end{aligned}$$

using a sample of 13,444 applications for a credit card. The complication in that example was that the sample was choice based. In the data set, 78.1 percent of the applicants are cardholders. In the population, at that time, the true proportion was roughly 23.2 percent, so the sample is substantially choice based on this variable. The sample was deliberately skewed in favor of cardholders for purposes of the original study [Greene (1992)]. The weights to be applied for the WESML estimator are $0.232/0.781 = 0.297$ for the observations with $C = 1$ and $0.768/0.219 = 3.507$ for observations with $C = 0$. Of the 13,444 applicants in the sample, 10,499 were accepted (given the credit cards). The "default rate" in the sample is 996/10,499 or 9.48 percent. This is slightly less than the population rate at the time, 10.3 percent. For purposes of a less complicated numerical example, we will ignore the choice-based sampling nature of the data set for the present. An orthodox treatment of both the selection issue and the choice-based sampling treatment is left for the exercises [and pursued in Greene (1992).]

We have formulated the cardholder equation so that it probably resembles the policy of credit scorers, both then and now. A major derogatory report results when a credit account that is being monitored by the credit reporting agency is more than 60 days late in payment. A minor derogatory report is generated when an account is 30 days delinquent. Derogatory reports are a major contributor to credit decisions. Contemporary credit processors such as Fair Isaacs place extremely heavy weight on the "credit score," a single variable that summarizes the credit history and credit-carrying capacity of an individual. We did not have access to credit scores at the time of this study. The selection equation

TABLE 17.19 Estimated Joint Cardholder and Default Probability Models

	Endogenous Sample Model			*Uncorrelated Equations*		
Variable/Equation	*Estimate*	*Standard Error*		*Estimate*	*Standard Error*	
		Cardholder Equation				
Constant	0.30516	0.04781	(6.38)	0.31783	0.04790	(6.63)
Age	0.00226	0.00145	(1.56)	0.00184	0.00146	(1.26)
Current Address	0.00091	0.00024	(3.80)	0.00095	0.00024	(3.94)
OwnRent	0.18758	0.03030	(6.19)	0.18233	0.03048	(5.98)
Income	0.02231	0.00093	(23.87)	0.02237	0.00093	(23.95)
SelfEmployed	−0.43015	0.05357	(−8.03)	−0.43625	0.05413	(−8.06)
Major Derogatory	−0.69598	0.01871	(−37.20)	−0.69912	0.01839	(−38.01)
Minor Derogatory	−0.04717	0.01825	(−2.58)	−0.04126	0.01829	(−2.26)
		Default Equation				
Constant	−0.96043	0.04728	(−20.32)	−0.81528	0.04104	(−19.86)
Dependents	0.04995	0.01415	(3.53)	0.04993	0.01442	(3.46)
Income	−0.01642	0.00122	(−13.41)	−0.01837	0.00119	(−15.41)
Expend/Income	−0.16918	0.14474	(−1.17)	−0.14172	0.14913	(−0.95)
Correlation	0.41947	0.11762	(3.57)	0.000	0.00000	(0)
Log Likelihood		−8660.90650			−8670.78831	

was given earlier. The default equation is a behavioral model. There is no obvious standard for this part of the model. We have used three variables, *Dependents*, the number of dependents in the household, *Income*, and *Exp_Income* which equals the ratio of the average credit card expenditure in the 12 months after the credit card was issued to average monthly income. Default status is measured for the first 12 months after the credit card was issued.

Estimation results are presented in Table 17.19. These are broadly consistent with the earlier results—the model with no correlation from Example 17.9 are repeated in Table 17.19. There are two tests we can employ for endogeneity of the selection. The estimate of ρ is 0.41947 with a standard error of 0.11762. The t ratio for the test that ρ equals zero is 3.57, by which we can reject the hypothesis. Alternatively, the likelihood ratio statistic based on the values in Table 17.19 is $2(8670.78831 - 8660.90650) = 19.76362$. This is larger than the critical value of 3.84, so the hypothesis of zero correlation is rejected. The results are as might be expected, with one counterintuitive result, that a larger credit burden, expenditure to income ratio, appears to be associated with lower default probabilities, though not significantly so.

