17

BINARY OUTCOMES AND DISCRETE CHOICES

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, 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 has 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 used in this area of study are inherently (and intrinsically) nonlinear. We have developed some of the elements of nonlinear modeling in Chapters 7 and 14. Those elements are combined in whole in the study of discrete choices. This chapter will focus on binary choices, where *the model* is the probability of an event. Many general treatments of nonlinear modeling in econometrics, in fact, focus on only this segment of the field. This is reasonable. Nearly the full set of results used more broadly, for specification, estimation, inference, and analysis can be developed and understood in this particular application. We will take that approach here. Several of the parts of nonlinear modeling will be developed in detail in this chapter, then invoked or extended in straightforward ways in the chapters to follow.

The models that are analyzed in this and Chapter 18 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 his or her underlying preferences by the choice that he or she makes. The choice(s) made will be affected by observable influences—this is, for example, 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]

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

This chapter and Chapter 18 will describe four broad frameworks for analysis. The first is the simplest:

Binary Choice: The individual faces two 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 mathematical convenience. This chapter will present a lengthy survey of models and methods for binary choices.

The binary choice case naturally extends to cases of more than two outcomes. For one example, in our our travel mode case, the individual choosing private transport might choose between private transport as driver and private transport as passenger, or public transport by train or by bus. Such multinomial (many named) choices are *unordered*. Another case is one that is a constant staple of the online experience. Instead of being asked a binary choice, "Did you like our service?", the hapless surfer will be asked an *ordered* multinomial choice, "On a scale from 1 to 5, how much did you like our service?"

Multinomial Choice: The individual chooses among more than two choices, once again, making the choice that provides the greatest utility. 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 again 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 labeled 0, 1, 2, 3, 4 to indicate the strength of preferences 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" only in a qualitative sense, 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 and in how changes in those attributes can help explain variation in 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 explain the choice. Chapter 18 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, the number of vehicles in a household's capital stock, 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 the occurrence rate of a disease in a population or the number of defects per unit of time in a production process. In these settings, we will be doing a more familiar sort of regression modeling. However, the models will still be constructed specifically to accommodate the discrete (and nonnegative) nature of the observed response variable and the modeling of probabilities of occurrences of events rather than some measure of the events themselves.

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 include several applications and 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. In general, because the outcome variable in the first three of these four cases is merely the name of an event, not the event itself, linear regression will be an inappropriate approach. In most cases, the method of estimation is maximum likelihood.[2] 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

[2] In the binary choice case, it is possible arbitrarily to assign two numerical values to the outcomes, typically 0 and 1, and "linearly regress" this constructed variable on the covariates. We will examine this strategy at some length with an eye to what information it reveals. The strategy would make little sense in the multinomial choice cases. Since the count data case is, in fact, a quantitative regression setting, the comparison of a linear regression approach to the intrinsically nonlinear regression approach is worth a close look.

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 Greene and Hensher (2010) provide, among many others, general surveys of discrete choice models and methods.[3]

17.2 MODELS FOR BINARY OUTCOMES

For purposes of studying individual behavior, we will construct models that link a 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:

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

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. Henceforth, we will use the shorthand,

$$\text{Prob}(Y = 1 \mid \mathbf{x}) = \text{Probability that event of interest occurs} \mid \mathbf{x},$$

and, naturally, $\text{Prob}(Y = 0 \mid \mathbf{x}) = [1 - \text{Prob}(Y = 1 \mid \mathbf{x})]$ is the probability that the event does not occur.

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

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

where *earnings* is *hourly wage* times *hours worked, education* is measured in years of schooling, and *kids* is a binary variable that 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 is the outcome of a market process whereby the demanders of labor services are willing to offer a wage based on expected marginal product, and individuals themselves make a decision whether or not to accept the offer depending on whether it exceeds 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 $y = 1$ (*in the labor force,* 428 observations) or 0 (*not in the labor force*, 325 observations). For example, we would be interested how and how significantly the presence of children in the household (*kids*) affects the labor force participation.

Models for explaining a binary 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

[3] There are dozens of book-length surveys of discrete choice models. Two others that are heavily oriented to an application of these methods are Train (2009) and Hensher, Rose, and Greene (2015).

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 dictates 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 ($Y = 0$). The event here is still qualitative, but now it is constructed as an indicator of a censoring (or not) of an underlying continuous variable, in this case, unobserved true demand. It will generally turn out that the models and techniques used in both cases (and, indeed, the underlying structure) are the same. Nonetheless, it is useful to examine both of them.

17.2.1 RANDOM UTILITY

An interpretation of data on individual choices is provided by a 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 underlying unobservable utilities. Hence, the observed indicator equals 1 if $U_a > U_b$ and 0 if $U_a \leq U_b$. If we define, $U = U_a - U_b$, then $Y = \mathbf{1}(U > 0)$ [where **1** (condition) equals 1 if condition is true and 0 if it is false]. This is precisely the same as the censoring case noted earlier.

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, such as party affiliation.

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 $U_a > U_b$. Because the outcome is ultimately driven by the random elements in the utility functions, we have

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

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:[4] 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.

17.2.2 THE 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 perceived 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 (there is a constant term in $\mathbf{x}$) and has either a logistic distribution 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

[4] 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 17.13 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, underlies Heckman (1979) and 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. The Di Maria et al. (2010) study of the light bulb puzzle in Example 17.4 is another example of an elaborate structural random utility model that produces a binary outcome. This application is also closely related to Rubin's (1974, 1978) potential outcomes model discussed in Section 8.5.

the net benefit of the purchase (i.e., net utility), only whether it is made or not. Therefore, our observation is

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

The statement in (17-3) is conveniently denoted $y = \mathbf{1}\,(y* > 0)$. In this formulation, $\mathbf{x}'\boldsymbol{\beta}$ is called the index function. The assumption of known variance of ε is an innocent normalization. Note, once again, the outcomes 0 and 1 are merely labels of the event. Now, suppose the variance of ε is, instead, an unrestricted parameter σ^2. The **latent regression** will be $y^* = \mathbf{x}'\boldsymbol{\beta} + \sigma\varepsilon*$, where now $\varepsilon*$ has variance one. 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."[5] The assumption of zero for the threshold in (17-4) is likewise innocent if the model contains a constant term (and not if it does not).[6] Let a be a 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

$$\text{Prob}(y^* > a\,|\,\mathbf{x}) = \text{Prob}(\alpha + \mathbf{x}'\boldsymbol{\beta} + \varepsilon > a\,|\,\mathbf{x}) = \text{Prob}[(\alpha - a) + \mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0\,|\,\mathbf{x}]. \qquad \textbf{(17-3)}$$

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 (17-4). The choice of zero is a normalization with no significance. With the two normalizations, then,

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

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

$$\text{Prob}(y^* > 0\,|\,\mathbf{x}) = \text{Prob}(\varepsilon < \mathbf{x}'\boldsymbol{\beta}\,|\,\mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta}), \qquad \textbf{(17-5)}$$

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 PROBABILITY

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

$$\begin{aligned} \text{Prob}(Y = 1\,|\,\mathbf{x}) &= F(\mathbf{x}, \boldsymbol{\beta}) \\ \text{Prob}(Y = 0\,|\,\mathbf{x}) &= 1 - F(\mathbf{x}, \boldsymbol{\beta}). \end{aligned} \qquad \textbf{(17-6)}$$

[5] In some treatments [e.g., Horowitz (1990) and Lewbel (2000)] it is more convenient to normalize one of the elements of $\boldsymbol{\beta}$ to equal 1 and leave σ free to vary. In the end, only $\boldsymbol{\beta}/\sigma$ is estimated, so this is inconsequential.

[6] Unless there is some compelling reason, binary choice models should not be estimated without constant terms.

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 partial effect of having children in the household on the probability of labor force participation. The challenge at this point is to devise a suitable specification for the right-hand side of the equation.

Our requirement is a model that will produce predictions consistent with the underlying theory in (17-5) and (17-6). For a given regressor vector, we would expect

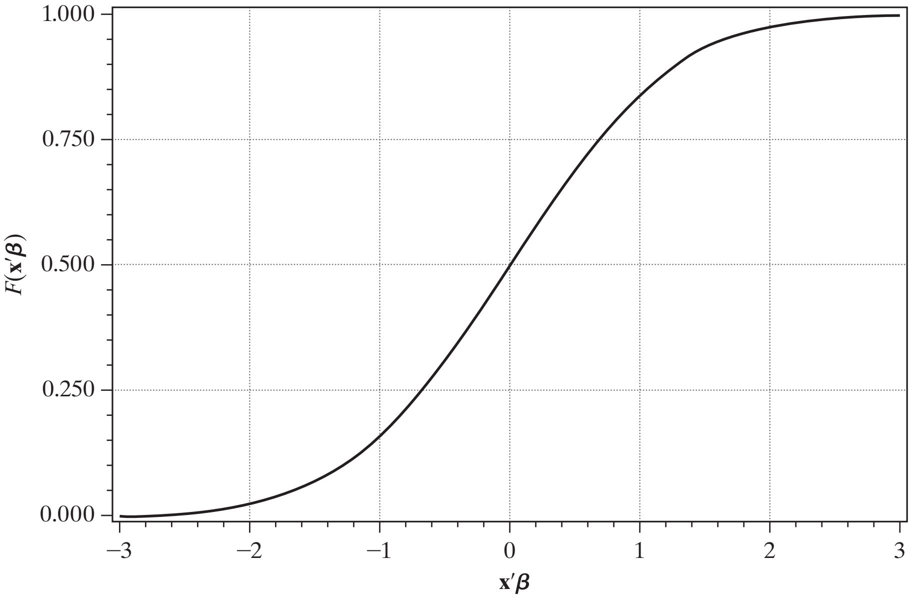
$$0 \leq \text{Prob}(Y = 1|\mathbf{x}) \leq 1, \tag{17-7}$$

$$\begin{aligned} \lim_{\mathbf{x}'\boldsymbol{\beta}\to-\infty} \text{Prob}(Y = 1|\mathbf{x}) &= 0, \\ \lim_{\mathbf{x}'\boldsymbol{\beta}\to+\infty} \text{Prob}(Y = 1|\mathbf{x}) &= 1. \end{aligned} \tag{17-8}$$

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**,[7]

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

FIGURE 17.1 Model for a Probability.



[7] The term "probit" derives from "probability unit," in turn from the use of inverse normal probability units in bioassay. See Finney (1971) and Greene and Hensher (2010, Ch. 4).

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

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

has also been used in many applications. We shall use the notation $\Lambda(.)$ to indicate the logistic distribution function. For this case, the density is $\Lambda(t)[1 - \Lambda(t)]$. This model is called the **logit** model for reasons we shall discuss below. Both of these distributions have the familiar bell shape of symmetric distributions and sigmoid shape shown in Figure 17.1. Other models which do not assume symmetry, such as the **Gumbel model** or Type I extreme value model,

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

complementary log log model,

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

and the Burr model,[8]

$$\text{Prob}(Y = 1|\mathbf{x}) = \left[\frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}\right]^{\gamma} = [\Lambda(\mathbf{x}'\boldsymbol{\beta})]^{\gamma},$$

have also been employed. Still other distributions have been suggested,[9] but the probit and logit models are by far 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.) For intermediate values of $\mathbf{x}'\boldsymbol{\beta}$, the two distributions tend to give very 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 might 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. As seen in the following example, the symmetric and asymmetric distributions can give somewhat different results, and here, the guidance on how to choose is unfortunately sparse. On the other hand, for estimation of the quantities usually of interest (partial effects), in the sample sizes typical in modern

[8] Or Scobit model for a skewed logit model; see Nagler (1994).

[9] See, for example, Maddala (1983, pp. 27–32), Aldrich and Nelson (1984), and *Stata* (2014).

research, it turns out that the different functional forms tend to give comfortably similar results. The choice of which $F(.)$ to use is ultimately less important than the choice of $\mathbf{x}$ and $\mathbf{x}'\boldsymbol{\beta}$. We will examine this proposition in more detail below.

17.2.4 PARTIAL EFFECTS IN BINARY CHOICE MODELS

Most analyses will be directed at examining the relationships between the covariates, $\mathbf{x}$, and the probability of the event, $\text{Prob}(Y = 1|\mathbf{x}) = F(y|\mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta})$, typically, the partial effects. Whatever distribution is used, it is important to note that the parameters of the model ($\boldsymbol{\beta}$), like those of any nonlinear model, are not necessarily the partial effects we are accustomed to analyzing. In general, via the chain rule,

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

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

$$\frac{\partial F(y|\mathbf{x})}{\partial \mathbf{x}} = \phi(\mathbf{x}'\boldsymbol{\beta}_{probit}) \times \boldsymbol{\beta}_{probit}.$$

For the logistic distribution,

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

so, in the logit model,

$$\frac{\partial F(y|\mathbf{x})}{\partial \mathbf{x}} = \Lambda(\mathbf{x}'\boldsymbol{\beta}_{logit})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta}_{logit})]\boldsymbol{\beta}_{logit}.$$

These values will vary with the values of $\mathbf{x}$. In index function models generally, the set of partial effects is a multiple of the coefficient vector.

As we will observe below in several applications, a common empirical regularity for estimates of probit and logit models is $\hat{\boldsymbol{\beta}}_{logit} \approx 1.6\hat{\boldsymbol{\beta}}_{probit}$. This might suggest quite a large difference between the two models, however, that would be misleading. As a general result, the partial effects produced by these two (and other) models will be nearly the same. Near the middle of the range of the probabilities, where $F(\mathbf{x}'\boldsymbol{\beta})$ is roughly 0.5, the logistic partial effects will be roughly $0.5(1 - 0.5)\boldsymbol{\beta}_{logit}$ while the probit partial effects will be roughly $0.4\boldsymbol{\beta}_{probit}$ (where 0.4 is the normal density at the point where the cdf equals 0.5). If the two partial effects are to be the same, then $0.25\boldsymbol{\beta}_{logit} = 0.4\boldsymbol{\beta}_{probit}$ or $\boldsymbol{\beta}_{logit} = 1.6\boldsymbol{\beta}_{probit}$. Observed estimates will vary around this general result. An example is shown in Table 17.1.

For computing partial effects one can evaluate the expressions at the sample means of the data, producing the partial effects at the averages (PEA),

$$PEA = \hat{\boldsymbol{\gamma}}(\overline{\mathbf{x}}) = f(\overline{\mathbf{x}}'\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}.$$

The means of the data do not always produce a realistic scenario for the computation. For example, the mean gender of 0.5 does not correspond to any individual in the sample. It is more common to evaluate the partial effects at every actual observation and use the sample average of the individual partial effects, producing the **average partial effects (APE)**. The desired computation would be

$$APE = \hat{\bar{\boldsymbol{\gamma}}} = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}. \tag{17-12}$$

It is usually, the "average partial effect," that is, the expected value of the partial effect, that is actually of interest. Let $\boldsymbol{\gamma}^0$ denote the population parameter. Then,

$$APE^0 = \boldsymbol{\gamma}^0 = E_x\left[\frac{\partial E[y\,|\,\mathbf{x}]}{\partial \mathbf{x}}\right]. \tag{17-13}$$

One might wonder whether the APE produces a different answer from the PEA. It is tempting to suggest that the difference is a small sample effect, but it is not, at least not entirely. Assume the parameters are known, and let the average partial effect for variable x_k be

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

We will 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

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

where $\Delta(\bar{\mathbf{x}})$ is the remaining higher-order terms. The first of the four terms is the partial effect at the sample means. The second term is zero. The third is an average of functions of the variances and covariances of the data and the curvature of the probability function at the means. The final term is the remainder. Little can be said to characterize these two terms in any particular sample. In applications, the difference is usually relatively small.

Another complication for computing partial effects in a nonlinear 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. It is not appropriate to apply (17-12) for the effect of a change in a dummy variable, or a change of state. The appropriate partial effect for a binary independent variable, say, d, would be

$$\text{PEA} = \text{Prob}[Y = 1\,|\,\bar{\mathbf{x}}_{(d)}, d = 1] - \text{Prob}[Y = 1\,|\,\bar{\mathbf{x}}_{(d)}, d = 0] \tag{17-14}$$

or

$$\text{APE} = \frac{1}{n}\sum_{i=1}^{n}[\text{Prob}(Y = 1\,|\,\mathbf{x}_{i,(d)}, d_i = 1) - \text{Prob}(Y = 1\,|\,\mathbf{x}_{i,(d)}, d_i = 0)],$$

where d denotes the other variables in the model excluding the dummy variable in question. 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 average difference in the two probabilities for the probit model is 0.374, whereas the derivative approximation is $0.222 \times 1.426 = 0.317$. In a larger sample, the differences are often very small. Nonetheless, the difference in the probabilities is the preferred computation, and is automated in standard software.

If the dummy variable in the choice model is a treatment as PSI is in the example below, then the APE would estimate the average treatment, ATE, for the population. But the average treatment on the treated, ATET, would require a change in the computation. If the treatment were exogenous (e.g., if students were carefully randomly assigned to PSI), then computing the APE over the subsample with $d_i = 1$, would be an appropriate estimator.[10] Any difference between ATE and ATET would then be attributable to systematic differences in $\mathbf{x}|d = 1$ and $\boldsymbol{x}|(d = 0 \; or \; d = 1)$. If the treatment were endogenous, then neither APE nor APE$|d = 1$ would be an appropriate estimator—indeed, the model itself would have to be extended. We will treat this case in Section 17.6.

17.2.5 ODDS RATIOS IN LOGIT MODELS

The odds *in favor* of an event is the ratio Prob($Y = 1$)/Prob($Y = 0$). For the logit model—the result is not meaningful for the other models considered—the odds "in favor of $Y = 1$" are

$$Odds = \frac{\text{Prob}(Y = 1|\mathbf{x})}{\text{Prob}(Y = 0|\mathbf{x})} = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})/[1 + \exp(\mathbf{x}'\boldsymbol{\beta})]}{1/[1 + \exp(\mathbf{x}'\boldsymbol{\beta})]} = \exp(\mathbf{x}'\boldsymbol{\beta}).$$

Consider the effect on the odds of the change of a dummy variable, d,

$$Odds\ Ratio = \frac{Odds(\mathbf{x}, d = 1)}{Odds(\mathbf{x}, d = 0)} = \frac{\left[\dfrac{\exp(\mathbf{x}'\boldsymbol{\beta} + \delta \times 1)/[1 + \exp(\mathbf{x}'\boldsymbol{\beta} + \delta \times 1)]}{1/[1 + \exp(\mathbf{x}'\boldsymbol{\beta} + \delta \times 1)]}\right]}{\left[\dfrac{\exp(\mathbf{x}'\boldsymbol{\beta} + \delta \times 0)/[1 + \exp(\mathbf{x}'\boldsymbol{\beta} + \delta \times 0)]}{1/[1 + \exp(\mathbf{x}'\boldsymbol{\beta} + \delta \times 0)]}\right]} = \exp(\delta).$$

Therefore, the change in the odds when a variable changes by one unit somewhat resembles a partial effect, though in fact it is not a derivative. "Odds ratios" are reported in many studies that are based on logit models. When the experiment of changing the variable in question, x_k, by one unit is meaningful, $\exp(\beta_k)$ for the respective coefficient reports the multiplicative change in the ratio. The proportional change would be $\exp(\delta) - 1$. [Received studies always report $\exp(\delta)$, not $\exp(\delta) - 1$.] If the experiment of a change in one unit is not meaningful, the odds ratio, like the simple partial effect, could be misleading. Note, in Example 17.8 (Table17.5) below, we have computed a partial effect for income of roughly -0.03. However, a change in income of a full unit in these data is not a meaningful experiment—the full range of values is about 1.0–3.0. The more useful calculation for a variable x_k is $\partial\text{Prob}(Y = 1|\mathbf{x})/\partial x_k \times dx_k$. In Example 17.8, for the income variable, $\text{d}x_k = 0.1$ would be more informative. A similar computation would be appropriate for the odds ratios, though it is unclear how that might be constructed independently of the specific change for a specific variable, in which case, the partial effect (or elasticity) might be more straightforward. The odds ratio is meaningful for a dummy variable, however. We examine an application in Example 17.11.