17.5.7 A MULTIVARIATE PROBIT MODEL

In principle, a multivariate probit model would simply extend (17-48) to more than two outcome variables just by adding equations. The resulting equation system, again analogous to the seemingly unrelated regressions model, would be

$$
\begin{aligned}
&y_m^* = \mathbf{x}_m'\boldsymbol{\beta}_m + \varepsilon_m,\ y_m = 1 \text{ if } y_m^* > 0, 0 \text{ otherwise}, m = 1, \ldots, M,\\
&E[\varepsilon_m \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_M] = 0,\\
&\text{Var}[\varepsilon_m \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_M] = 1,\\
&\text{Cov}[\varepsilon_j, \varepsilon_m \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_M] = \rho_{jm},\\
&(\varepsilon_1, \ldots, \varepsilon_M) \sim \text{N}_M[\mathbf{0}, \mathbf{R}].
\end{aligned}
$$

The joint probabilities of the observed events, $[y_{i1}, y_{i2} \ldots, y_{iM} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iM}]$, $i = 1, \ldots, n$ that form the basis for the log-likelihood function are the M-variate normal probabilities,

$$L_i = \Phi_M(q_{i1}\mathbf{x}_{i1}'\beta_1, \ldots, q_{iM}\mathbf{x}_{iM}'\beta_M, \mathbf{R}^*),$$

where

$$q_{im} = 2y_{im} - 1,$$
$$\mathbf{R}^*_{jm} = q_{ij}q_{im}\rho_{jm}.$$

The practical obstacle to this extension is the evaluation of the M-variate normal integrals and their derivatives. Some progress has been made on using quadrature for trivariate integration (see Section 14.9.6.c), but existing results are not sufficient to allow accurate and efficient evaluation for more than two variables in a sample of even moderate size. However, given the speed of modern computers, simulation-based integration using the GHK simulator or simulated likelihood methods (see Chapter 15) do allow for estimation of relatively large models. We consider an application in Example 17.23.[42]

The **multivariate probit model** in another form presents a useful extension of the random effects probit model for panel data (Section 17.4.2). If the parameter vectors in all equations are constrained to be equal, we obtain what Bertschek and Lechner (1998) call the "panel probit model,"

$$y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it},\ y_{it} = 1 \text{ if } y_{it}^* > 0, 0 \text{ otherwise}, i = 1, \ldots, n, t = 1, \ldots, T,$$
$$(\varepsilon_{i1}, \ldots, \varepsilon_{iT}) \sim \mathrm{N}[\mathbf{0}, \mathbf{R}].$$

The Butler and Moffitt (1982) approach for this model (see Section 17.4.2) has proved useful in many applications. But, their underlying assumption that $\mathrm{Cov}[\varepsilon_{it}, \varepsilon_{is}] = \rho$ is a substantive restriction. By treating this structure as a multivariate probit model with the restriction that the coefficient vector be the same in every period, one can obtain a model with free correlations across periods.[43] Hyslop (1999), Bertschek and Lechner (1998), Greene (2004 and Example 17.16), and Cappellari and Jenkins (2006) are applications.

Example 17.23 A Multivariate Probit Model for Product Innovations

Bertschek and Lechner applied the panel probit model to an analysis of the product innovation activity of 1,270 German firms observed in five years, 1984–1988, in response to imports and foreign direct investment. [See Bertschek (1995).] The probit model to be estimated is based

[42] Studies that propose improved methods of simulating probabilities include Pakes and Pollard (1989) and especially Börsch-Supan and Hajivassiliou (1993), Geweke (1989), and Keane (1994). A symposium in the November 1994 issue of *Review of Economics and Statistics* presents discussion of numerous issues in specification and estimation of models based on simulation of probabilities. Applications that employ simulation techniques for evaluation of multivariate normal integrals are now fairly numerous. See, for example, Hyslop (1999) (Example 17.15) which applies the technique to a panel data application with $T = 7$. Example 17.23 develops a five-variate application.

[43] By assuming the coefficient vectors are the same in all periods, we actually obviate the normalization that the diagonal elements of $\mathbf{R}$ are all equal to one as well. The restriction identifies $T - 1$ relative variances $\rho_{tt} = \sigma_T^2/\sigma_T^2$. This aspect is examined in Greene (2004).

TABLE 17.20 Estimated Pooled Probit Model

		Estimated Standard Errors				Marginal Effects		
Variable	*Estimate*[a]	*SE(1)*[b]	*SE(2)*[c]	*SE(3)*[d]	*SE(4)*[e]	*Partial*	*Std. Err.*	t *ratio*
Constant	−1.960	0.239	0.377	0.230	0.373	—	−−	—
log Sales	0.177	0.0250	0.0375	0.0222	0.0358	0.0683[f]	0.0138	4.96
Rel Size	1.072	0.206	0.306	0.142	0.269	0.413[f]	0.103	4.01
Imports	1.134	0.153	0.246	0.151	0.243	0.437[f]	0.0938	4.66
FDI	2.853	0.467	0.679	0.402	0.642	1.099[f]	0.247	4.44
Prod.	−2.341	1.114	1.300	0.715	1.115	−0.902[f]	0.429	−2.10
Raw Mtl	−0.279	0.0966	0.133	0.0807	0.126	−0.110[g]	0.0503	−2.18
Inv Good	0.188	0.0404	0.0630	0.0392	0.0628	0.0723[g]	0.0241	3.00

[a]Recomputed. Only two digits were reported in the earlier paper.
[b]Obtained from results in Bertschek and Lechner, Table 9.
[c]Based on the Avery et al. (1983) GMM estimator.
[d]Square roots of the diagonals of the negative inverse of the Hessian
[e]Based on the cluster estimator.
[f]Coefficient scaled by the density evaluated at the sample means
[g]Computed as the difference in the fitted probability with the dummy variable equal to one, then zero.

on the latent regression

$$y_{it}^* = \beta_1 + \sum_{k=2}^{8} x_{k,it}\beta_k + \varepsilon_{it},\ y_{it} = 1(y_{it}^* > 0),\ i = 1, \ldots, 1,270,\ t = 1984, \ldots, 1988,$$

where

y_{it} = 1 if a product innovation was realized by firm i in year t, 0 otherwise
$x_{2,it}$ = Log of industry sales in DM
$x_{3,it}$ = Import share = ratio of industry imports to (industry sales plus imports)
$x_{4,it}$ = Relative firm size = ratio of employment in business unit to employment in the industry (times 30)
$x_{5,it}$ = FDI share = Ratio of industry foreign direct investment to (industry sales plus imports)
$x_{6,it}$ = Productivity = Ratio of industry value added to industry employment
$x_{7,it}$ = Raw materials sector = 1 if the firm is in this sector
$x_{8,it}$ = Investment goods sector = 1 if the firm is in this sector

The coefficients on import share (β_3) and FDI share (β_5) were of particular interest. The objectives of the study were the empirical investigation of innovation and the methodological development of an estimator that could obviate computing the five-variate normal probabilities necessary for a full maximum likelihood estimation of the model.

Table 17.20 presents the single-equation, pooled probit model estimates.[44] Given the structure of the model, the parameter vector could be estimated consistently with any single period's data, Hence, pooling the observations, which produces a mixture of the estimators, will also be consistent. Given the panel data nature of the data set, however, the conventional standard errors from the pooled estimator are dubious. Because the marginal distribution

[44]We are grateful to the authors of this study who have generously loaned us their data for our continued analysis. The data are proprietary and cannot be made publicly available, unlike the other data sets used in our examples.

TABLE 17.21 Estimated Constrained Multivariate Probit Model (estimated standard errors in parentheses)

Coefficients	*Full Maximum Likelihood Using GHK Simulator*	*Random Effects* $\rho = 0.578$ (0.0189)
Constant	−1.797** (0.341)	−2.839 (0.534)
log Sales	0.154** (0.0334)	0.245 (0.0523)
Relative size	0.953** (0.160)	1.522 (0.259)
Imports	1.155** (0.228)	1.779 (0.360)
FDI	2.426** (0.573)	3.652 (0.870)
Productivity	−1.578 (1.216)	−2.307 (1.911)
Raw material	−0.292** (0.130)	−0.477 (0.202)
Investment goods	0.224** (0.0605)	0.331 (0.0952)
log-likelihood	−3522.85	−3535.55
	Estimated Correlations	
1984, 1985	0.460** (0.0301)	
1984, 1986	0.599** (0.0323)	
1985, 1986	0.643** (0.0308)	
1984, 1987	0.540** (0.0308)	
1985, 1987	0.546** (0.0348)	
1986, 1987	0.610** (0.0322)	
1984, 1988	0.483** (0.0364)	
1985, 1988	0.446** (0.0380)	
1986, 1988	0.524** (0.0355)	
1987, 1988	0.605** (0.0325)	

*Indicates significant at 95 percent level,
** indicates significant at 99 percent level based on a two-tailed test.

will produce a consistent estimator of the parameter vector, this is a case in which the cluster estimator (see Section 14.8.4) provides an appropriate asymptotic covariance matrix. Note that the standard errors in column SE(4) of the table are considerably higher than the uncorrected ones in columns 1–3.