[10] Use of linear regression with binary dependent variables to estimate treatment effects in randomized trials is discussed in Department of Health and Human Services, Office of Adolescent Health, Evaluation Technical Assistance Brief No. 6, December 2014, www.hhs.gov/ash/oah-initiatives/assets/lpm-tabrief.pdf (accessed June 2016).

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 the application is *GRADE*, which indicates 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 five sets of parameter estimates. The coefficients and average partial effects were computed for four probability models: probit, logit, Gompertz, and complementary log log and for the linear regression of *GRADE* on the covariates. The last four 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 average of the density function evaluated at the means of the variables. If one looked only at the coefficient estimates, then it would be natural to conclude that the five models had produced radically different estimates. But a comparison of the columns of average partial effects 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 and logit models.[11] 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 explain the finding. The index $\mathbf{x}'\boldsymbol{\beta}$ is not the random variable. The partial 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

TABLE 17.1 Estimated Probability Models

	Linear		*Logit*		*Probit*		*Comp. Log Log*		*Gompertz*	
Variable	*Coeff.*	*APE*	*Coeff.*	*APE*	*Coeff.*	*APE*	*Coeff.*	*APE*	*Coeff.*	*APE*
Constant	−1.498	–	−13.021	–	−7.452	–	−10.361	–	−7.141	–
GPA	0.464	0.464	2.826	0.363	1.626	0.361	2.293	0.413	1.584	0.319
TUCE	0.010	0.010	0.095	0.012	0.052	0.011	0.041	0.007	0.060	0.012
PSI[a]	0.379	0.379	2.379	0.358	1.426	0.374	1.562	0.312	1.616	0.411
Mean $f(\mathbf{x}'\boldsymbol{\beta})$		1.000		0.128		0.222		0.180		0.201

[a] Partial effects for PSI computed as average of [Prob(Grade $= 1|\mathbf{x}_{(PSI)}, PSI = 1) -$ Prob(Grade $= 1|\mathbf{x}_{(PSI)}, PSI = 0)$].

[11] 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.

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 partial 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.

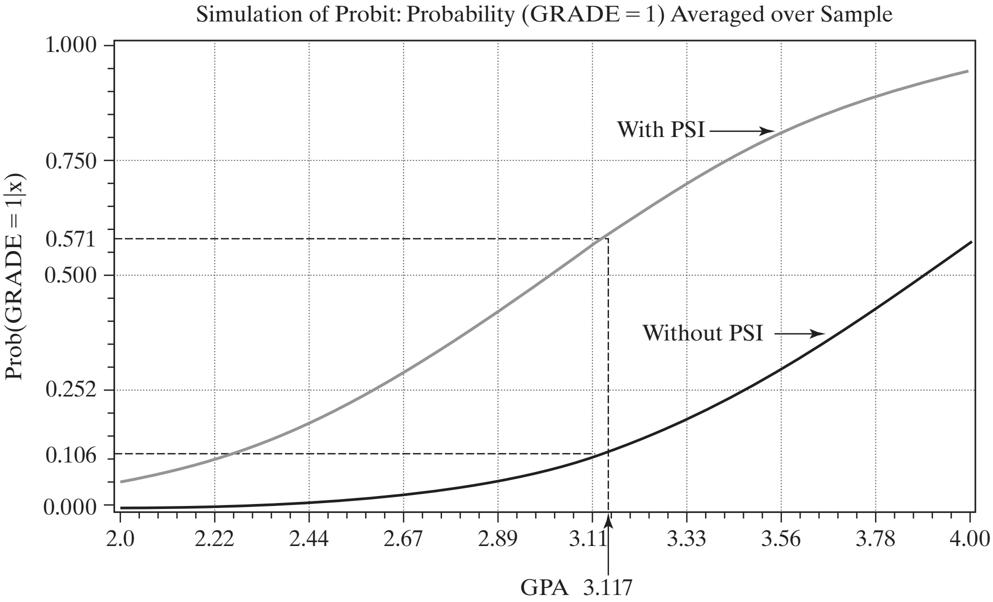
The computation of effects of dummy variables in binary choice settings is an important (one might argue, the most important) element of the analysis. One way to analyze the effect of a dummy variable on the whole distribution is to compute Prob($Y = 1$) over the range of $\mathbf{x}'\boldsymbol{\beta}$ (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 (21.938):

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

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

Figure 17.2 shows these two functions plotted over the range of *GPA* observed in the sample, 2.0 to 4.0. The partial 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 estimate of the partial effect at the mean is 0.468. But of course, this calculation does not show the wide range of differences displayed in Figure 17.2. The APE averages over the entire distribution, and equals 0.374. This latter figure is probably more representative of the desired effect. (In the typical application with a much larger sample, the differences in these results will usually be much smaller.)

FIGURE 17.2 Effect of *GPA* on Predicted Probabilities.

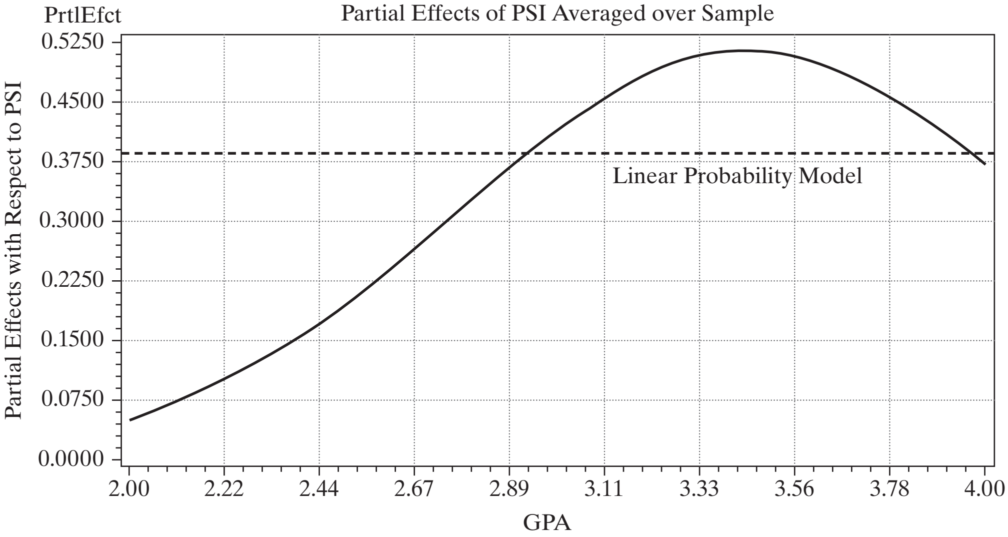


The odds ratio for the *PSI* variable is exp(2.379) = 10.6. This would imply that the odds of a grade increase for those who take the *PSI* are more than 10 times the odds for a student who does not. From Figure 17.2, for the average student, the odds ratio would appear to be about (0.571/0.429)/(0.106/0.894) = 11.1, which is essentially the same result. The partial effect of *PSI* for that student is 0.571 − 0.106 = 0.465. It is clear from Figure 17.2, however, that the partial effect of *PSI* varies greatly depending on the *GPA*. The odds ratio, being a constant, will mask that aspect of the results. The plot in Figure 17.2 is suggestive, but imprecise. A more direct analysis would examine the effect of *PSI* on the probability as it varies with *GPA*. Figure 17.3 shows that effect. The unsurprising conclusion is that the impact of *PSI* is greatest for students in the middle of the grade distribution, not at the low end, which might have been expected. We also see that the marginal benefit of *PSI* actually begins to diminish for the students with the highest *GPA*s, probably because they are most likely already to have *GRADE* = 1. [Figure 17.3 also shows the estimated effect from the linear probability, model (Section 17.2.6) which, like the odds ratio, oversimplifies the relationship.]

Example 17.4 The Light Bulb Puzzle: Examining Partial Effects

The *light bulb puzzle* refers to an observed sluggishness by consumers in adopting energy efficient and environmentally less harmful CFL (compact fluorescent light) bulbs in spite of their advantageous cost and environmental impacts. Di Maria, Ferreira, and Lazarova (2010) examined a survey of Irish energy consumers to learn about the underlying preferences that seem to be driving this puzzling outcome. The authors develop a model of utility maximization over consumption of conventional lighting and CFL lighting. Utility is derived from two sources, consumption of the lighting (in lumens) and environmental impact, *I*. Determination of the binary outcome, "adopt CFL," is based on maximizing utility from the two sources, subject to the costs of adoption, including effort. Individual heterogeneity enters the utility calculation (as a random component) through differences in environmental preferences, perceived costs, understanding of the technology, the costs of the effort in adoption, and differences in individual discount rates.

FIGURE 17.3 Effect of *PSI* on *GRADE* by *GPA*.



The empirical analysis is based on a survey of 1,500 Irish lighting consumers in the 2001 Urban Institute Ireland National Survey on Quality of Life. Inputs to the adoption model are in three components:

Environmental Interest:[12]

Support of Kyoto Protocol (1–4), Importance of Environment (1, 2, 3),
Knowledge of Environment (0, 1).

Demographics:

Age, Gender, Marital Status, Family Size, Education (4 levels), Income

Housing Attributes:

Rural, Own/Rent, Detached or Semidetached Number of Rooms,
House Built Before the 1960s.

The authors report coefficient estimates for probit models with standard errors and partial effects evaluated at the means of the data. Among the statistically significant results reported are partial effects of 0.098 for support of the Kyoto Protocal, 0.044 for the Importance of the Environment, and 0.115 for Knowledge of the Environment. Overall, about 30% of the sample are adopters. The environmental interest variables, therefore, are found to exert a very large influence. The mean values of these variables are 3.05, 2.51, and 0.85, respectively. Thus, starting from the base of 3.05, increased support for Kyoto increases the acceptance rate from about 0.30 to about 0.398, or roughly a third. For the *Importance* variable, the change from the average to the highest would be about 0.5, and the partial effect is 0.044, so the probability would increase by about 0.022 from a base of about 0.3, or about 7.3%, a much smaller increase. For the *Knowledge* variable, the partial effect is 0.115. Increasing this variable from 0 to 1 would increase the probability from 0.3 by about 0.115, or, again, by about one-third.

The average income in the sample is €22,987. The log of the mean is about 10. An increase in the log of income of one unit would take it to 11, or income of about €62,500, which is larger than the maximum in the sample. A more reasonable experiment might be to raise income by about 10%, in which case the log income rises by about 0.095.The partial effect for log income is 0.073. An increase in the log of income of 0.095 would be associated with an increase in the average probability of $0.095 \times 0.073 = 0.007$. This would correspond to a 2.3% increase in the probability, from 0.30 to 0.307.

The authors report an experiment with the marginal effects: "As robustness checks we first estimated the marginal effects associated with the coefficients in Table 5 at different levels of income (1st, 25th, 50th, 75th, and 99th percentile) and educational attainment. The marginal impacts discussed above increase monotonically with the level of income and education, but these increases are not statistically significant." That is, they examined the changes in the partial effect of education associated with changes in income. Superficially, this is an estimation of $\partial[\partial \text{Prob}(\text{Adopt} = 1)/\partial \text{Education}]/\partial \text{income}$. This is the analysis in Figure 17.3.

17.2.6 THE LINEAR PROBABILITY MODEL

The binary outcome suggests a regression model,

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

with

$$E[y\,|\,\mathbf{x}] = \{0 \times [1 - F(\mathbf{x}, \boldsymbol{\beta})]\} + \{1 \times [F(\mathbf{x}, \boldsymbol{\beta})]\} = F(\mathbf{x}, \boldsymbol{\beta}).$$

[12] The authors used a principal component for the three measures in one specification of the model, but the preferred specification used the three environmental variables separately.

This implies the regression model,

$$\begin{aligned} y &= E[y|\mathbf{x}] + (y - E[y|\mathbf{x}]) \\ &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon. \end{aligned}$$

The **linear probability model (LPM)** 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,

$$\mathrm{Var}[\varepsilon|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}(1 - \mathbf{x}'\boldsymbol{\beta}).$$

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.[13] 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. Five of the 32 observations in Example 17.3 predict negative probabilities. (This failure of the model to adhere to the basic assumptions of the theory is sometimes labeled "incoherence.")

In spite of the list of shortcomings, the LPM has been used in a number of recent studies. The principal motivation is that it appears to reliably reproduce the partial effects obtained from the formal models such as probit and logit—often only the signs and statistical significance are of interest. Proponents of the LPM argue that it produces a good approximation to the partial effects in the nonlinear models. The authors of the study in Example 17.5 state that they obtained similar results from a logit model (in the 2002 version, a probit model in the 2003 version). If that is always the case, and given the restrictiveness and incoherence of the linear specification, what is the LPM's advantage? Proponents point to two:

1. **Simplicity.** This is, of course, dubious because modern software requires merely the press of a different button or two for nonlinear models. The argument gains more currency in models that contain endogenous variables. We will return to this case below.
2. **Robustness.** The assumptions of normality or logisticality (?) are fragile while linearity is distribution free. This remains actually to be verified. Researchers disagree on the appropriateness of the LPM. For discussion, see Lewbel, Dong, and Yang (2012) and Angrist and Pischke (2009).

Example 17.5 Cheating in the Chicago School System—An LPM

Jacob and Levitt (2002, 2003) used a binary choice model to detect cheating by teachers on behalf of their students in the Chicago school system. The study developed a method of detecting whether test results had been altered. The model used to generate the final results

[13] There is a deeper peculiarity about this formulation. In the regression models we have examined up to this point, the disturbance, ε, is assumed to embody the independent variation of influences (other variables) that are generated outside the model. Because the disturbance in this model arises only tautologically through the need to have y on the LHS of the equation equal y on the RHS, there is no room in the linear probability model for left-out variables to explain some of the variation in y. For a given $\mathbf{x}$, ε cannot vary independently of $\mathbf{x}$. Although the least squares residuals, e_i, are algebraically orthogonal to $\mathbf{x}_i$, it is difficult to construct a statistical understanding of independence or uncorrelatedness of ε_i and $\mathbf{x}_i$.

in the study is an LPM for the variable "Indicator of classroom cheating." In one of the main results in the paper, the authors report (2002, p. 41): "[T]eachers are roughly 6 percentage points more likely to cheat for students who scored in the second quartile (between the 25th and 50th percentile) in the prior year, as compared to students scoring at the third or fourth quartiles." The coefficient on the relevant variable in the LPM is 0.057, or roughly 6%. This seems like a moderate result. However, only about 1% of the observations in their sample are actually classified as having cheated, overall. As such, if 1% is the baseline, the "6 percentage points" is actually a 600% increase! The moderate result is actually extreme. The result is not surprising, however. The linear probability model forces the probability function to have the same slope all the way from zero to one. It is clear from Figure 17.1, however, that in the extreme tails, such as $F(.) = 0.01$, the function will be much flatter than in the center of the distribution.[14] Unless the entire distribution of the data is confined to the extreme ends of the range, having to accommodate the middle of the distribution will make the LPM highly inaccurate in the tails.[15] An implication of this restriction is shown in Figure 17.3.

17.3 ESTIMATION AND INFERENCE FOR 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,

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

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

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

Taking logs, we obtain

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

The likelihood equations are

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

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

[14] This result appears in the 2002 (NBER) version of the paper, but not in the 2003 version.

[15] See Wooldridge (2010, pp. 562–564).

[16] 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 $\ln L = \Sigma_i \ln F(q_i\mathbf{x}_i'\boldsymbol{\beta})$.

far are relatively straightforward to calibrate. For the logit model, by inserting (17-10) 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.[17] This implication also bears some similarity to the least squares normal equations if we view the term $y_i - \Lambda_i$ as a residual.[18] For the probit model, the log likelihood is

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

The first-order conditions for maximizing ln L are

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

Using the device suggested in footnote 16, we can reduce this to

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

where $q_i = 2y_i - 1$.

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

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

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. 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 $\lambda(y_i, \mathbf{x}_i'\boldsymbol{\beta}) = \lambda_i$ 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,

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

This matrix is also negative definite for all values of $\boldsymbol{\beta}$. The proof is less obvious than for the logit model.[19] It suffices to note that the scalar part in the summation is $\text{Var}[\varepsilon|\varepsilon \le \boldsymbol{\beta}'\mathbf{x}] - 1$ when $y = 1$ and $\text{Var}[\varepsilon|\varepsilon \ge -\boldsymbol{\beta}'\mathbf{x}] - 1$ when $y = 0$. The unconditional variance is one. Because truncation always reduces variance—see

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

[18] 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 19.3.2.

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

Theorem 18.2—in both cases, the variance is between zero and one, so the value is negative.[20]

The asymptotic covariance matrix for the maximum likelihood estimator can be estimated by using the negative 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 $(\mathbf{B})^{-1}$ where

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

where $g_i = (y_i - \mathbf{\Lambda}_i)$ 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,

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

Once again, the scalar part of the expression is always negative [note in (17-20) 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. Because 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 covariance matrix of the QMLE (see Section 14.11 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 studies based on the probit model.[21] (Indeed, it is ubiquitous in the contemporary literature.) 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. 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.[22] 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

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

[21] For example, Fernandez and Rodriguez-Poo (1997), Horowitz (1993), and Blundell, Laisney, and Lechner (1993).

[22] See Section 14.11 and Freedman (2006).

basis for robust estimation techniques" (1982a, p. 4). His very useful result is that, if the QMLE converges to a probability limit, then the sandwich estimator can be used under certain circumstances 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. It is true, however, that the robust estimator does appropriately estimate the asymptotic covariance for the parameter vector that is estimated by maximizing the log likelihood, whether that is $\boldsymbol{\beta}$ or something else. In practice, because the model is generally reasonably specified, the correction usually makes little difference.

Similar considerations apply to the cluster correction of the asymptotic covariance matrix for the MLE described in Section 14.8.2. For data with clustered structure, the estimator is

$$\mathbf{V} = \frac{C}{C-1}\left(-\sum_{c=1}^{C}\sum_{t=1}^{N_c}\frac{\partial^2 \ln f_{ct}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}\partial\hat{\boldsymbol{\theta}}'}\right)^{-1}\left[\sum_{c=1}^{C}\left(\sum_{t=1}^{N_c}\frac{\partial \ln f_{ct}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}}\right)\left(\sum_{t=1}^{N_c}\frac{\partial \ln f_{ct}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}'}\right)\right] \left(-\sum_{c=1}^{C}\sum_{t=1}^{N_c}\frac{\partial^2 \ln f_{ct}(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}\partial\hat{\boldsymbol{\theta}}'}\right)^{-1}. \quad \textbf{(17-24)}$$

(The analogous form will apply for a panel data arrangement with n groups and T_i observations in group i.) The matrix provides an appropriate estimator for the asymptotic variance for the MLE. Whether the MLE, itself, estimates the parameter vector of interest when the observations are correlated (clustered) is a separate issue.