The pooled estimator is consistent, so the further development of the estimator is a matter of (1) obtaining a more efficient estimator of $\boldsymbol{\beta}$ and (2) computing estimates of the cross-period correlation coefficients. The FIML estimates of the model can be computed using the GHK simulator.[45] The FIML estimates and the random effects model using the Butler and Moffit (1982) quadrature method are reported in Table 17.21. The correlations reported are based on the FIML estimates. Also noteworthy in Table 17.21 is the divergence of the random effects estimates from the FIML estimates. The log-likelihood function is −3535.55 for the random effects model and −3522.85 for the unrestricted model. The chi-squared statistic for the nine restrictions of the equicorrelation model is 25.4. The critical value from the chi-squared table for nine degrees of freedom is 16.9 for 95 percent and 21.7 for 99 percent significance, so the hypothesis of the random effects model would be rejected.

17.6 SUMMARY AND CONCLUSIONS

This chapter has surveyed a large range of techniques for modeling a binary choice variable. The model for choice between two outcomes provides the framework for a

[45]The full computation required about one hour of computing time. Computation of the single-equation (pooled) estimators required only about 1/100 of the time reported by the authors for the same models, which suggests that the evolution of computing technology may play a significant role in advancing the FIML estimators.

large proportion of the analysis of microeconomic data. Thus, we have given a very large amount of space to this model in its own right. In addition, many issues in model specification and estimation that appear in more elaborate settings, such as those we will examine in the next chapter, can be formulated as extensions of the binary choice model of this chapter. Binary choice modeling provides a convenient point to study endogeneity in a nonlinear model, issues of nonresponse in panel data sets, and general problems of estimation and inference with longitudinal data. The binary probit model in particular has provided the laboratory case for theoretical econometricians such as those who have developed methods of bias reduction for the fixed effects estimator in dynamic nonlinear models.

We began the analysis with the fundamental parametric probit and logit models for binary choice. Estimation and inference issues such as the computation of appropriate covariance matrices for estimators and partial effects are considered here. We then examined familiar issues in modeling, including goodness of fit and specification issues such as the distributional assumption, heteroscedasticity and missing variables. As in other modeling settings, endogeneity of some right-hand variables presents a substantial complication in the estimation and use of nonlinear models such as the probit model. We examined the problem of endogenous right-hand-side variables, and in two applications, problems of endogenous sampling. The analysis of binary choice with panel data provides a setting to examine a large range of issues that reappear in other applications. We reconsidered the familiar pooled, fixed and random effects estimator estimators, and found that much of the wisdom obtained in the linear case does not carry over to the nonlinear case. The incidental parameters problem, in particular, motivates a considerable amount of effort to reconstruct the estimators of binary choice models. Finally, we considered some multivariate extensions of the probit model. As before, the models are useful in their own right. Once again, they also provide a convenient setting in which to examine broader issues, such as more detailed models of endogeneity nonrandom sampling, and computation requiring simulation.

Chapter 18 will continue the analysis of discrete choice models with three frameworks: unordered multinomial choice, ordered choice, and models for count data. Most of the estimation and specification issues we have examined in this chapter will reappear in these settings.

Key Terms and Concepts

- Attributes
- Attrition bias
- Average partial effect
- Binary choice model
- Bivariate probit
- Butler and Moffitt method
- Characteristics
- Choice-based sampling
- Chow test
- Complementary log log model
- Conditional likelihood function
- Control function
- Event count
- Fixed effects model
- Generalized residual
- Goodness of fit measure
- Gumbel model
- Heterogeneity
- Heteroscedasticity
- Incidental parameters problem
- Index function model
- Initial conditions
- Interaction effect
- Inverse probability weighted (IPW)
- Lagrange multiplier test
- Latent regression
- Likelihood equations
- Likelihood ratio test

- Linear probability model
- Logit
- Marginal effects
- Maximum likelihood
- Maximum simulated likelihood (MSL)
- Method of scoring
- Microeconometrics
- Minimal sufficient statistic
- Multinomial choice
- Multivariate probit model
- Nonresponse bias
- Ordered choice model
- Persistence
- Probit
- Quadrature
- Qualitative response (QR)
- Quasi-maximum likelihood estimator (QMLE)
- Random effects model
- Random parameters logit model
- Random utility model
- Recursive model
- Robust covariance estimation
- Sample selection bias
- Selection on unobservables
- State dependence
- Tetrachoric correlation
- Unbalanced sample