Example 17.6 Robust Covariance Matrices for Probit and LPM Estimators

In Example 7.6, we considered nonlinear least squares estimation of a loglinear model for the number of doctor visits variable shown in Figure 14.6. The data are drawn from the Riphahn et al. (2003) data set in Appendix Table F7.1. We will continue that analysis here by fitting a more detailed model for the binary variable *Doctor* = **1** (*DocVis* > 0). The index function for the model is

$$\text{Prob}(\textit{Doctor} = 1 \mid \mathbf{x}_{it}) = F(\beta_1 + \beta_2\,\textit{Age}_{it} + \beta_3\,\textit{Educ}_{it} + \beta_4\,\textit{Income}_{it} + \beta_5\,\textit{Kids}_{it} + \beta_6\,\textit{Health Satisfaction}_{it} + \beta_7\,\textit{Marital Status}_{it}).$$

The data are an unbalanced panel of 27,326 household-years in 7,293 groups. We will examine the 3,377 observations in the 1994 wave, then the full data set. Descriptive statistics for the variables in the model are given in Table 17.2. (We will use these data in

TABLE 17.2 Descriptive Statistics for Binary Choice Model

	Full Panel: n = 27,326				***1994 Wave: n = 3,377***	
Variable	***Mean***	***Standard Deviation***	***Minimum***	***Maximum***	***Mean***	***Standard Deviation***
Doctor	0.629	0.483	0	1	0.658	0.474
Age	43.526	11.330	25	64	42.627	11.586
Education	11.321	2.325	7	18	11.506	2.403
Income	0.352	0.177	0.0015	3.0671	0.445	0.217
Kids	0.403	0.490	0	1	0.388	0.487
Health Sat.	6.786	2.294	0	10	6.643	2.215
Married	0.759	0.428	0	1	0.710	0.454

several examples to follow.) Table 17.3 presents two sets of estimates for each of the probit model and the linear probability model. The 1994 wave of the panel is used for the top panel of results. The comparison is between the conventional standard errors and the robust standard errors. These would be the White estimator for the LPM and the robust estimator in (14-36) for the MLE. In both cases, there is essentially no difference in the estimated standard errors. This would be the typical result. The lower panel shows the impact of correcting the standard errors of the pooled estimator in a panel. The robust standard errors are based on (17-24). In this case, there is a tangible difference, though perhaps less than one might expect. The correction for clustering produces a 20% to 50% increase in the standard errors.

17.3.2 HYPOTHESIS TESTS

The full menu of procedures is available for testing hypotheses about the coefficients. The simplest method for a single restriction would be the usual t tests, using the standard errors from the estimated asymptotic covariance matrix for the MLE. Based on the asymptotic normal distribution of the estimator, we would use the standard normal table rather than the t table for critical points. (See the several previous examples.) 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

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

TABLE 17.3 Estimates for Binary Choice Models

	Cross Section Estimates, 1994 Wave					
	Probit Model			***Linear Probability Model***		
Variable	***Coefficient***	***Standard Error***	***Robust Std. Error***	***Coefficient***	***Std. Error***	***Robust Std. Error***
Constant	1.69384	0.18199	0.18063	1.05062	0.05986	0.05840
Age	0.00448	0.00240	0.00238	0.00147	0.00080	0.00079
Education	−0.01205	0.01002	0.01002	−0.00448	0.00343	0.00351
Income	−0.09149	0.11187	0.11473	−0.02671	0.03842	0.04016
Kids	−0.24557	0.05514	0.05541	−0.08398	0.01874	0.01907
Health Sat.	−0.18503	0.01201	0.01187	−0.05800	0.00363	0.00319
Married	0.10571	0.06134	0.06131	0.03666	0.02055	0.02040
	Full Panel Data Pooled Estimates					
Variable	***Coefficient***	***Std. Error***	***Clustered Std. Error***	***Coefficient***	***Std. Error***	***Clustered Std. Error***
Constant	1.46973	0.06538	0.08687	0.99472	0.02246	0.02988
Age	0.00617	0.00082	0.00107	0.00213	0.00029	0.00037
Education	−0.01527	0.00360	0.00499	−0.00587	0.00127	0.00180
Income	−0.02838	0.04746	0.05727	−0.00285	0.01667	0.02031
Kids	−0.12993	0.01868	0.02354	−0.04508	0.00656	0.00837
Health Sat.	−0.17466	0.00396	0.00490	−0.05757	0.00126	0.00141
Married	0.06591	0.02103	0.02762	0.02363	0.00730	0.00958

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

$$W = \hat{\boldsymbol{\beta}}_M' \mathbf{V}_M^{-1} \hat{\boldsymbol{\beta}}_M, \tag{17-25}$$

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 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,

$$\ln L_0 = n[P \ln P + (1 - P)\ln(1 - P)], \tag{17-26}$$

where P is the proportion of the observations that have dependent variable equal to 1. These tests of models ML1 and ML2 are shown in Table 17.9 in Example 17.14.

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. Note, again, in Example 17.14, the log likelihood for the logit model is −1,991.13 while for the probit model (not shown) it is −1,990.36. This might suggest a preference for the probit model, but one could not carry out a test based on these results.

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 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, which gives

$$\text{LM} = \left(\sum_{i=1}^{n} g_i \mathbf{x}_i\right)' \left[\sum_{i=1}^{n} E[-h_i]\mathbf{x}_i\mathbf{x}_i'\right]^{-1} \left(\sum_{i=1}^{n} g_i \mathbf{x}_i\right), \tag{17-27}$$

where $E[-h_i]$ is defined in (17-21) for the logit model and in (17-23) for the probit model. One could use the robust estimator in Section 13.3.1 instead.

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

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

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 14.4.6 is also convenient. For any of the models considered (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

$$\text{LM} = n\left[\frac{1}{n}\mathbf{i}'(\mathbf{G}\mathbf{X})(\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{G}')\mathbf{i}\right] = nR_{\mathbf{i}}^2, \tag{17-28}$$

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.

Example 17.7 Testing for Structural Break in a Logit Model

The probit model in Example 17.6, based on Riphahn, Wambach, and Million (2003), is

$$\text{Prob}(\mathit{DocVis}_{it} > 0) = \Phi(\beta_1 + \beta_2\, \mathit{Age}_{it} + \beta_3\, \mathit{Education}_{it} + \beta_4\, \mathit{Income} + \beta_5\, \mathit{Kids}_{it} + \beta_6\, \mathit{HealthSat}_{it} + \beta_7\, \mathit{Married}_{it}).$$

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 6.6.2. Because 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 (based on the 1994 wave of the datra) are shown in Table 17.4. The chi-squared statistic for the likelihood ratio test is

$$\text{LR} = -2(-1{,}990.534 - (-1{,}117.587 - 840.246)) = 65.402.$$

The 95% critical value for seven degrees of freedom is 14.067. To carry out the Wald test for this hypothesis there are two numerically identical ways to proceed. First, using the estimates

TABLE 17.4 Estimated Models for Pooling Hypothesis

	Pooled Sample		*Male*		*Female*	
Variable	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*
Constant	1.69384	0.18199	1.51850	0.23388	1.80570	0.30341
Age	0.00448	0.00240	0.00509	0.00331	0.00031	0.00374
Education	−0.01205	0.01002	−0.01351	0.01309	0.00842	0.01645
Income	−0.09149	0.11187	0.09350	0.15627	−0.30374	0.16447
Kids	−0.24557	0.05514	−0.28068	0.07676	−0.26567	0.08357
Health Sat.	−0.18503	0.01201	−0.19514	0.01635	−0.16289	0.01797
Married	0.10571	0.06134	0.13027	0.08862	0.08212	0.08862
ln *L*	−1,990.534		−1,117.587		−840.246	
Sample Size	3,377		1,812		1,565	

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

$$W = [\hat{\boldsymbol{\beta}}_{Male} - \hat{\boldsymbol{\beta}}_{Female}]'[\text{Est.Asy.Var}(\hat{\boldsymbol{\beta}}_{Male}) + \text{Est.Asy.Var}(\hat{\boldsymbol{\beta}}_{Female})]^{-1}[\hat{\boldsymbol{\beta}}_{Male} - \hat{\boldsymbol{\beta}}_{Female}] = 64.6942.$$

Another way to obtain the same result is to add to the pooled model the original seven variables now multiplied by the *Female* dummy variable. We use the augmented **X** matrix $\mathbf{X}^* = [\mathbf{X}, \textit{female} \times \mathbf{X}]$. The model with 14 variables is now estimated, and a test of the pooling hypothesis is done by testing the joint hypothesis that the coefficients on these seven additional variables are zero. The Lagrange multiplier test is carried out by using this augmented model as well. To apply (17-28), the necessary derivatives are in (17-18). For the probit model, the derivative matrix is simply $\mathbf{G}^* = \text{diag}[\lambda_i]$ from (17-20). For the LM test, the vector $\boldsymbol{\beta}$ that is used is the one for the restricted model. Thus, $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}'_{Pooled}, 0, 0, 0, 0, 0, 0, 0)'$. The estimated values that appear in $\mathbf{G}^*$ are simply those obtained from the pooled model. Then,

$$\text{LM} = \mathbf{i}'\mathbf{G}^*\mathbf{X}^*[(\mathbf{X}^{*\prime}\mathbf{G}^{*\prime})(\mathbf{G}^*\mathbf{X}^*)]^{-1}\mathbf{X}^{*\prime}\mathbf{G}^{*\prime}\mathbf{i} = 65.9686.$$

The pooling hypothesis is rejected by all three procedures.

17.3.3 INFERENCE FOR 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. We have three methods of computing asymptotic standard errors for these: the delta method, the method of Krinsky and Robb, and bootstrapping. All three methods can be found in applications in the received literature. Discussion of the various methods and some related issues appears in Dowd, Greene, and Norton (2014).

17.3.3.a The Delta Method

To compute standard errors, we can use the linear approximation approach discussed in Section 4.6. For the predicted probabilities,

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

where

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

The estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ can be any of those 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

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

which depends on the particular **x** vector used. This result is also useful when a partial effect is computed for a dummy variable. In that case, the estimated effect is

$$\Delta\hat{F} = [\hat{F}|(d = 1)] - [\hat{F}|(d = 0)].$$

The estimator of the asymptotic variance would be

$$\text{Est.Asy.Var}[\Delta\hat{F}] = [\partial\Delta\hat{F}/\partial\hat{\boldsymbol{\beta}}]'\mathbf{V}[\partial\Delta\hat{F}/\partial\hat{\boldsymbol{\beta}}], \quad \textbf{(17-29)}$$

where

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

For the other partial effects, let $\hat{\boldsymbol{\gamma}}(\mathbf{x}) = \hat{f}(\mathbf{x}'\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}$. Then

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

The matrix of derivatives (the Jacobian) is

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

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

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

For the logit model, $\hat{f}(\mathbf{x}'\beta) = \hat{\Lambda}(\mathbf{x})[1 - \hat{\Lambda}(\mathbf{x})]$, so

$$\frac{d\hat{f}(\mathbf{x}'\hat{\boldsymbol{\beta}})}{dz} = [1 - 2\hat{\Lambda}(\mathbf{x})]\left(\frac{d\hat{\Lambda}(\mathbf{x})}{dz}\right) = [1 - 2\hat{\Lambda}(\mathbf{x})]\hat{\Lambda}(\mathbf{x})[1 - \hat{\Lambda}(\mathbf{x})].$$

Collecting terms, we obtain

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

As before, the value obtained will depend on the $\mathbf{x}$ vector used. A common application sets $\mathbf{x}$ at $\overline{\mathbf{x}}$, the means of the data.

The average partial effects would be computed as

$$\bar{\hat{\boldsymbol{\gamma}}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial F(\mathbf{x}_i'\hat{\boldsymbol{\beta}})}{\partial\mathbf{x}_i} = \left[\frac{1}{n}\sum_{i=1}^{n}f(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\right]\hat{\boldsymbol{\beta}}.$$

The preceding estimator appears to be the mean of a random sample. It would be if it were based on the true $\boldsymbol{\beta}$. But the n terms based on the same $\hat{\boldsymbol{\beta}}$ are correlated. The delta method must 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 would use

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

where

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

The estimator of the asymptotic covariance matrix for the APE is simply

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

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.

17.3.3.b An Adjustment to the Delta Method

The delta method treats the data as *fixed in repeated samples*. If, instead, the APE were treated as a parameter to be estimated—that is, a feature of the population from which $(y_i, \mathbf{x}_i)$ are randomly drawn—then the asymptotic variance would account for the variation in $\mathbf{x}_i$ as well.[23] In the application, then, there are two sources of variation: the first is the sampling variation of the parameter estimator of $\boldsymbol{\beta}$ and the second is the sampling variability due to the variation in $\mathbf{x}$.[24] An appropriate asymptotic variance for the APE would be the sum of the two terms.[25]

Assume for the moment that $\boldsymbol{\beta}$ is known. Then, the APE is

$$\overline{\boldsymbol{\gamma}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial F(\mathbf{x}_i'\boldsymbol{\beta})}{\partial \mathbf{x}_i} = \left[\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i'\boldsymbol{\beta})\right]\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\gamma}_i.$$

Based on the sample of observations on the partial effects, the natural estimator of the variance of each of the K estimated partial effects would be

$$\hat{\sigma}^2_{\gamma,k} = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}(\gamma_k(\mathbf{x}_i) - \overline{\gamma}_k)\right]^2 = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}(PE_{i,k} - APE_k)\right]^2.$$[26]

The asymptotic variance of the partial effects estimator is intended to reflect the variation of the parameter estimator, $\hat{\boldsymbol{\beta}}$, whereas the preceding 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, $\hat{\gamma}_k(\mathbf{x}_i) = \hat{\beta}_k\Lambda(\mathbf{x}_i'\hat{\boldsymbol{\beta}})[1 - \Lambda(\mathbf{x}_i'\hat{\boldsymbol{\beta}})] = \hat{\boldsymbol{\beta}}_k\hat{\delta}_i$, and $\hat{\delta}_i$ is the same for all k. It follows that

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

The delta method would use, instead, the kth diagonal element of

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

To account for the variation of the data as well, the variance estimator would be the sum of these two terms.

The impact of the adjustment is data dependent. In our experience, it is usually minor. (It is trivial in the example below.) We do note that the APEs are sometimes computed for specific configurations of $\mathbf{x}$, or specific values, or specific subsets of observations. In these cases, the appropriate adjustment, if any, is unclear.

[23] For example, see equation (17-13).

[24] The two sources of variation are the disturbances (the random part of the random utility model) and the variation of the observed sample of $\mathbf{x}_i$. This does raise a question as to the meaning of the standard errors, robust or otherwise, computed for the linear probability model.

[25] See Wooldridge (2010, p. 467 and 2011, pp. 184–186) for formal development of this result.

[26] See, for example, Contoyannis et al. (2004, p. 498), who reported computing the "sample standard deviation of the partial effects."

17.3.3.c The Method of Krinsky and Robb

The method of Krinsky and Robb was described in Section 15.3. For present purposes, we will apply the method as follows. The MLEs of the model parameters are $\hat{\boldsymbol{\beta}}$ and $\mathbf{V}$. We will draw a random sample of R draws from the multivariate normal population with this mean and variance. This is done by first computing the Cholesky decomposition of $\mathbf{V} = \mathbf{CC}'$ where $\mathbf{C}$ is a lower triangular matrix. With this in hand, we draw R standard multivariate normal vectors $\mathbf{w}_r$, then $\hat{\boldsymbol{\beta}}(r) = \hat{\boldsymbol{\beta}} + \mathbf{C}\mathbf{w}_r$. With each $\hat{\boldsymbol{\beta}}(r)$, we compute the partial effects, either APE or PEA, $\hat{\boldsymbol{\gamma}}(r)$. The estimator of the asymptotic variance is the empirical variance of this sample of R observations,

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}] = \frac{1}{R}\sum_{r=1}^{R} (\hat{\boldsymbol{\gamma}}(r) - \bar{\hat{\boldsymbol{\gamma}}})^2.$$

Note that Krinsky and Robb will accommodate the sampling variability of $\hat{\boldsymbol{\beta}}$ but not the sample variation in $\mathbf{x}_i$ considered in the preceding adjustment to the delta method.

17.3.3.d Bootstrapping

Bootstrapping is described in Section 15.4. It is essentially the same as Krinsky and Robb save that the sample of draws of $\hat{\boldsymbol{\beta}}\ (r)$ is obtained by repeatedly sampling n observations from the data with replacement and reestimating the model with each. In principle, bootstrapping will automatically account for the extra variation due to the data discussed in Section 17.3.2b.

Example 17.8 STANDARD ERRORS FOR PARTIAL EFFECTS

Table 17.5 shows estimates of a simple probit model,

$$\text{Prob}(\mathit{DocVis}_{it} > 0) = \Phi(\beta_1 + \beta_2\,\mathit{Age}_{it} + \beta_3\,\mathit{Education}_{it} + \beta_4\,\mathit{Income}_{it} + \beta_5\,\mathit{Kids}_{it} + \beta_6\,\mathit{HealthSat}_{it} + \beta_7\,\mathit{Married}_{it}).$$

We report the average partial effects and the partial effects at the means. These results are based on the 1994 wave of the panel in Example 17.7. The sample size is 3,377. As noted earlier, the APEs and PEAs differ slightly, but not enough that one would draw a different conclusion about the population from one versus the other. In computing the standard errors for the APEs, we used the delta method without the adjustment in Section 17.3.2b. When that adjustment is made, the results are almost identical. The only change is the standard error for the coefficient on health satisfaction which changes from 0.00361 to 0.00362.

TABLE 17.5 Comparison of Estimators of Partial Effects

	Probit Model		*Average Partial Effects*		*Partial Effects at Means*	
Variable	*Coefficient*	*Std. Error*	*Avg. Partial Effect*	*Std. Error*	*Partial Effect at Means*	*Std. Error*
Constant	1.69384	0.18199				
Age	0.00448	0.00240	0.00150	0.00080	0.00161	0.00086
Education	−0.01205	0.01002	−0.00404	0.00336	−0.00433	0.00360
Income	−0.09149	0.11187	−0.03067	0.03749	−0.03290	0.04022
Kids	−0.24557	0.05514	−0.08358	0.01890	−0.08830	0.01982
Health Sat.	−0.18503	0.01201	−0.06202	0.00362	−0.06653	0.00426
Married	0.10571	0.06134	0.02086	0.02086	0.04801	0.02206

TABLE 17.6 Comparison of Methods for Computing Standard Errors for Average Partial Effects

Variable	*Avg. Partial Effect*	*Std. Error Delta Method*	*Std. Error Krinsky and Robb**	*Std. Error Bootstrap**
Age	0.00150	0.00080	0.00081	0.00080
Education	−0.00404	0.00336	0.00336	0.00372
Income	−0.03067	0.03749	0.03680	0.04065
Kids	−0.08358	0.01890	0.01839	0.02032
Health Sat.	−0.06202	0.00361	0.00384	0.00372
Married	0.02086	0.02086	0.01971	0.02248

*100 Replications.

Table 17.6 compares the three methods of computing standard errors for average partial effects. These results, in a moderate sized data set, in a typical application, are consistent with the theoretical proposition that any of the three methods should be useable. The choice could be based on convenience.

Example 17.9 Hypothesis Tests About Partial Effects

Table 17.7 presents the maximum likelihood estimates for the probit model,

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

(The column labeled “Interaction Model” is the estimates of the model in Example 17.14.) The *t* ratios listed are used for testing the hypothesis that the coefficient or partial effect is zero. The similarity of the *t* statistics for the coefficients and the partial effects is typical. The interpretation differs, however. Consider the test of the hypothesis that the coefficient on *Kids* is zero. The value of −4.45 leads to rejection of the null bypothesis. The same hypothesis about the average partial effect produces the same conclusion. The question is, what should be the conclusion if these tests conflict? If the *t* ratio on the APE for *Kids* were 0.45, then the tests would conflict. And, because

$$\text{APE }(Kids) = \beta_{kids} \times E[density \,|\, \mathbf{x}],$$

TABLE 17.7 Estimates for Binary Choice Models

	Cross Section Estimation, 1994 Wave						
	Probit Model				***Average Partial Effects***		
Variable	*Coefficient*	*Std. Error*	*t Ratio*	*(Interaction Model)*	*Estimate*	*Std. Error*	*t Ratio*
Constant	1.69384	0.18199	9.31	1.98542	–	–	–
Age	0.00448	0.00240	1.86	−0.00177	0.00150	0.00080	−1.86
Education	−0.01205	0.01002	−1.20	−0.03466	−0.00404	0.00336	−1.20
Income	−0.09149	0.11187	−0.82	−0.09903	−0.03067	0.03749	−0.82
Kids	−0.24557	0.05514	−4.45	−0.24976	−0.08358	0.01890	−4.42
Health Sat.	−0.18503	0.01201	−15.40	−0.18527	−0.06202	0.00362	−17.15
Married	0.10571	0.06134	1.72	−0.10598	0.03571	0.02086	1.71
Age × Educ.				0.00055			

the conflict would be fundamental. We have already rejected the hypothesis that β_{kids} equals zero, so the only way that the APE can equal zero is if the second term is zero. But the second term is positive by construction—the density must be positive. Worse, if the expected density were zero, then all the other APEs would be zero as well. The natural way out of the dilemma is to base tests about relevance of variables on the structural model, not on the partial effects. The implication runs in the direction from the structure to the partial effects, not the reverse. That leaves a question. Is there a use for the standard errors for the partial effects? Perhaps not for hypothesis tests, but for developing confidence intervals as in the next example.

Example 17.10 Confidence Intervals for Partial Effects

Continuing the development of Section 17.3.3, the usual approach could be taken for forming a confidence interval for the APE. For example, based on the results in Table 17.7, we would estimate the APE for Kids to be $-0.08358 \pm 1.96\,(0.0189) = [-0.12062 - 0.0465]$. As we noted in Example 17.3, the single estimate of the APE might not capture the interesting variation in the partial effect as other variables change. Figure 17.4 below reproduces the APE for PSI as it varies with GPA in the example of the performance in economics courses. We have added to Figure 17.3 confidence intervals for the APE of PSI for a set of values of GPA ranging from 2 to 4 to show a confidence region.

Example 17.11 Inference about Odds Ratios

The results in Table 17.8 are obtained for a logit model for *GRADE* in Example 17.3. (The coefficient estimates appear in Table 17.1.)

We are interested in the odds ratios for this model, which as we saw in Section 17.2.5, would be computed as $\exp(\hat{\beta}_k)$ for each estimate. Williams (2015) reports the following post-estimation results for this model using Version 11 (and later) of *Stata*. (Some detail has been omitted.)

FIGURE 17.4 Confidence Region for Average Partial Effect.

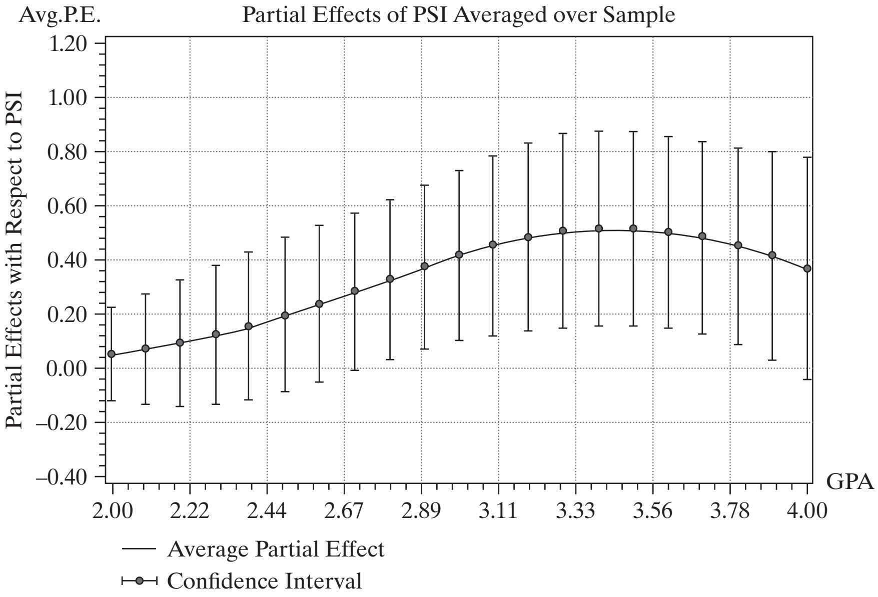


TABLE 17.8 Estimated Logit Model

Variable	*Coefficient*	*Std. Error*	*t Ratio*	*P Value*	*95% Confidence Lower*	*Interval Upper*
Constant	−13.0213	4.93132	−2.64	0.0083	−22.6866	−3.3561
GPA	2.82611	1.26294	2.24	0.0252	0.35079	5.3014
TUCE	0.09516	0.14155	0.67	0.5014	−0.18228	0.37260
PSI	2.37869	1.06456	2.23	0.0255	0.29218	4.46520

```
--------------------------------------------------------------------------------
  grade |    Odds Ratio   Std. Err.     z      P>|z|       [95%     Interval]
                                                          conf.
--------------------------------------------------------------------------------
    gpa |     16.87972    21.31809   2.24    0.035    1.420194   200.6239
   tuce |     1.098832    .1556859   0.67    0.501    .8333651   1.451502
    psi |     10.79073    11.48743   2.23    0.025    1.339344   86.93802
--------------------------------------------------------------------------------
```

This result from a widely used software package provides context to consider what is reported and how to interpret it. The estimated odds ratios appear in the first column. To obtain the standard errors, we would use the delta method. The Jacobian for each coefficient is d[exp($\hat{\beta}_k$)]/d $\hat{\beta}_k$ = exp($\hat{\beta}_k$), so the standard error would just be the odds ratio times the original estimated standard error. Thus, 21.31809 = 16.87972 × 1.26294. But the z is not the ratio of the odds ratio to the estimated standard error. It is the z ratio for the original coefficient. On the other hand, it would make no sense to test the hypothesis that the odds ratio equals zero, because it must be positive. Perhaps the meaningful test would be against the value 1.0, but 2.24 is not equal to (16.87972 − 1)/21.31898 either. The 2.24 and the P value next to it are simply carried over from the original logit model. The implied test is that the odds ratio equals one—it is implied by the equality of the coefficient to zero. The confidence interval would typically be computed as we did in the previous example, but again, the values shown are not equal to 16.87972 ± 1.96 (21.31809). They are equal to exp(0.35079) to exp(5.30143) which is the confidence interval from the original coefficient. This is logical—we have estimated a 95% confidence interval for β, so these values do provide a 95% interval for the exponent. In Section 4.8.3, we considered whether this would be the shortest 95% confidence interval for a prediction of y from ln y, which is what we have done here, and discovered that it is not. On the other hand, it is unclear what utility that is not provided by the coefficient would be provided by the confidence interval for the odds ratio. Finally, as noted earlier, the odds ratio is useful for the conceptual experiment of changing the variable by one unit. For the *GPA* which ranges from 2 to 4 and for *PSI* which is a dummy variable, these would seem appropriate. *TUCE* is a test score that ranges around 30. A unit change in *TUCE* might not be as interesting.

17.3.4 INTERACTION EFFECTS

Models with **interaction effects**, such as

$$\begin{aligned}\text{Prob}(DocVis_{it} > 0) = \Lambda(&\beta_1 + \beta_2 Age_{it} + \beta_3\, Education_{it} + \beta_4\, Income_{it}\\ &+ \beta_5\, Kids_{it} + \beta_6\, Health_{it} + \beta_7\, Married_{it} + \beta_8\, Age_{it} \times Education_{it}),\end{aligned}$$

have attracted considerable attention in recent applications of binary choice models.[27] A practical issue concerns the computation of partial effects by standard computer packages. Write the model as

$$\text{Prob}(DocVis_{it} > 0) = \Lambda(\beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \beta_5 x_{5it} + \beta_6 x_{6it} + \beta_7 x_{7it} + \beta_8 x_{8it}).$$

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

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

$$PE_8 = \partial\,\text{Prob}(DocVis > 0)/\partial x_8 = \beta_8\Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})],$$

which is what common computer packages will dutifully report. The problem is that $x_8 = x_2x_3$, and PE_8 in the previous equation is *not* the partial effect for x_8—there is no meaningful partial effect for x_8 because $x_8 = x_2x_3$. Moreover, the partial effects for x_2 and x_3 will also be misreported by the rote computation. To revert back to our original specification,

$$\partial\text{Prob}(DocVis > 0|\mathbf{x})/\partial\,Age = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})](\beta_2 + \beta_8\,Education),$$
$$\partial\text{Prob}(DocVis > 0|\mathbf{x})/\partial\,Education = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})](\beta_3 + \beta_8\,Age),$$

and what is computed as $\partial\text{Prob}(DocVis > 0|\mathbf{x})/\partial(Age \times Education)$ is meaningless. The practical problem motivating Ai and Norton (2004) was that the computer package does not know that x_8 is x_2x_3, so it computes a partial effect for x_8 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 he or she is using, perhaps by programming the computations themselves.[28]

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

$$\partial^2 E[y|\mathbf{x}]/\partial x_2\partial x_3 = \beta_8,$$

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

$$\partial^2 E[y|\mathbf{x}]/\partial x_2\partial x_3 = \{\Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})]\}\beta_8 + \{\Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})][1 - 2\Lambda(\mathbf{x}'\boldsymbol{\beta})]\}(\beta_2 + \beta_8\,Education)(\beta_3 + \beta_8 Age).$$

Not only is β_8 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 naïve term PE_8 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_8 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 $-\mathbf{x}'\boldsymbol{\beta}\phi(\mathbf{x}'\boldsymbol{\beta}) = 0$.)

[28] The practical issue is now widely understood. Modern computer packages are able to understand model specifications stated in structural form. For our example, rather than compute x_8, the user would literally specifically the instruction to the software as $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_2{*}x_3$ (not computing x_8) and the computation of partial effects would be done accordingly.

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 $E[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$. The results obtained in that study were general, and will apply to the application here, where the nonlinear regression is $E[y|\mathbf{x}] = \Lambda(\mathbf{x}'\boldsymbol{\beta})$ or $\Phi(\mathbf{x}'\boldsymbol{\beta})$.

Example 17.12 Interaction Effect

We added an interaction term, *Age* × *Education*, to the model in Example 17.9. The model is now

$$\text{Prob}(DocVis_{it} > 0) = \Phi(\beta_1 + \beta_2\, Age_{it} + \beta_3\, Education_{it} + \beta_4\, Income_{it} + \beta_5\, Kids_{it} \\ + \beta_6\, Health_{it} + \beta_7\, Married_{it} + \beta_8\, Age_{it} \times Education_{it}).$$

Estimates of the model parameters appear in Table 17.6. Estimation of the probit model produces an estimate of β_8 of 0.00055. It is not clear what this measures. From the correctly specified and estimated model (with the explicit interaction term), the estimated partial effect for education is $\phi(\mathbf{x}'\boldsymbol{\beta})(\beta_3 + \beta_8 Age) = -0.00392$. By fitting the model with x_8 instead of x_2 times x_3, we obtain the first term as the (erroneous) partial effect of education, -0.01162. This implies that the second term, $\phi(\mathbf{x}'\boldsymbol{\beta})\beta_8 Age$, is $-0.00392 + 0.01162 = 0.00770$. As noted, the naïve calculation produces a value that has little to do with the desired result.

17.4 MEASURING GOODNESS OF FIT FOR BINARY CHOICE MODELS

There have been many fit measures suggested for discrete response models.[29] The general intent is to devise a counterpart to the R^2 in linear regression. The R^2 for a linear model provides two useful measures. First, when computed as $1 - \mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$, it measures the success of the estimator at optimizing (minimizing) the fitting criterion, $\mathbf{e}'\mathbf{e}$. That is the interpretation of R^2 as the proportion of the variation of y that is explained by the model. Second, when computed as $Corr^2(y, \mathbf{x}'\mathbf{b})$, it measures the extent to which the predictions of the model are able to mimic the actual data. Fit measures for discrete choice models are based on the same two ideas. We will discuss several.

17.4.1 FIT MEASURES BASED ON THE FITTING CRITERION

Most applications of binary choice modeling use a maximum likelihood estimator. The log-likelihood function itself is the fitting criterion, so as a starting point for considering the performance of the estimator, $\ln L_{\text{MLE}} = \Sigma_{i=1}^{n}[(1 - y_i)\ln(1 - \hat{P}_i) + y_i \ln \hat{P}_i]$ is computed using the MLEs of the parameters. Following the first motivation for R^2, the hypothesis that all the slopes in the model are zero is often interesting. The log likelihood computed with only a constant term will be $\ln L_0 = n[P_0 \ln P_0 + P_1 \ln P_1]$ where n is the sample size and P_j is the sample proportion of zeros or ones. (*Note:* $\ln L_0$ is based only on the sample proportions, so it will be the same regardless of the model.) McFadden's (1974) "Pseudo R^2" or "likelihood ratio index" is

$$R^2_{Pseudo} = LRI = 1 - \frac{\ln L_{MLE}}{\ln L_0}.$$

[29] 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). See, as well, Long and Freese (2006, Sec. 3.5) for a catalog of fit measures for discrete dependent variable models.

This measure has an intuitive appeal in that it is bounded by zero and one and it increases when variables are added to the model.[30] If all the slope coefficients (but not the constant term) are zero, then R^2_{Pseudo} equals zero. Unlike R^2, there is no way to make R^2_{Pseudo} reach one. Moreover, the values between zero and one have no natural interpretation. If $P(\mathbf{x}_i'\boldsymbol{\beta})$ is a proper cdf, 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 $\boldsymbol{\beta}$. 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.[31]

Notwithstanding all of the preceding, this statistic is very commonly reported with empirical results, with references to "fit" and even "proportion of variation explained." A "degrees of freedom correction," $\overline{R}^2_{Pseudo} = 1 - \dfrac{\ln L_{MLE} - K}{\ln L_0}$, has been suggested, as well as some similar ad hoc "adjustments," such as the "Cox and Snell $R^2_{CS} = 1 - \exp(-(\ln \mathrm{L}_M - \ln \mathrm{L}_0)/n)$. We note, however, none of these are fit measures in the familiar sense, and they are not R^2-like measures of explained variation. As a final note, another shortcoming of these measures is that they are based on a particular estimation criterion. There are other estimators for binary choice models, as shown in Example 17.14.

The pseudo R^2 will be most useful for comparing one model to another. If the models are nested, then the log-likelihood function is the natural choice, as examined in the next section. For more general cases, researchers often use one of the information criteria, typically the Akaike Information Criterion,

$$AIC = -2\ln L + 2K \qquad \text{or} \qquad AIC/n,$$

or Schwartz's Bayesian Information Criterion,

$$BIC = -2\ln L + K\ln n \qquad \text{or} \qquad BIC/n.$$

In general, a lower IC value suggests a better model. In comparing nonnested models, some care is needed in interpreting this result, however.

17.4.2 FIT MEASURES BASED ON PREDICTED VALUES

Fit measures based on the predicted probabilities rather than the log likelihood have also been suggested. For example, Efron (1978) proposed a direct counterpart to R^2,

$$R^2_{Efron} = 1 - \frac{\Sigma_{i=1}^n (y_i - \hat{P}_i)^2}{\Sigma_{i=1}^n (y_i - \overline{y})^2}.$$

[30] The log likelihood for a binary choice model must be negative as it is a sum of logs of probabilities. The model with fewer variables is a restricted version of the larger model so it must have a smaller log likelihood. Thus, the log-likelihood function increases when variables are added to the model, and the LRI must be between zero and one. For models with continuous variables, the log likelihood can be positive, so these appealing results are not assured.

[31] See McKenzie (1998) for an application and discussion.

The ambiguity in this measure comes from treating $(y_i - \hat{P}_i)$ as a quantitative residual when the y_i is actually only a label of the outcome. Ben-Akiva and Lerman (1985) and Kay and Little (1986) suggested a fit measure that is keyed to the prediction rule,

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

which can be written as a simple weighted average of the mean predicted probabilities of the two outcomes, $R^2_{\text{BL}} = P_0\hat{\bar{P}}_0 + P_1\hat{\bar{P}}_1$. A 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) and Tjur (2009) have suggested an alternative measure, the *coefficient of discrimination*, that directly considers this failure,

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

This 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

$$\hat{y} = 1 \quad \text{if } \hat{F} > F^* \text{ and 0 otherwise.} \qquad \textbf{(17-30)}$$

(In information theory, this is labeled a *confusion matrix*.) 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. Consider, for example, the naïve predictor

$$\hat{y} = 1 \quad \text{if } P > 0.5 \text{ and 0 otherwise,} \qquad \textbf{(17-31)}$$

where P is the simple proportion of ones in the sample. This rule will always predict correctly 100 P% of the observations, which means that the naïve 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.[32] 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 a $\hat{P}$ 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

[32] See Amemiya (1981).

have $y = 1$, but we will also increase the number of times that we *incorrectly* classify as ones observations that have $y = 0$.[33] In general, any prediction rule of the form in (17-30) 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, incorrectly classifying an applicant as a bad risk is not the same as incorrectly classifying a bad risk as a good one.[34] Changing F^* will always reduce the probability of one type of error 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.

17.4.3 SUMMARY OF FIT MEASURES

The likelihood ratio index and various modifications of it are related to the likelihood ratio statistic for testing the hypothesis that the coefficient vector is zero. Cramer's measure is oriented more toward the relationship between the fitted probabilities and the actual values. It is usefully tied to the standard prediction rule $\hat{y} = \mathbf{1}[\hat{P} > 0.5]$. Whether these have a close relationship to any type of fit in the familiar sense is uncertain. In some cases, it appears so. But the maximum likelihood estimator, on which many of 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 linear regression model (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 17.13 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 confusion matrix 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.[35] The model predicts 491 of 690, or 71.2%, of the observations correctly, although the likelihood ratio index is only 0.083. A naïve model, which always predicts that $y = 0$ because $P < 0.5$, predicts 487 of 690, or 70.6%, 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 naïve predictor.[36]

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

[33] 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.

[34] See Boyes, Hoffman, and Low (1989).

[35] 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.