Exercises

1. A binomial probability model is to be based on the following index function model:

$$y^* = \alpha + \boldsymbol{\beta} d + \varepsilon,$$
$$y = 1, \text{ if } y^* > 0,$$
$$y = 0 \text{ otherwise.}$$

The only regressor, d, is a dummy variable. The data consist of 100 observations that have the following:

		y	
		0	1
d	0	24	28
	1	32	16

Obtain the maximum likelihood estimators of α and β, and estimate the asymptotic standard errors of your estimates. Test the hypothesis that β equals zero by using a Wald test (asymptotic t test) and a likelihood ratio test. Use the probit model and then repeat, using the logit model. Do your results change? (*Hint*: Formulate the log-likelihood in terms of α and $\delta = \alpha + \beta$.)

2. Suppose that a linear probability model is to be fit to a set of observations on a dependent variable y that takes values zero and one, and a single regressor x that varies continuously across observations. Obtain the exact expressions for the least squares slope in the regression in terms of the mean(s) and variance of x, and interpret the result.
3. Given the data set

y	1	0	0	1	1	0	0	1	1	1
x	9	2	5	4	6	7	3	5	2	6,

estimate a probit model and test the hypothesis that x is not influential in determining the probability that y equals one.

4. Construct the Lagrange multiplier statistic for testing the hypothesis that all the slopes (but not the constant term) equal zero in the binomial logit model. Prove that the Lagrange multiplier statistic is nR^2 in the regression of $(y_i = p)$ on the x's, where p is the sample proportion of 1's.

5. The following hypothetical data give the participation rates in a particular type of recycling program and the number of trucks purchased for collection by 10 towns in a small mid-Atlantic state:

Town	1	2	3	4	5	6	7	8	9	10
Trucks	160	250	170	365	210	206	203	305	270	340
Participation%	11	74	8	87	62	83	48	84	71	79

The town of Eleven is contemplating initiating a recycling program but wishes to achieve a 95 percent rate of participation. Using a probit model for your analysis,
   a. How many trucks would the town expect to have to purchase to achieve its goal? (*Hint*: You can form the log-likelihood by replacing y_i with the participation rate (e.g., 0.11 for observation 1) and $(1 - y_i)$ with 1—the rate in (17-22).
   b. If trucks cost $20,000 each, then is a goal of 90 percent reachable within a budget of $6.5 million? (That is, should they *expect* to reach the goal?)
   c. According to your model, what is the marginal value of the 301st truck in terms of the increase in the percentage participation?
6. A data set consists of $n = n_1 + n_2 + n_3$ observations on y and x. For the first n_1 observations, $y = 1$ and $x = 1$. For the next n_2 observations, $y = 0$ and $x = 1$. For the last n_3 observations, $y = 0$ and $x = 0$. Prove that neither (17-18) nor (17-20) has a solution.
7. Prove (17-30).
8. In the panel data models estimated in Section 17.4, neither the logit nor the probit model provides a framework for applying a Hausman test to determine whether fixed or random effects is preferred. Explain. (*Hint*: Unlike our application in the linear model, the incidental parameters problem persists here.)

Applications

1. Appendix Table F17.1 provides Fair's (1978) *Redbook* survey on extramarital affairs. The data are described in Application 1 at the end of Chapter 18 and in Appendix F. The variables in the data set are as follows:

id = an identification number
C = constant, value = 1
yrb = a constructed measure of time spent in extramarital affairs
$v1$ = a rating of the marriage, coded 1 to 4
$v2$ = age, in years, aggregated
$v3$ = number of years married
$v4$ = number of children, top coded at 5
$v5$ = religiosity, 1 to 4, 1 = not, 4 = very
$v6$ = education, coded 9, 12, 14, 16, 17, 20,
$v7$ = occupation
$v8$ = husband's occupation

and three other variables that are not used. The sample contains a survey of 6,366 married women, conducted by *Redbook* magazine. For this exercise, we will analyze,

first, the binary variable

$$A = 1 \text{ if } yrb > 0, 0 \text{ otherwise.}$$

The regressors of interest are v_1 to v_8; however, not necessarily all of them belong in your model. Use these data to build a binary choice model for A. Report all computed results for the model. Compute the marginal effects for the variables you choose. Compare the results you obtain for a probit model to those for a logit model. Are there any substantial differences in the results for the two models?