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

Example 17.14 Fit Measures for a Logit Model

Table 17.9 presents estimates of a logit model for the specification in Example 17.12. Results ML1 are the MLEs for the full model. ML2 is a restricted version from which *Age*, *Education,* and *Health* are excluded. The variables removed are highly significant; the chi-squared statistic for the four restrictions is 2(2,137.06 − 1,991.13) = 291.86. The critical value for 95% from the chi-squared table with four degrees of freedom is 9.49, so the excluded variables significantly contribute to the likelihood for the data. We consider the fit of the model based on the measures suggested earlier. The results labeled NLS in Table 17.9 were computed by nonlinear least squares, rather than MLE. The criterion function is $SS(\mathbf{b}_{NLS}) = \Sigma_i(y_i - \Lambda(\boldsymbol{\beta}'\mathbf{x}_i))^2$. We are interested in how the fit obtained by this alternative estimator compares to that obtained by the MLE. Table 17.10 shows the various scalar fit measures. Note, first, the log likelihood strongly favors ML1. The nonlinear least squares estimates appear rather different from the MLEs but produce nearly the same log likelihood. However, the statistically significant coefficients, on *Kids*, *Health,* and *Married*, are actually almost the same, which would explain the finding. The information criteria favor ML1 as might be expected. The predictive influence of the excluded variables in ML2 is clear in the scalar measures, which generally rise from about 0.01 to 0.10. The Ben-Akiva and Lerman measure does not discriminate between the two specifications. Cramer and the others are essentially the same. Based on the confusion matrices, the count R^2 underscores the difficulty of summarizing the fit of the model to the data. The two models do essentially equally well, though, at predicting different outcomes. ML1 predicts the zeros much better than ML2, but at the cost of many more erroneous predictions of the observations with y equal to one. Overall, the results for this model are typical. The ambiguity of the overall picture suggests the difficulty of constructing a single scalar measure of fit for a binry choice model. The comparison between ML1 and ML2 provided by the Cramer or the other measures seems appropriate. However, it is unclear how to interpret the 0.10 value for the fit measures. It obviously does not reflect a "proportion of explained variation." Nor, however, does it (or the pseudo R^2) have any connection to the ability of the model to predict the outcome variable—the standard predictor obtains a 67.3% success rate. But the naïve predictor, Doctor = 1, will predict correctly 2,222/3,377 or 65.8% of the cases, so the full model improves the success rate from 65.8% to 67.3%

TABLE 17.9 Estimated Parameters for Logit Model for Prob (Doctor=1) (Absolute values of z statistics in parentheses for model ML1)

	Maximum Likelihood ML1	*Maximum Likelihood ML2*	*Nonlinear Least Squares NLS*
Constant	3.18430 (4.00)	0.85360	2.98328
Age	−0.00097 (0.05)	0.00000	0.00294
Education	−0.05054 (0.18)	0.00000	−0.03707
Income	−0.15076 (0.81)	−0.52235	−0.09437
Kids	−0.41358 (4.50)	−0.57608	−0.42014
Health	−0.30957 (14.9)	0.00000	−0.30032
Married	0.17415 (1.71)	0.37995	0.17301
Age × Education	0.00072 (0.47)	0.00000	0.00028

TABLE 17.10 Fit Measures for Estimated Logit Models

	ML1	***ML2***	***NLS***
Based on the log likelihood			
Ln L_0	−2,169.27	−2,169.27	−2,169.27
Ln L_M	−1,991.13	−2,137.06	−1,991.41
Chi squared[df]	356.28[7]	64.41[3]	
Pseudo R^2	0.08212	0.01484	0.0819923
Adjusted Pseudo R^2	0.07889	0.01162	0.0787654
AIC	3,998.27	4,290.13	3,998.81
AIC/n	1.18397	1.27040	1.18413
BIC	4,047.26	4,339.12	4,047.81
BIC/n	1.19848	1.28491	1.19864
Based on the predicted outcomes			
Cramer R^2	0.09840	0.01867	0.09644
Cox-Snell R^2	0.10013	0.01889	0.09998
Efron R^2	0.09736	0.01827	0.09750
Ben-Akiva – Lerman R^2	0.54992	0.54992	0.54954
Count R^2	0.67338	0.65591	0.67516
Confusion Matrix	$\begin{bmatrix} 289 & 866 & 1155 \\ 237 & 1985 & 2222 \\ 526 & 2851 & 3377 \end{bmatrix}$	$\begin{bmatrix} 17 & 1138 & 1155 \\ 24 & 2198 & 2222 \\ 41 & 3336 & 3377 \end{bmatrix}$	$\begin{bmatrix} 285 & 870 & 1155 \\ 227 & 1995 & 2222 \\ 512 & 2865 & 3377 \end{bmatrix}$

17.5 SPECIFICATION ANALYSIS

In the linear regression model, we considered two important specification problems: the effect of omitted variables and the effect of heteroscedasticity. In the linear regression 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. In the context of a binary choice model, they find the following:

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 model, $y = \mathbf{1}[(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon) > 0]$, are heteroscedastic, then the maximum likelihood estimators are inconsistent and

the covariance matrix is inappropriate. This is in contrast to the linear regression case, where heteroscedasticity only affects the estimated asympotic variance of the estimator.

In both of these cases (and others), the impact of the specification error on estimates of partial effects and predictions is less clear, but probably of greater interest.

Any of the three methods of hypothesis testing discussed here can be used to analyze these two specification problems. The Lagrange multiplier test has the advantage that it can be carried out using the estimates from the restricted model, which might bring a saving in computational effort for the test for heteroscedasticity.[37] 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. It is assumed that the null model is nested in the alternative. 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 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 has a limiting chi-squared distribution with degrees of freedom equal to the number of restrictions.

17.5.1 OMITTED VARIABLES

The hypothesis to be tested is

$$\begin{aligned} H_0&: y^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon,\\ H_1&: y^* = \mathbf{x}_1'\boldsymbol{\beta} + \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon, \end{aligned}$$

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 (17-27) or (17-28). For a logit model, for example, the test is carried out as follows: (1) Fit the null model by ML; (2) Compute the fitted probabilities using the null model and the "residuals," $e_i = y_i - P_{i,0}$ arranged in diagonal matrix $\mathbf{E}$; (3) The LM statistic is $\mathbf{1}'\mathbf{EX}(\mathbf{X}'\mathbf{E}^2\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}'\mathbf{1}$. As usual, this can be computed as n times an uncentered R^2, here in the regression of a column of ones on variables $e_i\mathbf{x}_i$. The likelihood ratio test is equally straightforward. Using the estimates of the two models, the statistic is simply $2(\ln L_1 - \ln L_0)$. The Wald statistic would be based on estimates of the alternative model and is computed as in (17-25).

[37] 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).

17.5.2 HETEROSCEDASTICITY

We use the standard formulation analyzed by Harvey (1976)[38] (see Section 14.10.3), $\text{Var}[\varepsilon|\mathbf{z}] = [\exp(\mathbf{z}'\boldsymbol{\gamma})]^2$. We will obtain results specifically for the probit model; the logit or other models are essentially the same.

The starting point is an extension of the binary choice model,

$$y* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon, y = \mathbf{1}(y* > 0),$$
$$E[\varepsilon|\mathbf{x},\mathbf{z}] = 0, \text{Var}[\varepsilon|\mathbf{x},\mathbf{z}] = [\exp(\mathbf{z}'\boldsymbol{\gamma})]^2.$$

There is an ambiguity in the formulation of the model. A nonlinear index function, probit model (with no suggestion of heteroscedasticity),

$$y^{**} = \frac{\mathbf{x}'\boldsymbol{\beta}}{\exp(\mathbf{z}'\boldsymbol{\gamma})} + \varepsilon, \quad y = \mathbf{1}(y** > 0), \varepsilon \sim \text{N}[0,1],$$

leads to the identical log likelihood and the identical estimated parameters. It is not possible to distinguish heteroscedasticity from this nonlinearity in the conditional mean function.[39] Unlike the linear regression model, in this binary choice context, the data contain no direct (identifying) information about scaling, or variation of the dependent variable. (Hence, the *observational equivalence* of the two specifications.) The (identical) signs of $y*$ and $y**$ are unaffected by the variance function. More broadly, the binary choice model creates an ambiguity in the distinction between heteroscedasticity and variation in the mean of the underlying regression.

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

$$\frac{\partial\,\text{Prob}(y = 1|\mathbf{x},\mathbf{z})}{\partial w_k} = \left\{\phi\left[\frac{\mathbf{x}'\boldsymbol{\beta}}{\exp(\mathbf{z}'\boldsymbol{\gamma})}\right]\frac{1}{\exp(\mathbf{z}'\boldsymbol{\gamma})}\right\}(\beta_k - (\mathbf{x}'\boldsymbol{\beta})\gamma_k). \tag{17-32}$$

Only the first (second) term applies if w_k appears only in $\mathbf{x}$ ($\mathbf{z}$). This implies that the simple coefficient may differ greatly from the effect that is of interest in the estimated model. This effect is clearly visible in the next example.[40]

The log likelihood is

$$\ln L = \sum_{i=1}^{n}\left\{y_i \ln F\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\right) + (1 - y_i)\ln\left[1 - F\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\right)\right]\right\}. \tag{17-33}$$

[38] See Knapp and Seaks (1992) for an application. Other formulations are suggested by Fisher and Nagin (1981), Hausman and Wise (1978), Horowitz (1993), and Khan (2013).

[39] See Khan (2013) for extensive discussion of this observational equivalence. Manski (1988) notes this as well.

[40] Wooldridge (2010, pp. 602–603) develops the identification issue in terms of the *average structural function* [Blundell and Powell (2004)]; $\text{ASF}(\mathbf{x}) = E_{\mathbf{z}}[\Phi(\exp(-\mathbf{z}'\boldsymbol{\gamma})\mathbf{x}'\boldsymbol{\beta})]$. Under this interpretation, the partial effect is $\partial\text{ASF}(\mathbf{x})/\partial\mathbf{x} = E_{\mathbf{z}}[\phi(\exp(-\mathbf{z}'\boldsymbol{\gamma})\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}]$. The Average Structural Function treats $\mathbf{z}$ and $\mathbf{x}$ differently (even if they share variables). This computes the function for a fixed $\mathbf{x}$, averaging over the sample values of $\mathbf{z}$. The empirical estimator would be $\partial A\hat{S}F(\mathbf{x})/\partial\mathbf{x} = (1/n)\Sigma_{i=1}^{n}\phi[\exp(-\mathbf{z}_i'\hat{\boldsymbol{\gamma}})\mathbf{x}'\hat{\boldsymbol{\beta}}]\hat{\boldsymbol{\beta}}$. The author suggests “the uncomfortable conclusion is that we have no convincing way of choosing” between (17-32) and this alternative result. Recent applications generally report (17-32), notwithstanding this alternative interpretation. One advantage of interpretation (17-32) is that it explicitly examines the effect of variation in $\mathbf{z}$ on the response probability, particularly in the typical case in which $\mathbf{z}$ and $\mathbf{x}$, have variables in common.

To be able to estimate all the parameters, **z** cannot have a constant term. The derivatives are

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

$$\frac{\partial \ln L}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} \left[\frac{f_i(y_i - F_i)}{F_i(1 - F_i)} \right] \exp(-\mathbf{z}_i'\boldsymbol{\gamma})\mathbf{z}_i(-\mathbf{x}_i'\boldsymbol{\beta}). \tag{17-34}$$

If the model is estimated assuming that $\boldsymbol{\gamma} = \mathbf{0}$, then we can easily test for homoscedasticity. Let g_i equal the bracketed function in (17-34), $\mathbf{G} = \text{diag}(g_i)$ and

$$\mathbf{w}_i = \begin{bmatrix} \mathbf{x}_i \\ (-\mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{z}_i \end{bmatrix}, \tag{17-35}$$

computed at the maximum likelihood estimator, assuming that $\boldsymbol{\gamma} = \mathbf{0}$. Then, the LM statistic is

$$\text{LM} = \mathbf{i}'\mathbf{GW}[(\mathbf{W}'\mathbf{G})(\mathbf{GW})]^{-1}\mathbf{W}'\mathbf{Gi} = nR^2,$$

where the regression is of a column of ones on $g_i\mathbf{w}_i$. Wald and likelihood ratio tests of the hypothesis that $\boldsymbol{\gamma} = \mathbf{0}$ are also straightforward based on maximum likelihood estimates of the full model.

Davidson and MacKinnon (1981) 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 pick up some other form of misspecification, however, including perhaps the simple omission of **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. The problem in the binary choice context stems partly from the ambiguous interpretation of the role of **z** in the model discussed earlier.

Example 17.15 Specification Test 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 following specification [see Wooldridge (2010, p. 580)]:[41]

$$\text{Prob}[LFP = 1] = F(Constant, Other\ Income, Education, Experience, Experience^2, Age, Kids\ Under\ 6, Kids\ 6\ to\ 18).$$

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 -401.3022. The chi-squared statistic is, therefore, 227.142. The critical value from the chi-squared distribution with seven degrees of freedom is 14.07, so the joint hypothesis that the coefficients on *Other Income, etc.* are all zero is rejected.

Consider the alternative hypothesis, that the constant term and the coefficients on *Other Income, etc.* are the same whether the individual resides in a city (CITY = 1) or not (CITY = 0), against the alternative that an altogether different equations apply for the two

[41] Other income is computed as family income minus the wife's hours times the wife's reported wage, divided by 1,000. This produces several small negative values. In the interest of comparability to the received application, we have left these values intact.

TABLE 17.11 Estimated Coefficients

		Homoscedastic		***Heteroscedasti***	
		Estimate (Std. Err.)	***Partial Effect****	***Estimate (Std. Err.)***	***Partial Effect****
Constant	β_1	0.27008 (0.5086)	–	0.25140 (0.4548)	–
Other Inc.	β_2	−0.01202 (0.0048)	−0.00362 (0.0014)	−0.01075 (0.0044)	−0.00362 (0.0014)
Education	β_3	0.13090 (0.0253)	0.39370 (0.0072)	0.11734 (0.0255)	0.03949 (0.0072)
Exper	β_4	0.12335 (0.0187)	0.02558 (0.0022)	0.11190 (0.0197)	0.02599 (0.0022)
Exper^2	β_5	−0.00189 (0.0006)		−0.00171 (0.0006)	
Age	β_6	−0.05285 (0.0085)	−0.01590 (0.0024)	−0.04774 (0.0089)	−0.01607 (0.0024)
Kids < 6	β_7	−0.86833 (0.1185)	−0.26115 (0.0131)	−0.77151 (0.1356)	−0.25968 (0.0318)
Kids 6–18	β_8	0.03600 (0.0438)	0.01083 (0.0319)	0.02800 (0.0390)	0.00943 (0.0130)
City	γ	0.00000		−0.17446 (0.1541)	0.00843 (0.0075)
ln *L*		−401.302		−400.641	

*Average partial effects and estimated standard errors include both mean ($\boldsymbol{\beta}$) and variance ($\boldsymbol{\gamma}$) effects.

groups of women. To test this hypothesis, we would use a counterpart to the Chow test of Section 6.4.1 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 and the seven variables listed above—is −401.302. The log likelihoods for this model based on the 484 observations with CIT = 1 and the 269 observations with CIT = 0 are −255.552 and −142.727, respectively. The log likelihood for the unrestricted model with separate coefficient vectors is thus the sum, −398.279. The chi-squared statistic for testing the eight restrictions of the pooled model is twice the difference, 6.046. The 95% critical value from the chi-squared distribution with 8 degrees of freedom is 15.51, so at this significance level, the hypothesis that the constant terms and the other coefficients are all the same is not rejected.

Table 17.11 presents estimates of the probit model with a correction for heteroscedasticity of the form $\text{Var}[\varepsilon_i] = [\exp(\gamma CITY)]^2$. The three tests for homoscedasticity give

$$\begin{aligned}
\text{LR} &= 2[-400.641 - (-401.302)] = 1.322,\\
\text{LM} &= 1.362 \text{ based on the BHHH estimator},\\
\text{Wald} &= (-1.13)^2 = 1.276.
\end{aligned}$$

The 95% critical value for one restrictions is 3.84 so the three tests are consistent in not rejecting the hypothesis that γ equals zero.

17.5.3 DISTRIBUTIONAL ASSUMPTIONS

One concern about the models suggested here is that the choice of the particular distribution is itself vulnerable to a specification error. For example, the problem arises if a probit model is analyzed when a logit model would be appropriate.[42] It might seem logical to test the hypothesis of the model along with the other specification analyses one might do. Alternatively, a more robust, less parametric specification might be attractive. The substantive difference between probit and logit coefficient estimates in the preceding examples (e.g., Example 17.3) is misleading. The difference masks the underlying scaling of

[42] See, for example, Ruud (1986).

the distributions. The partial effects generated by the models are typically almost identical. This is a widely observed result that suggests that concerns about biases in the coefficients due to the wrong distribution might be misplaced. The other element of the analysis is the predicted probabilities. Once again, the scaling of the coefficients by the different models disguises the typical similarity of the predicted probabilities of the different parametric models. A broader question concerns the specific distribution compared to a semi- or nonparametric alternative. Manski's (1988) maximum score estimator [and Horowitz's (1992) smoothed version], Klein and Spady's (1993) semiparametric (kernel function based), and Khan's (2013) heteroscedastic probit model are a few of the less heavily parameterized specifications that have been proposed for binary choice models. Frolich (2006) presents a comprehensive survey of nonparametric approaches to binary choice modeling, with an application to Portuguese female labor supply.

The linear probability model is not offered as a robust alternative specification for the choice model. Proponents of the linear probability model argue only that the linear regression delivers a reliable approximation to the partial effects of the underlying true probability model.[43] The robustness aspect is speculative. The approximation does appear to mimic the nonlinear results in many cases. In terms of the relevant computations, partial effects and predicted probabilities, the various candidates seem to behave similarly. An essential ingredient is often the curvature in the tails that allows predicted probabilities to mimic the features of unbalanced samples. From this standpoint, the linear model would seem to be the less robust specification. (See Example 17.5.) It is precisely this rigidity of the LPM (as well as the parametric models) that motivates the nonparametric approaches such as the local likelihood logit approach advocated by Frolich (2006).

Example 17.16 Distributional Assumptions

Table 17.12 presents estimates of the model in Example 17.36 based on the linear probability model and four alternative specifications. Only the estimated partial effects are shown in the table. The probit estimates match the authors' results. The correspondence of the various results is consistent with the earlier observations. Generally, the models produce similar results. The linear probability model does stand alone for two of the seven results, for the market share and productivity variables.

TABLE 17.12 Estimated Partial Effects in a Model of Innovation

				Complementary	
	Linear	*Probit*	*Logit*	*Log Log*	*Gompertz*
Log Sales	0.05198	0.06573	0.06766	0.06457	0.06639
Share	0.09492	0.39812	0.43993	0.33011	0.49826
Imports	0.45284	0.42080	0.41101	0.43734	0.40304
FDI	1.07787	1.05890	1.08753	0.99556	1.12929
Productivity	−0.55012	−0.86887	−1.01060	−0.85039	−0.87471
Raw Material	−0.09861	−0.10569	−0.09635	−0.10626	−0.10615
Investment	0.07879	0.07045	0.06758	0.07704	0.06356

[43] Chung and Goldberger (1984), Stoker (1986, 1992), and Powell (1994) (among others) consider general cases in which $\boldsymbol{\beta}$ can be consistently estimated "up to scale" using ordinary least squares. For example, Stoker (1986) shows that if $\mathbf{x}$ is multivariate normally distributed, then the LPM would provide a consistent estimator of the slopes of the probability function under very general specifications.

17.5.4 CHOICE-BASED SAMPLING

In some studies, 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.[44] 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 exogenous 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 $\mathbf{B}$.) 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.

Example 17.17 Credit Scoring

In Example 7.12, 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*, *OwnRent* (whether the applicant owns hi or her home), *Self-Employed* (whether he or she is self-employed), and how long the applicant has lived at his or her current address. The number of major and minor derogatory reports (60-day and 30-day delinquencies) are also very influential variables in credit scoring. The probit model we will use to "model the model" is

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

[44] For example, Boyes, Hoffman, and Low (1989) and Greene (1992).

TABLE 17.13 Estimated Card Application Equation (*t* ratios in parentheses)

	Unweighted			*Weighted*		
Variable	*Estimate*	*Std. Error*		*Estimate*	*Std. 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)

In the data set, 78.1% of the applicants are cardholders. In the population, at that time, the true proportion was roughly 23.2%, 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 17.13 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. Because lenders are strongly focused on default probability, past evidence of default behavior will be a major consideration.

17.6 TREATMENT EFFECTS AND ENDOGENOUS VARIABLES IN BINARY CHOICE MODELS

Consider the binary choice model with endogenous right-hand side variable, T,

$$y* = \mathbf{x}'\boldsymbol{\beta} + T\gamma + \varepsilon, y = \mathbf{1}(y* > 0), \text{Cov}(T, \varepsilon) \neq 0.$$

We examine the two leading cases:

1. T is an endogenous dummy variable that indicates some kind of treatment or program participation such as graduating from high school or college, receiving some kind of job training, purchasing health insurance, etc.[45]
2. T is an endogenous continuous variable. Because the model is not linear, conventional instrumental variable estimators such as two-stage least squares (2SLS) are not appropriate. We consider the alternative estimators based on the maximum likelihood estimator.

[45] Discussion appears in Angrist (2001) and Angrist and Pischke (2009, 2010).

17.6.1 ENDOGENOUS TREATMENT EFFECT

A structural model in which a treatment effect will be correlated with the unobservables is

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

The correlation between u and ε induces the endogeneity of T in the equation for y. We are interested in two effects: (1) the causal treatment effect of T on Prob(y $= 1|\mathbf{x}, T)$, and (2) the partial effects of $\mathbf{x}$ and $\mathbf{z}$ on Prob($\mathbf{y} = 1|\mathbf{x}, \mathbf{z}, T$) in the presence of the endogenous treatment.

This **recursive model** is a bivariate probit model (Section 17.9.5). The log likelihood is constructed from the joint probabilities of the observed outcomes. The four possible outcomes and associated probabilities are obtained as the marginal probabilities for T times the conditional probabilities for $y|T$. Thus, $P(y = 1, T = 1) = P(y = 1|T = 1)P(T = 1)$. The marginal probability for $T = 1$ is just $\Phi(\mathbf{z}_i'\boldsymbol{\alpha})$, whereas the conditional probability is the bivariate normal probability divided by the marginal, $\Phi_2(\mathbf{x}_i'\boldsymbol{\beta} + \gamma, \mathbf{z}_i'\boldsymbol{\alpha}, \rho)/\Phi(\mathbf{z}_i'\boldsymbol{\alpha})$. The product returns the bivariate normal probability. The other three terms in the log likelihood are derived similarly. The four terms are

$$\begin{aligned}
P(y = 1, T = 1|\mathbf{x}, \mathbf{z}) &= \Phi_2(\mathbf{x}'\boldsymbol{\beta} + \gamma, \mathbf{z}'\boldsymbol{\alpha}, \rho),\\
P(y = 1, T = 0|\mathbf{x}, \mathbf{z}) &= \Phi_2(\mathbf{x}'\boldsymbol{\beta} + \gamma, -\mathbf{z}'\boldsymbol{\alpha}, -\rho),\\
P(y = 0, T = 1|\mathbf{x}, \mathbf{z}) &= \Phi_2[-(\mathbf{x}'\boldsymbol{\beta} + \gamma), \mathbf{z}'\boldsymbol{\alpha}, -\rho],\\
P(y = 0, T = 0|\mathbf{x}, \mathbf{z}) &= \Phi_2[-(\mathbf{x}'\boldsymbol{\beta} + \gamma), -\mathbf{z}'\boldsymbol{\alpha}, \rho).
\end{aligned}$$

The log likelihood is then

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \rho) = \sum_{i=1}^{n} \ln \text{Prob}(y = y_i, T = T_i|\mathbf{x}, \mathbf{z}).$$

Estimation is discussed in Section 17.9.5. The model looks like a conventional simultaneous-equations model; the difference arises from the nonlinear transformation of (y^*,T^*) that produces the observed (y,T). One implication is that whereas for identification of a linear model of this form, there would have to be at least one variable in $\mathbf{z}$ that is not in $\mathbf{x}$, that is not the case here. The model is identified partly through the nonlinearity of the functional form. (See the commentary in Example 17.18.)

The *treatment effect (TE)* is derived from the marginal distribution of *y,*

$$\begin{aligned}
\text{TE} &= \text{Prob}(y = 1|\mathbf{x}, T = 1) - \text{Prob}(y = 1|\mathbf{x}, T = 0)\\
&= \Phi(\mathbf{x}'\boldsymbol{\beta} + \gamma) - \Phi(\mathbf{x}'\boldsymbol{\beta}).
\end{aligned}$$

The *average treatment effect (ATE)*, will be estimated by averaging the estimates of TE over the sample observations. The *treatment effect on the treated (ATET)* would be based on the conditional probability, Prob$(y = 1|T = 1)$,

$$\text{TET} = \Phi\left[\frac{(\mathbf{x}'\boldsymbol{\beta} + \gamma) - \rho(\mathbf{z}'\boldsymbol{\alpha})}{\sqrt{1 - \rho^2}}\right] - \Phi\left[\frac{(\mathbf{x}'\boldsymbol{\beta}) - \rho(\mathbf{z}'\boldsymbol{\alpha})}{\sqrt{1 - \rho^2}}\right].$$

The ATET is computed by averaging this quantity over the sample observations for which $T_i = 1$.[46]

To compute the average partial effects for the exogenous variables, we will require

$$\begin{aligned}&\text{Prob}(y = 1|\mathbf{x}, \mathbf{z}, T = 0)\ \text{Prob}(T = 0|\mathbf{z}) +\\ &\text{Prob}(y = 1|\mathbf{x}, \mathbf{z}, T = 1)\ \text{Prob}(T = 1|\mathbf{z})\\ &\qquad = \Phi_2(\mathbf{x}'\boldsymbol{\beta} + \gamma, \mathbf{z}'\boldsymbol{\alpha}, \rho) + \Phi_2(\mathbf{x}'\boldsymbol{\beta}, -\mathbf{z}'\boldsymbol{\alpha}, -\rho\end{aligned}$$

The partial effects for **x** and **z** are then

$$\frac{\partial\, \text{Prob}(y = 1|\mathbf{x}, \mathbf{z})}{\partial\begin{pmatrix}\mathbf{x}\\ \mathbf{z}\end{pmatrix}} = \frac{\partial[\Phi_2(\mathbf{x}'\boldsymbol{\beta} + \gamma, \mathbf{z}'\boldsymbol{\alpha}, \rho) + \Phi_2(\mathbf{x}'\boldsymbol{\beta}, -\mathbf{z}'\boldsymbol{\alpha}, -\rho)]}{\partial\begin{pmatrix}\mathbf{x}\\ \mathbf{z}\end{pmatrix}}.$$

Expressions for the derivatives appear in Section 17.9. This is a fairly intricate calculation. It is automated or conveniently computed in contemporary software, however. We can interpret $\partial\text{Prob}(y = 1|\mathbf{x}, \mathbf{z})/\partial\mathbf{x}$ as a *direct effect* and $\partial\text{Prob}(\text{y} = 1|\mathbf{x}, \mathbf{z})/\partial\mathbf{z}$ as an indirect effect on y that is transmitted through T. For variables that appear in both **x** and **z**, the total effect is the sum of the two. The computations are illustrated in Example 17.19 below.

Example 17.18 An Incentive Program for Quality Medical Care

Scott, Schurer, Jensen, and Sivey (2009) examined an incentive program for Australian general practitioners to provide high quality care in diabetes management. The specific outcome of interest is ordering HbA1c tests as part of a diabetes consultation. The treatment of interest is participation in the incentive program.

A pay-for-performance program, the Practice Incentive Program (PIP) was superimposed on the Australian fee for service system in 1999 to encourage higher quality of care in chronic diseases including diabetes. Program participation by general practitioners (GPs) was voluntary. The quality of care outcome is whether the HbA1c test is administered. Analysis is conducted with a unique data set on GP consultations. The authors compare the average proportion of HbA1c tests ordered by GPs who have joined the incentive scheme with the average proportion of tests ordered by GPs who have not joined, while controlling for key sources of unobserved heterogeneity. A key assumption here is that HbA1c tests are undersupplied in the absence of the PIP scheme and therefore more frequent HbA1c testing is related to higher quality management. The endogenous nature of general practitioners' participation in the PIP is addressed by applying a bivariate probit model, using exclusion restrictions to aid identification of the causal parameters.

The GP will join the PIP if the utility from joining is positive. Utility depends on the additional income from joining the PIP, from the diabetes sign-on payment and negatively on the costs of accreditation and establishing the requisite IT systems. GPs will increase quality of care if the utility of doing so is positive, which partly depends on PIP membership. The bivariate probit model used is

$$\begin{aligned} Y_{ij}* &= \alpha_1 + \boldsymbol{\beta}_1'\mathbf{X}_{ij} + \beta_{PIP}\, PIP_{ij} + u_{1ij}\\ PIP_{ij}* &= \alpha_2 + \boldsymbol{\beta}_2{}'\mathbf{X}_{\text{ij}} + \boldsymbol{\pi}'\mathbf{I}_{ij} + u_{2ij},\end{aligned}$$

where $\quad Y_{ij} = \mathbf{1}$ (GP j ordered an HbA1c test in recorded consultation i),
and $\quad PIP_{ij} = \mathbf{1}$ (Practice in which GP_j works has joined the *PIP* program).

[46] See Jones (2007).

The authors calculate the marginal treatment effect of PIP using $ME_{PIP} = \beta_{PIP}\, \phi(\hat{\boldsymbol{\beta}}_1'\bar{\mathbf{x}})$.[47] Regarding the specification, they note "[a]lthough the model is formally identified by its nonlinear functional form, as long as the full rank condition of the data matrix is ensured (Heckman, 1978; Wilde, 2000), we introduce exclusion restrictions to aid identification of the causal parameter β_{PIP} (Maddala, 1983); Monfardini and Radice (2008). The row vector $\mathbf{I}_{ij}$ captures the variables in the PIP participation equation (5) but excluded from the outcome equation (4)."

Marginal effects for PIP status are reported (in Table II) for two treatment groups. For the first group, the estimated effect is roughly 0.2. In year 1 of the data set, before the PIP was introduced, the average proportion of HbA1c tests conducted was 13%. After the reform was introduced, the average diabetes patient therefore faced a probability of 32% of receiving an HbA1c test during an average encounter in a practice that has joined the PIP. The result from a univariate probit model that treated PIP as exogenous produced a corresponding value of only 0.028.

Example 17.19 Moral Hazard In German Health Care

Riphahn, Wambach, and Million (2003) examined health care utilization in a panel data set of German households. The main objective of the study was to consider evidence of moral hazard. The authors considered the joint determination of hospital and doctor visits in a bivariate count data model. The model assessed whether purchase of Add-on insurance was associated with heavier use of the health care system. All German households have some form of health insurance. In our data, roughly 89% have the compulsory public form. Some households, typically higher income, can opt, instead, for private insurance. The "Add-on" insurance, that is available to those who have the compulsory public insurance, provides coverage for additional benefits, such as certain prevention programs and additional dental coverage. We will construct a small model to suggest the computations of treatment effects in a recursive bivariate probit model. The structure for one of the two count variables is

$$Hospital* = \beta_1 + \beta_2\, Age + \beta_3\, Working + \beta_4\, Health + \gamma\, Addon + \varepsilon,$$

$$Addon* = \alpha_1 + \alpha_2\, Age + \alpha_3\, Education + \alpha_4\, Income + \alpha_5\, Married + \alpha_6\, Kids + \alpha_7\, Health + u.$$

Hospital is constructed as $\mathbf{1}$(*Hospital Visits* > 0) while *Add-On* $= \mathbf{1}$(*Household has Add-On Insurance*). Estimation is based, once again, on the 1994 wave of the data.

Estimation results are shown in Table 17.14. We find that the only significant determinant of hospital visitation is *Health* (measured as self-reported Health Satisfaction). The crucial parameter is γ, the coefficient on *Add-On*. The value of 0.04131 for APE(Add-On) is the estimated average treatment effect. We find, as did Riphahn, that the data do not appear to support the hypothesis of moral hazard. The *t* ratio on *Add-On* in the regression is only 0.16, far from significant. On the other hand, the estimated value, 0.04131, is not trivial. The mean value of *Hospital* is 0.091; 9.1% of this sample had at least one hospital visit in 1994. On average, if the subgroup of *Add-On* policy holders visited the hospital with 0.04 greater probability, this represents, using 0.091 as the base, an increase of 44% in the rate. That is actually quite large. For comparison purposes, the 2SLS estimates of this model are shown in the last column. (The authors of the application in Example 17.6 used 2SLS for estimation of their recursive bivariate probit model.) As might be expected, the 2SLS estimates provide a good approximation to the average partial effects of the exogenous variables. However, it produces an estimate for the causal *Add-On* effect that is three times as large as the FIML estimate, and has the wrong sign.

[47] The calculation of ME_{PIP} treats *PIP* as if it were continuous and differentiates the probability. This approximates $\Phi(\hat{\boldsymbol{\beta}}_1'\bar{\mathrm{x}} + \beta_{\mathrm{PIP}}) - \Phi(\hat{\boldsymbol{\beta}}_1'\bar{\mathrm{x}})$ as suggested earlier. The authors note: "An alternative is to calculate the difference in the probabilities of an HbA1c test in a consultation in which the practice participates in the PIP, and a practice that does not. Our method assumes the treatment indicator to be continuous to be able to use the delta method. We compared the two methods and the magnitude of the marginal effect is the same." (There is, in fact, no obstacle to using the delta method for the difference in the probabilities. See equation (17-29).) The authors computed the TE at the means of the data rather than averaging the TE values over the observations.

TABLE 17.14 Estimates of Recursive Bivariate Probit Model

	Add-On			*Hospital*				
Variable	*Estimate*	*Std. Error*	*t Ratio*	*Estimate*	*Std. Error*	*t Ratio*	*APE*	*2SLS*
Constant	−3.64543	0.42225	−8.63	−0.56009	0.18342	−3.05		0.24352
Health	0.00452	0.02552	0.18	−0.14258	0.01412	−10.10	−0.02195	−0.02505
Working				0.00728	0.07223	0.10	0.00112	0.00121
Add-On				0.23389	1.43618	0.16	0.04131	−0.11826
Age	0.00884	0.00568	1.56	0.00210	0.00292	0.72	0.00034*	0.00035
Education	0.07896	0.02030	3.89					
Income	0.48428	0.23142	2.09					
Married	−0.09885	0.13584	−0.73					
Kids	0.21025	0.13142	1.60					
ρ				−0.01363	0.60432	−0.02		
Log likelihood function		−1296.40433						
Estimation based on $N = 3377, K = 13$								

*Average Treatment Effect. Estimated ATET is 0.03861

17.6.2 ENDOGENOUS CONTINUOUS VARIABLE

If the endogenous variable in the recursive model is continuous, the structure is

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

In the model for labor force participation in Example 17.15, Family income is endogenous.

17.6.2.a IV and GMM Estimation

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. Starting from

$$\begin{aligned}
E[\varepsilon_i|\mathbf{z}_i, \mathbf{x}_i] &= 0,\\
E[T_i\mathbf{z}_i] &\neq \mathbf{0},
\end{aligned}$$

a natural instrumental variable estimator would be based on the moment condition,

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

(In this formulation, $\mathbf{z}_i*$ would contain only the variables in $\mathbf{z}_i$ not also contained in $\mathbf{x}$.) However, y_i^* is not observed, y_i is. The approach that was used in Avery et al. (1983), Butler and Chatterjee (1997), and Bertschek and Lechner (1998) is to assume that the instrumental variables are orthogonal to the residual, $[y - \Phi(\mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i)]$; that is,

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

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.)

17.6.2.b Partial ML Estimation

Simple probit estimation based on y_i and $(\mathbf{x}_i, T_i)$ will not consistently estimate $(\boldsymbol{\beta}, \gamma)$ because of the correlation between T_i and ε_i induced by the correlation between u_i and ε_i. The maximum likelihood estimator is based on the full specification of the model, including the bivariate normality assumption that underlies the endogeneity of T. One possibility is to use the partial reduced form obtained by inserting the first equation in the second. This becomes a probit model with probability $\text{Prob}(y_i = 1 \,|\, \mathbf{x}_i, \mathbf{z}_i) = \Phi(\mathbf{x}_i'\boldsymbol{\beta}* + \mathbf{z}_i^{*\prime}\boldsymbol{\alpha}^*)$. This will produce a consistent estimator 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 would estimate a mixture of $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^*$ for any variable that appears in both $\mathbf{x}_i$ and $\mathbf{z}_i$.) Newey (1987) suggested a minimum chi-squared estimator that does estimate all parameters. Linear regression of T_i on $\mathbf{z}_i$ produces estimates of $\boldsymbol{\alpha}$ and σ_u^2, which suggests a third possible estimator, based on a two-step MLE. But there is no method of moments estimator of ρ or γ produced by this procedure, so this estimator is incomplete.

17.6.2.c Full Information Maximum Likelihood Estimation

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 T_i, which we write as the product of the conditional and the marginal densities,

$$f(y_i, T_i) = f(y_i \,|\, T_i) f(T_i).$$

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

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

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

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

Therefore,

$$\text{Prob}[y_i = 1 \,|\, \mathbf{x}_i, T_i] = \Phi\left[\frac{\mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i + (\rho/\sigma_u)u_i}{\sqrt{1 - \rho^2}}\right]. \tag{17-36}$$

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

$$\ln L = \sum_{i=1}^{n}\left\{\ln\Phi\left[(2y_i - 1)\left(\frac{\mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i + (\rho/\sigma_u)(T_i - \mathbf{z}_i'\boldsymbol{\alpha})}{\sqrt{1 - \rho^2}}\right)\right] + \ln\left[\frac{1}{\sigma_u}\phi\left(\frac{T_i - \mathbf{z}_i'\boldsymbol{\alpha}}{\sigma_u}\right)\right]\right\}. \tag{17-37}$$

Some convenience can be obtained by rewriting the log-likelihood function as

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

where $\tilde{\boldsymbol{\beta}} = \left(1/\sqrt{1-\rho^2}\right)\boldsymbol{\beta}$, $\tilde{\gamma} = \left(1/\sqrt{1-\rho^2}\right)\gamma$ and $\tau = \left(\rho/\sqrt{1-\rho^2}\right)$. The delta method can be used to recover the original parameters and appropriate standard errors after estimation.[48]

Partial effects are derived from the first term in (17-37),

$$\frac{\partial \text{Prob}(y = 1 \mid \mathbf{x}, T, \mathbf{z})}{\partial \begin{pmatrix} \mathbf{x} \\ T \\ \mathbf{z} \end{pmatrix}} = \frac{\partial \Phi\left(\dfrac{\mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i + (\rho/\sigma_u)(T_i - \mathbf{z}_i'\boldsymbol{\alpha})}{\sqrt{1-\rho^2}}\right)}{\partial \begin{pmatrix} \mathbf{x} \\ T \\ \mathbf{z} \end{pmatrix}}$$

$$= \phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta} + \gamma T_i + (\rho/\sigma_u)(T_i - \mathbf{z}_i'\boldsymbol{\alpha})}{\sqrt{1-\rho^2}}\right)\frac{1}{\sqrt{1-\rho^2}}\begin{pmatrix} \boldsymbol{\beta} \\ \gamma + \rho/\sigma_u \\ -(\rho/\sigma_u)\boldsymbol{\alpha} \end{pmatrix}.$$

17.6.2.d Residual Inclusion and Control Functions

A further simplification of the log-likelihood function is obtained by writing

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

$\tilde{u}_i = (T_i - \mathbf{z}_i'\boldsymbol{\alpha})/\sigma_u$. This "residual inclusion" form suggests a two-step approach. The parameters in the linear regression, $\boldsymbol{\alpha}$ and σ_u, can be consistently estimated by a linear regression of T on $\mathbf{z}$. The scaled residual $\hat{\tilde{u}}_i = (T_i - \mathbf{z}_i'\mathbf{a})/s_u$ 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, so it can be ignored. The second-step log likelihood is, then,

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

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. Because this is a two-step estimator, the resulting estimator of the asymptotic covariance matrix would be adjusted using the Murphy and Topel (2002) results in Section 14.7. Bootstrapping the entire apparatus (i.e., both steps—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 suggests how the residual inclusion method proceeds. The general approach of residual inclusion for nonlinear models with endogenous variables is explored in detail by Terza, Basu, and Rathouz (2008).

17.6.2.e A Control Function Estimator

In the residual inclusion estimator noted earlier the endogeneity of T in the probit model is mitigated by adding the estimated residual to the equation—in the presence

[48] Recent applications of this estimator have referred to it as *instrumental variable probit* estimation. The estimator is a full information maximum likelihood estimator.

of the residual, T is no longer correlated with ε. We took this approach in estimating a linear model in Section 8.4.2. Blundell and Powell (2004) label the foregoing the **control function** approach to accommodating the endogeneity. The residual inclusion estimator suggested here was proposed by Rivers and Vuong (1988). 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,

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

Note that (17-36) 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% and remains significant, but the coefficient on other income increases by more than tenfold.

Example 17.20 Labor Supply Model

In Examples 5.2, 17.1, and 17.15, we examined a labor supply model for married women using Mroz's (1987) data on labor supply. The wife's labor force participation equation suggested in Example 17.15 is

$$\text{Prob}[\mathit{LFP} = 1] = F(\mathit{Constant}, \mathit{Other\ Income}, \mathit{Education}, \mathit{Experience}, \mathit{Experience}^2,$$
$$\mathit{Age}, \mathit{Kids\ Under}\ 6, \mathit{Kids}\ 6\ \mathit{to}\ 18).$$

The *Other Income* (non-wife's) would likely be jointly determined with the LFP decision. We model this with

$$\mathit{Other\ Income} = \alpha_1 + \alpha_2\, \mathit{Husband's\ Age} + \alpha_3\, \mathit{Husband's\ Education} + \alpha_4\, \mathit{City}$$
$$+ \alpha_5\, \mathit{Kids\ Under}\ 6 + \alpha_6\, \mathit{Kids}\ 6\ \mathit{to}\ 18 + u.$$

As before, we use the Mroz (1987) labor supply data described in Example 5.2. Table 17.15 reports the naïve single-equation and full information maximum likelihood estimates of the parameters of the two equations. The third set of results is the two-step estimator detailed in Section 17.6.2d. Standard errors for the maximum likelihood estimators are based on the derivatives of the log-likelihood function. Standard errors for the two-step estimator are computed using 50 bootstrap replications. (Both steps are computed for the bootstrap replications.)

Comparing the two sets of probit estimates, it appears that the (assumed) endogeneity of the *Other Income* is not substantially affecting the estimates. The results are nearly the same. There are two 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 $(0.18777/0.13625)^2 = 1.378^2 = 1.899$. The critical value from the chi-squared table for one degree of freedom would be 3.84, so we would not reject the hypothesis of exogeneity. 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 $-401.30 + (-2{,}844.103) = -3{,}245.405$. The

TABLE 17.15 Estimated Labor Supply Model

	Probit		*FIML*			*2-Step Control Function*	
Variable	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*	*APE*	*Estimate*	*Std. Err.*
LFP Equation for Wife							
Constant	0.27008	0.50859	0.21277	0.51736		0.21811	0.50719
Education	0.13090	0.02525	0.14571	0.02689	0.05693	0.14816	0.02900
Experience	0.12335	0.01872	0.12299	0.01851	0.04805	0.12521	0.01868
Experience^2	−0.00189	0.00060	−0.00192	0.00060	−0.00075	−0.00196	0.00053
Age	−0.05285	0.00848	−0.04878	0.00951	−0.01906	−0.04970	0.00914
Kids Under 6	−0.86833	0.11852	−0.83049	0.12684	−0.32447	−0.84568	0.13693
Kids 6–18	0.03600	0.04348	0.04781	0.04214	0.01868	0.04855	0.05240
Non-wife Inc.	−0.01202	0.00484	−0.02761	0.01254	−0.01079	−0.02798	0.01500
Residual						0.01795	0.01572
Non-wife Income Equation							
Constant			−10.6816	4.34481		−10.5492	
Hus. Age			0.23009	0.07089		0.22818	
Hus. Education			1.35361	0.12978		1.34613	
City			3.54202	0.91338		3.62319	
Kids Under 6			1.36755	0.67056		1.36403	
Kids 6–18			0.67856	0.36160		0.67573	
σ			10.5708	0.15966		10.61312	
ρ			0.18777	0.13625			
ln *L*	−401.302		−3244.556			−2844.103	

log likelihood for the combined model is −3,244.556. Twice the difference is 0.849, which is also well under the 3.84 critical value, so on this basis as well, we would not reject the null hypothesis that $\rho = 0$. As would now be expected, the three sets of estimates are nearly the same. The estimate of −0.02761 for the coefficient on *Other Income* implies that a $1,000 increase reduces the LFP by about 0.028. Because the participation rate is about 0.57, the $1,000 increase suggests a reduction in participation of about 4.9%. The mean value of other income is roughly $20,000, so the 5% increase in *Other Income* is associated with a 5% decrease in LFP, or an elasticity of about one.

17.6.3 ENDOGENOUS SAMPLING

We have encountered several instances of nonrandom sampling in the binary choice setting. In Example 17.17, 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 skewed in favor of observations with $C = 1$ to enrich the data set. A second type of nonrandom sampling arises in the analysis of nonresponse/attrition in the GSOEP in Example 17.29 below. Here, 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 construct a model for the second of these that relies on an assumption of selection on a set of certain observables—the variables that enter 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 17.17. In that application, we examined a binary choice model,

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

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($\textit{Default} = 1|\mathbf{z}, C = 1$). 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 of cardholders is not randomly drawn from the full population—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

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

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

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

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

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

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 17.17 and continued in Example 17.21 and hundreds of applications since.

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 17.6.1. Partial effects and post-estimation 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 17.9.3.

Example 17.21 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% of the applicants are cardholders. In the population, at that time, the true proportion was roughly 23.2%, 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.[49] 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%. This is slightly less than the population rate at the time, 10.3%. 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

[49] See Greene (1992).

TABLE 17.16 Estimated Joint Cardholder and Default Probability Models

	Endogenous Sample Model			*Uncorrelated Equations*		
Variable/Equation	*Estimate*	*Std. Error*	*(t)*	*Estimate*	*Std. 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)	0.00000		
Log Likelihood	−8,660.90650			−8,670.78831		

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 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 17.16. These are broadly consistent with the earlier results—the models with no correlation from Example 17.9 are repeated in Table 17.16. 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 17.16 is $2(8{,}670.78831 - 8{,}660.90650) = 19.76362$. 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.

17.7 PANEL DATA MODELS

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 large, high-quality panel data sets on microeconomic behavior has supported 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

$$\begin{aligned} y^*_{it} &= \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \dots, n, t = 1, \dots, T_i, \\ y_{it} &= \mathbf{1}(y^*_{it} > 0). \end{aligned} \tag{17-38}$$

Most of the interesting cases to be analyzed will start from our familiar common effects model,

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

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 of *strict exogeneity*, that $f(u_i|\mathbf{X}_i)$ is not dependent on $\mathbf{X}_i$, 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 the **fixed effects model**. As before, the distinction does not relate to any intrinsic characteristic of the effect itself.

As we shall see shortly, this modeling framework 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; the fixed effects model relaxes these assumptions, but the natural estimator in this case encounters an **incidental parameters problem** that renders the maximum likelihood estimator inconsistent even when the model is correctly specified.

17.7.1 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.[50] If the fixed effects model is appropriate, then results for omitted variables, including the Yatchew and Griliches (1984) result, apply. The pooled MLE that ignores fixed effects will be inconsistent—possibly wildly so. (*Note:* Because 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 new variable with unknown properties.)

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

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

[50] We could begin the analysis by establishing the assumptions within which we can estimate the parameters of interest ($\boldsymbol{\beta}$) by treating the panel as a long cross section. The point of the exercise, however, is that those assumptions are unlikely to be met in any realistic application.

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.[51] This would be a pseudo MLE because 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 $y_{it}|\mathbf{x}_{it}$. (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

$$PE(\mathbf{x}_{it}, u_i) = \partial \text{Prob}[y_{it} = 1|\mathbf{x}_{it}, u_i]/\partial \mathbf{x}_{it} = \boldsymbol{\beta} \times f(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i),$$

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 (2010) shows that the end result, assuming normality of both v_{it} and u_i is $E_u[PE(\mathbf{x}_{it}, u_i)] = \boldsymbol{\delta}\phi(\mathbf{x}_{it}'\boldsymbol{\delta})$. 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 14.8.2 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 that might be of interest in its own right; and (3) the FIML estimator is available in contemporary software and is no more difficult to estimate than 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.

17.7.2 RANDOM EFFECTS

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

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

$$E[u_i|\mathbf{X}] = 0; \text{Cov}[u_i, u_j|\mathbf{X}] = \text{Var}[u_i|\mathbf{X}] = \sigma_u^2, \quad \text{if } i = j; 0 \text{ otherwise},$$
$$\text{Cov}[v_{it}, u_j|\mathbf{X}] = 0 \text{ for all } i, t, j,$$

[51] This result is explored at length in Wooldridge (2010).

and $\mathbf{X}$ indicates all the exogenous data in the sample, $\mathbf{x}_{it}$ for all i and t.[52] Then,

$$E[\varepsilon_{it}|\mathbf{X}] = 0,$$
$$\text{Var}[\varepsilon_{it}|\mathbf{X}] = \sigma_v^2 + \sigma_u^2 = 1 + \sigma_u^2,$$

and

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

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

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

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,

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

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

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

So,

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

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

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

Inserting this result in (17-40) produces

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

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:

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

[52] See Wooldridge (2010) for discussion of this strict exogeneity assumption.

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

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

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 form for random effects for the binary choice model. Collecting all the terms, we have reduced it to

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

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 how to do the outer integration. **Butler and Moffitt's quadrature method** assuming that u_i is normally distributed is detailed in Section 14.14.4.

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.[53]

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,

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

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

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

Note that u_i is free of any unknown parameters. For example, for normally distributed u, by this transformation, θ is σ_u and now, $u \sim N[0, 1]$. The function is smooth, continuous, and

[53] See Greene (2007b).

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

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

This function is maximized with respect to $\boldsymbol{\beta}$ 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.

For testing the hypothesis of the restricted, pooled model, a Lagrange multiplier approach that does not require estimation of the full random effects model will be attractive. Greene and McKenzie (2015) derived an LM test specifically for the random effects model. Let λ_{it} equal the derivative with respect to the constant term under H_0, defined in (17-20), and let $\tau_{it} = -(q_{it}\mathbf{x}_{it}'\boldsymbol{\beta})\lambda_{it} - \lambda_{it}{}^{2}$. Then,

$$\mathbf{g}_i = \begin{bmatrix} \displaystyle\sum_{t=1}^{T_i} \lambda_{it}\mathbf{x}_{it} \\ \displaystyle\frac{1}{2}\left(\sum_{t=1}^{T_i}\tau_{it}\right) + \frac{1}{2}\left(\sum_{t=1}^{T_i}\lambda_{it}\right)^2 \end{bmatrix}.$$

Finally, $\mathbf{g}_i'$ is the ith row of the $n \times (K + 1)$ matrix $\mathbf{G}$. The LM statistic is LM $= \mathbf{i}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{i} = nR^2$ in the regression of a column of ones on $\mathbf{g}_i$. The first K elements of $\mathbf{i}'\mathbf{G}$ equal zero as they are the score of the log likelihood under H_0. Therefore, the LM statistic is the square of the $(K + 1)$ element of $\mathbf{i}'\mathbf{G}$ times the last diagonal element of the matrix $(\mathbf{G}'\mathbf{G})^{-1}$. Wooldridge (2010) proposes an omnibus test of the null of the pooled model against the more general model that contains lagged values of $\mathbf{x}_{it}$ and/or y_{it}. The two steps of the test are: (1) Pooled probit estimation of the null model; and (2) Pooled probit estimation of the augmented model Prob$(y_{it} = 1) = \Phi(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma u_{i,t-1})$ based on observations $t = 2, \ldots, T_i$ where $u_{it} = (y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta})$. The test is a simple Wald, LM, or LR test of the hypothesis that γ equals zero.

We have examined two approaches to estimation of a probit model with random effects. GMM estimation is a third 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 17.36 will use the Bertschek and Lechner data.)

17.7.3 FIXED EFFECTS

The fixed effects model is

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

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 will 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 $(Y_{it} = y_{it} \mid \mathbf{x}_{it}) = P(q_{it}z_{it})$.

In our previous application of this model, in the linear regression case, we found that estimation of the parameters was simplified 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.[54] The method for estimation of nonlinear fixed effects models such as the probit and logit models is detailed in Section 14.9.6.d.[55]

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).] 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%, 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%, 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 Section 14.9.6.d

[54] See, for example, Maddala (1987), p. 317.

[55] Fernandez-Val (2009) reports using that method to fit a probit model for 500,000 groups.

is one of these special cases.) The available mix of theoretical results and Monte Carlo evidence suggests that for binary choice estimation of static models, plim $\hat{\boldsymbol{\beta}}_{FE} = S(T)\boldsymbol{\beta}$ where $S(2) = 2$, $S(T + 1) < S(T)$ and $\lim_{T\to\infty} S(T) = 1$.[56] The issue is much less clear for dynamic models—there is little small T wisdom, though the large T result appears to apply as well.

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).

17.7.3.a A Conditional Fixed Effects Estimator

Why does the incidental parameters problem arise here and not in the linear regression model?[57] 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 $f(y_{it}|\mathbf{X}_i)$ is a function of α_i, $f(y_{it}|\mathbf{X}_i, \bar{y}_i)$ 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

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

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

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

[56] For example, Hahn and Newey (2002), Fernandez-Val (2009), Greene (2004), Katz (2001), Han (2002) and others.

[57] 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 $S_i = \sum_{t=1}^{T_i} y_{it}$.[58]

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.[59] When we take logs, these terms (and these observations) will drop out. But suppose that $y_{i1} = 0$ and $y_{i2} = 1$. Then

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

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 (1, 0) 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.[60] Hausman's (1978) specification test is a natural one to use here, however. Under the null hypothesis of homogeneity, both Chamberlain's conditional maximum likelihood

[58] 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.

[59] In the probit model when we encounter this situation, the individual constant term cannot be estimated and the group is removed from the sample. The same effect is at work here.

[60] 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.

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,[61] whereas Chamberlain's estimator is consistent and efficient. The Hausman test can be based on the chi-squared statistic,

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

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 17.22 Binary Choice Models for Panel Data

In Example 17.6, we fit a pooled binary logit model $y = \mathbf{1}(DocVis > 0)$ 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 ranging from 1 to 7. Table 17.17 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 marked "U" is the unconditional (inconsistent) estimator. The one marked "C" is Chamberlain's consistent estimator. Note for all three fixed effects estimator 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% 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 17.18. The average of the fixed effects constant terms is used to obtain a constant term for the unconditional fixed effects case. No estimator is available for the conditional 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.

Example 17.23 Fixed Effects Logit Model: 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

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

TABLE 17.17 Estimated Parameters for Panel Data Binary Choice Models

			Variable					
Model	***Estimate***	**ln *L***	***Constant***	***Age***	***Income***	***Kids***	***Education***	***Married***
Logit	$\boldsymbol{\beta}$		**0.25112**	**0.02071**	**−0.18630**	**−0.22947**	**−0.04557**	**0.08530**
Pooled	St. Err.	−17673.09	0.09114	0.00129	0.07509	0.02954	0.00565	0.03328
	Rob.SE[a]		0.12827	0.00174	0.09160	0.03831	0.00808	0.04531
Logit R.E.	$\boldsymbol{\beta}$	−16277.04	**0.06447**	**0.03416**	**0.00237**	**−0.26127**	**−0.05786**	**0.02707**
$\rho = 0.41503$	St. Err.		0.16391	0.00225	0.11299	0.04589	0.01071	0.05328
Logit	$\boldsymbol{\beta}$	−9452.55		**0.10469**	**−0.05712**	**−0.08828**	**−0.11673**	**−0.05761**
F.E.(U)[b]	St. Err.			0.00726	0.17844	0.07447	0.06880	0.10619
Logit	$\boldsymbol{\beta}$	−6299.02		**0.08471**	**−0.04732**	**−0.07767**	**−0.09084**	**−0.05229**
F.E.(C)[c]	St. Err.			0.00650	0.15891	0.06228	0.05668	0.09304
Probit	$\boldsymbol{\beta}$	−17670.93	**0.15501**	**0.01283**	**−0.11666**	**−0.14118**	**−0.02811**	**0.05226**
Pooled	St. Err.		0.05652	0.00079	0.04635	0.01822	0.00350	0.02046
	Rob.SE[a]		0.07959	0.00107	0.05647	0.02361	0.00501	0.02790
Probit:RE[d]	$\boldsymbol{\beta}$	−16273.96	**0.03410**	**0.02014**	**−0.00267**	**−0.15377**	**−0.03371**	**0.01629**
$\rho = 0.44788$[e]	St. Err.		0.09635	0.00132	0.06670	0.02704	0.00629	0.03135
Probit:RE[f]	$\boldsymbol{\beta}$	−16274.06	**0.03447**	**0.02013**	**−0.00261**	**−0.15359**	**−0.03379**	**0.01749**
$\rho = 0.44768$	St. Err.		0.06337	0.00090	0.05212	0.02030	0.00394	0.02280
Probit	$\boldsymbol{\beta}$	−9453.47		**0.06249**	**−0.03155**	**−0.04818**	**−0.07222**	**−0.03298**
F.E.(U)	St. Err.			0.00432	0.10749	0.04457	0.04074	0.06364

[a] Robust, “cluster” corrected standard error.

[b] Unconditional fixed effects estimator.

[c] Conditional fixed effects estimator.

[d] Butler and Moffitt estimator.

[e] Probit LM statistic = 1011.43.

[f] Maximum simulated likelihood estimator.

TABLE 17.18 Estimated Partial Effects for Panel Data Binary Choice Models

Model	*Age*	*Income*	*Kids*	*Education*	*Married*
Logit, P[a]	0.00472	−0.04238	−0.05272	−0.01037	0.01951
Logit: RE,Q[b]	0.00705	0.00049	−0.05461	−0.01193	0.00560
Logit: F,U[c]	0.02570	−0.01402	−0.02167	−0.02865	−0.01404
Logit: F,C[d]	—	—	—	—	—
Probit, P[a]	0.00475	−0.04315	−0.05267	−0.01040	0.01942
Probit RE.Q[b]	0.00550	−0.00073	−0.04226	−0.00920	0.00445
Probit:RE,S[e]	0.00694	−0.00090	−0.05362	−0.01166	0.00605
Probit: F,U[c]	0.01312	−0.00662	−0.01012	−0.01516	−0.00688

[a] Pooled estimator.
[b] Butler and Moffitt estimator.
[c] Unconditional fixed effects estimator.
[d] Conditional fixed effects estimator. Partial effects not computed.
[e] Maximum simulated likelihood estimator.

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

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

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 17.19.

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 Section 14.15.7) in which the classes are defined by different constant terms—the remaining parameters in the model 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 17.19 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 17.19. He reported 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

TABLE 17.19 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

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.

17.7.3.b 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 coefficients. 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 (2009) approach that involves correcting the bias in the FE MLE.

The Mundlak (1978) approach[62] augments (17-44) as follows:

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

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

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

(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 $\overline{\mathbf{x}}_i$ produces a random effects formulation. As in the

[62] See also Chamberlain (1984) and Wooldridge (2010).

linear model (see Sections 11.5.6 and 11.5.7), it also suggests a means of testing for fixed versus random effects. Because $\boldsymbol{\delta} = \mathbf{0}$ produces the pure random effects model, a joint Wald test of the null hypothesis that $\boldsymbol{\delta}$ equals zero can be used.

Example 17.24 Panel Data Random Effects Estimators

Example 17.22 presents several 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 17.20 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 113.35 with 5 degrees of freedom. The critical value from the chi-squared table for 95% significance is 11.07, so the hypothesis that $\boldsymbol{\delta}$ equals zero, that is, the hypothesis of the random effects model (restrictions), is rejected. The two log likelihoods are −16,273.96 for the REM and −16,222/04 for the augmented REM. The LR statistic would be twice the difference, or 103.4. 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($\hat{\boldsymbol{\beta}}_{FE,MLE} - \boldsymbol{\beta}$) 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 FE, MLE 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

TABLE 17.20 Estimated Random Effects Models

	Basic Random Effects		*Mundlak Formulation*			
	Estimate	*Std. Error*	*Estimate*	*Std. Error*	*Mean*	*Std. Error*
Constant	0.03410	(0.09635)	0.37496	(0.10501)		
Age	0.02014	(0.00132)	0.05032	(0.00357)	−0.03656	(0.00384)
Income	−0.00267	(0.06770)	−0.02863	(0.09325)	−0.35365	(0.13991)
Kids	−0.15377	(0.02704)	−0.04195	(0.03752)	−0.22516	(0.05499)
Education	−0.03371	(0.00629)	−0.05450	(0.03307)	0.02391	(0.03374)
Married	0.01629	(0.03135)	−0.02661	(0.05180)	0.14689	(0.06606)

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 well 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 17.7.4), where the problem is possibly more severe than in the static case. Research in this area is ongoing.

17.7.4 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 Arellano (2001) suggest 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.[63]

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 so 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,

$$\begin{gathered}\text{Prob}(y_{i1} = 1 \,|\, \mathbf{x}_{i1}, \mathbf{z}_i, \alpha_i) = \Phi(\mathbf{x}_{i1}'\boldsymbol{\delta} + \mathbf{z}_i'\boldsymbol{\tau} + \theta\alpha_i)\\ \text{Prob}(y_{it} = 1 \,|\, \mathbf{x}_{it}, y_{i,t-1}, \alpha_i) = \Phi(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \alpha_i), t = 2, \ldots, T_i,\end{gathered}$$

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

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

[63] 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 Fernandez-Val (2009). Stewart (2006) and Arulampalam and Stewart (2007) provide several results for practitioners.

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 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 $D_{it} = 1$ and $\gamma = \theta - 1$ in period 1 and 0 in every other period, $C_{it} = 1 - D_{it}$. Then, the two parts may be combined in

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

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,

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

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

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

The source of the instruments $\mathbf{z}_i$ is unclear. Wooldridge (2005) simplifies the model a bit by using, instead, a Mundlak approach, using the group means of the time-varying variables as $\mathbf{z}$. The resulting random effects formulation is

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

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

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

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.[64] 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 17.25 A Dynamic Model for Labor Force Participation and Disability

Gannon (2005) modeled the relationship between labor force participation and disability in Ireland with a panel data set, *The Living in Ireland Survey 1995–2000*. The sample begins in 1995 with 7,254 individuals, but with attrition, shrinks to 3,670 in 2000. The dynamic probit model is

$$y_{it}^* = b_0 + b_1 y_{i,t-1} + b_2 D_{it} + b_3 D_{i,t-1} + b_4 z_{it} + \alpha_i + \varepsilon_{it},\ y_{it} = \mathbf{1}(y_{it}^* > 0),$$

where y_{it} is the labor force participation indicator and D_{it} is an indicator of disability. The related covariates are gathered in z_{it}. The lagged value of D_{it} helps distinguish longer-term disabilities from those recently acquired. Unobserved time-invariant individual effects are captured by the common effect, α_i. The lagged dependent variable helps distinguish between the impact of the individual effect and the inertia of past participation. Variables in z_{it} include age, residence region, education, marital status, children, and unearned income.

The starting point of the analysis is a pooled probit model without the common effect (with standard errors corrected for the clustering at the individual level). The pooled model leaves two interesting questions:

1. Do the control variables adequately account for the unobserved characteristics?
2. Does past disability affect participation directly as in the model, or through some different channel that affects past participation?

The author adopts Wooldridge's (2005) (Mundlak) form of the random effects model we examined in Section 17.7.3.b and Example 17.24 to deal with the unobserved heterogeneity and the initial conditions problem. Thus, the initial value of y_{it} and the group means of time-varying variables are added to the random effects model,

$$y_{it}^* = b_1 y_{i,t-1} + b_2 D_{it} + b_3 D_{i,t-1} + b_4 z_{it} + \alpha_0 + \alpha_1 y_{i0} + \boldsymbol{\alpha}_2' \bar{\mathbf{x}}_i + a_i + \varepsilon_{it},\ y_{it} = \mathbf{1}(y_{it}^* > 0).$$

The resulting model is now estimated using the Butler and Moffitt method for random effects.

Example 17.26 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

[64] 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) 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).

study were the years 1979–1985 of the *Panel Study of Income Dynamics*. A sample of 1,812 continuously married couples was 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 to 2, 3 to 5, and 6 to 17 years old. Hyslop's formulation, in general terms, is

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

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

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

(stochastic specification)

$$\begin{gathered}\eta_i \,|\, \mathbf{X}_i \sim \text{N}[0, \sigma_\eta^2], \\ v_{i0} \,|\, \mathbf{X} \sim \text{N}[0, \sigma_0^2], \\ w_{it} \,|\, \mathbf{X}_i \sim \text{N}[0, \sigma_\text{w}^2], \\ v_{it} = \rho v_{i,t-1} + w_{it}, \sigma_\eta^2 + \sigma_w^2 = 1, \\ \text{Corr}[v_{i0}, v_{it}] = \rho^t, \quad t = 1, \ldots, T-1.\end{gathered}$$

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

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

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.

17.7.5 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 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 Example 17.23, Heckman and Singer's (1984b) model provides a less stringent 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 17.27 Semiparametric Models of Heterogeneity

We have extended the random effects and fixed effects logit models in Example 17.22 by fitting the Heckman and Singer (1984b) model. Table 17.21 shows the specification search and the results under different specifications. The first column of results shows the estimated fixed effects model from Example 17.22. 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

TABLE 17.21 Estimated Heterogeneity Models

		Number of Classes				
	Fixed Effect	***1***	***2***	***3***	***4***	***5***
β_1	0.10475	0.02071	0.03033	0.03368	0.03408	0.03416
	(0.08476)					
β_2	−0.06097	−0.18592	0.02555	−0.00580	−0.00635	−0.01363
	(−0.05038)					
β_3	−0.08841	−0.22947	−0.24708	−0.26388	−0.26590	−0.26626
	(−0.07776)					
β_4	−0.11671	−0.04559	−0.05092	−0.05802	−0.05975	−0.05918
	(−0.09082)					
β_5	−0.05732	0.08529	0.04297	0.03794	0.02923	0.03070
	(−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	−0.03674	−1.96750
				(0.46749)	(0.46341)	(0.19573)
α_4					−4.03970	−0.25588
					(0.02636)	(0.40930)
α_5						−6.48191
						(0.01396)
Mean	−2.62334	0.25111	0.02361	0.05506	0.06369	0.05471
Std. Dev.	3.13415	0.00000	1.15866	1.40723	1.48707	1.62143
ln *L*	−9458.638	−17673.10	−16353.14	−16278.56	−16276.07	−16275.85
	(−6299.02)					
AIC/N	1.00349	1.29394	1.19748	1.19217	1.19213	1.19226

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 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 14.15.5, 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.

17.7.6 MODELING PARAMETER HETEROGENEITY

In Section 11.10, 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 14.15 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 14.17, has been on linear models.

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

$$\begin{aligned} y_{it}^* &= \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T_i, \\ y_{it} &= \mathbf{1}(y_{it}^* > 0), \end{aligned}$$

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

$$\begin{aligned} y_{it}^* &= \mathbf{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T_i, \\ y_{it} &= \mathbf{1}(y_{it}^* > 0), \end{aligned}$$

where $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Gamma}\mathbf{u}_i$ and $\boldsymbol{\Gamma}$ 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 17.28. Example 17.28 also presents a similar model that assumes that the distribution of $\boldsymbol{\beta}_i$ is discrete rather than continuous.

Example 17.28 Parameter Heterogeneity in a Binary Choice Model

We have extended the logit model for doctor visits from Example 17.14 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:

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

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 17.27 to find a preferred specification. Table 17.22 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.[65])

Figure 17.5 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

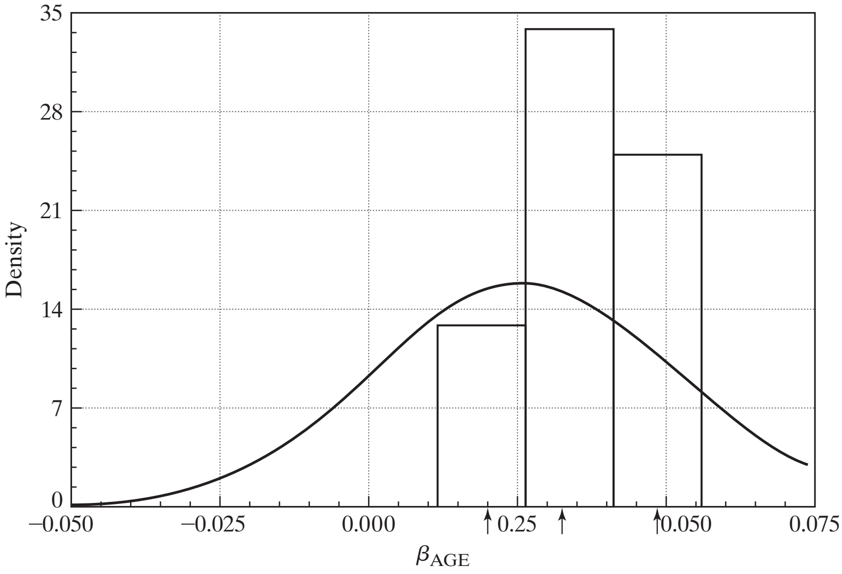
[65] Nonreplicability is an ongoing challenge in empirical work in economics. (See, for instance, Example 17.14.) The problem is particularly acute in analyses that involve simulation such as Monte Carlo studies and random parameter models. In the interest of replicability, we note that the random parameter estimates in Table 17.22 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.

TABLE 17.22 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	−0.03496	0.81651	0.96605	−0.18579	−1.52595
	(0.09114)	(0.07553)	(0.01654)	(0.43757)	(0.23907)	(0.43498)
Age	0.02071	0.02631	0.02533	0.04906	0.03225	0.01998
	(0.00129)	(0.00110)	(0.00042)	(0.00695)	(0.00315)	(0.00626)
Income	−0.18592	−0.00436	0.10737	−0.27917	−0.06863	0.45487
	(0.07506)	(0.06245)	(0.03828)	(0.37149)	(0.16748)	(0.31153)
Kids	−0.22947	−0.17461	0.55520	−0.28385	−0.28336	−0.11708
	(0.02954)	(0.02452)	(0.02387)	(0.14279)	(0.06640)	(0.12363)
Education	−0.04559	−0.04051	0.03792	−0.02530	−0.05734	−0.09385
	(0.00565)	(0.00475)	(0.00134)	(0.02777)	(0.01247)	(0.02797)
Married	0.08529	0.01462	0.07070	−0.10875	0.02533	0.23571
	(0.03329)	(0.027417)	(0.01736)	(0.17228)	(0.07593)	(0.14369)
Class	1.00000	1.00000		0.34833	0.46181	0.18986
Prob.	(0.00000)	(0.00000)		(0.03850)	(0.02806)	(0.02234)
ln *L*	−17673.10	−16271.72		−16265.59		

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, 0.461, between the midpoints of the intervals between the center mass point and the two extremes. With a width

FIGURE 17.5 Distribution of AGE Coefficient.



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.

17.7.7 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;
- 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, see Section 4.7.4) 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 $E[h|\mathbf{x}, R] = E[h|\mathbf{x}]$. 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, (**x**,**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 $P(R = 1 | h, \mathbf{x}, \mathbf{z}) = P(R = 1 | \mathbf{x}, \mathbf{z})$. 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 **x** and **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: $P(R = 1 | h, \mathbf{x}, \mathbf{z}) = P(R = 1 | \mathbf{z})$ where **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, $P(R_{i1} = 1 | \mathbf{x}_{i1}, \mathbf{z}_{i1}) = 1$. Wooldridge suggests using a probit model for $P(R_{it} = 1 | \mathbf{x}_{i1}, \mathbf{z}_{i1}), t = 2, \ldots, T$ 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, because the first step is not used to produce an estimated parameter vector in the second. Wooldridge (2002) shows that the standard errors computed without the adjustment are "conservative" in that they are larger than they would be with the adjustment.

Example 17.29 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 17.6. 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 17.6 could be a bit misleading—the nonresponse pattern is quite different from simple attrition. For example, 364 of the 3,874 individuals who responded at wave 1 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 17.23. 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:

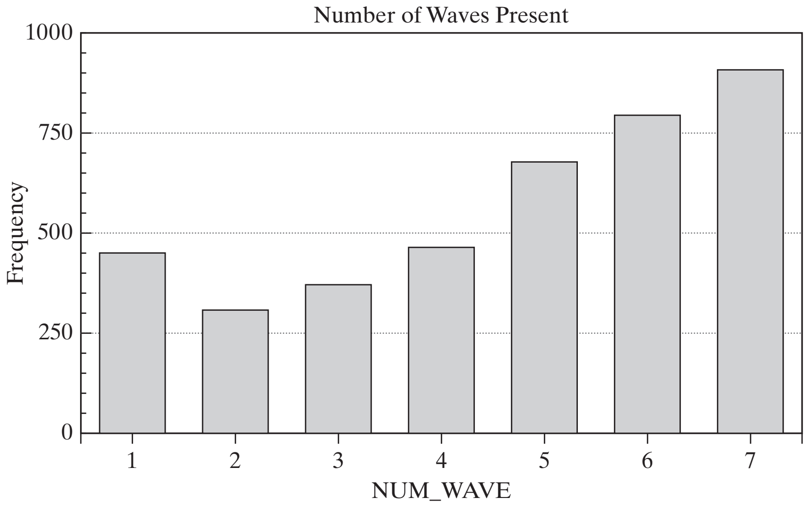
$$\begin{aligned} x_{i1} &= \textit{constant}, \textit{age}, \textit{income}, \textit{educ}, \textit{kids}, \textit{married} \\ z_{i1} &= \textit{female}, \textit{handicapped dummy}, \textit{percentage handicapped}, \\ &\quad \textit{university}, \textit{working}, \textit{blue collar}, \textit{white collar}, \textit{public servant}, y_{i1} \\ y_{i1} &= \textit{DoctorVisits} > 0 \textit{ in period } 1. \end{aligned}$$

This first-year data vector is used as the observed explanatory variables in probit models for waves 2 to 7 for the 3,874 individuals who were present at wave 1. There are 3,874 observations for each of these probit models, because all were observed at wave, 1. Fitted probabilities for R_{it} are computed for waves 2 to 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 17.23 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 17.23 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		

FIGURE 17.6 Number of Waves Responded for Those Present at Wave 1.



17.9 SPATIAL BINARY CHOICE MODELS

Section 11.7 presented a model of spatial interaction among sample observations. In an application, Bell and Bockstael (2000) constructed a spatial hedonic regression model of house prices that were influenced by attributes and by neighborhood effects. We considered two frameworks for the regression model: spatial autoregression (SAR),

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \rho\Sigma_{j=1}^{n} w_{ij}y_j + \varepsilon_i, \text{ or, for all } n \text{ observations, } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon},$$

and spatial autocorrelation (SAC),

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \text{ where } \varepsilon_i = \rho\Sigma_{j=1}^{n} w_{ij}\varepsilon_j + u_i, \text{ or } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} = \rho\mathbf{W}\boldsymbol{\varepsilon} + \mathbf{u}.$$

Both cases produce a generalized regression model with full $n \times n$ covariance matrix when y is a continuous random variable. The model frameworks turn on the crucial spatial correlation parameter, ρ, and the specification of the contiguity matrix, **W**, which defines the form of the spatial correlation. In Bell and Bockstael's application, in the sample of 1,000 home sales, the elements of **W** (in one of several specifications) are

$$W_{ij} = \frac{\mathbf{1}(\text{Home } i \text{ and } j \text{ are } < 600 \text{ meters apart})}{\text{Distance between homes } i \text{ and } j}; W_{ii} = 0.$$

(The rows of **W** are standardized.) Conditioned on the value of ρ, this produces a generalized regression model that is estimated by GMM or maximum likelihood.

We are interested in extending the idea of spatial interaction to a binary outcome.[66] Some received examples are:

- Garrett, Wagner, and Wheelock (2005) examined banks' choices of branch banking;
- McMillen (1992) examined factors associated with high (or low) crime rates in neighborhoods of Columbus, Ohio;

[66] Smirnov (2010) provides a survey of applications of spatial models to nonlinear regression settings.

- Pinske and Slade (2006) examined operation decisions (open/closed) for a panel of copper mines;
- Flores-Lagunes and Schnier (2012) extended Heckman's (1979) two-step estimator to include spatial effects in both the selection (probit) and regression steps. They apply the method to a sample of 320 observations on trawl fishing in which only 207 are fully reported (selected).
- Klier and McMillen (2008) analyzed county-wide data on auto supply plant location decisions in the U.S. Midwest. An industry that serviced the auto manufacturing centered around Detroit was earlier oriented west-east from Chicago to New York. During the mid-20th century, entry took place along an axis running from south to north (along with an historic internal migration in the U.S. that accompanied the decline of the coal industry). Klier and McMillen examined data on counties and whether an auto supplier was located in the county, a binary outcome.

The model framework is a binary choice model,

$$y_i* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, y_i = \mathbf{1}(y_i* > 0).$$

The distribution for most applications will be the normal or logistic leading to a probit or logit model. A model of spatial autoregression would be

$$y_i* = \mathbf{x}_i'\boldsymbol{\beta} + \rho\Sigma_{j=1}^n w_{ij}y_j* + \varepsilon_i, y_i = \mathbf{1}(y_i* > 0).$$

Based on a random utility interpretation, it would be difficult to motivate spatial interaction based on the latent utilities.[67] The spatial autoregression model based on the observed outcomes instead would be

$$y_i* = \mathbf{x}_i'\boldsymbol{\beta} + \rho\Sigma_{j=1}^n w_{ij}y_j* + \varepsilon_i, y_i = \mathbf{1}(y_i* > 0).$$

This might seem more reasonable; however, this model is incoherent—it is not possible to insure that Prob$(y_i = 1|\mathbf{x}_i)$ lies between zero and one. A spatial error model used in several applications is

$$y_i* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i; \varepsilon_i = \rho\Sigma_{j=1}^n w_{ij}\varepsilon_j + u_i, u_i \sim \text{N}[0, 1], y_i = \mathbf{1}(y_i* > 0).$$

Pinske and Slade (1998, 2006) and McMillen (1992) use this framework to construct a GMM estimator based on the generalized residuals, λ_i, defined in (17-20). Solving for the reduced form,

$$\boldsymbol{\varepsilon} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u}.$$

The full covariance matrix for the n observations would be

$$\text{Var}[\boldsymbol{\varepsilon}] = \sigma_u^2[(\mathbf{I} - \rho\mathbf{W})'(\mathbf{I} - \rho\mathbf{W})]^{-1} = \sigma_u^2\mathbf{D}(\rho).$$

(Note that $\sigma_u^2 = 1$.) Then,

$$y_i* = \mathbf{x}_i'\boldsymbol{\beta} + \Sigma_{j=1}^n \mathbf{D}_{ij}(\rho)u_j, y_i = \mathbf{1}(y_i* > 0).$$

[67] But Klier and McMillen (2008, p. 462) note, "The assumption that the latent variable depends on spatially lagged values of the latent variable may be disputable in some settings. In our example, we are assuming that the propensity to locate a new supplier plant in a county depends on the propensity to locate plants in nearby counties, and it does *not* depend simply on whether new plants have located nearby. The assumption is reasonable in this context because of the forward-looking nature of plant location decisions."

The marginal probability is

$$\text{Prob}(y_i = 1 | \mathbf{x}_i) = \text{Prob}(\mathbf{x}_i'\boldsymbol{\beta} + \Sigma_{j=1}^{n}\mathbf{D}_{ij}(\rho)u_j > 0)$$
$$= F\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\Sigma_{j=1}^{n}[\mathbf{D}_{ij}(\rho)]^2}\right) = F[\mathbf{x}_i^*(\mathbf{D}, \rho)'].$$

This corresponds to the heteroscedastic probit model in Section 17.5.2. (The difference here is that the observations are all correlated.) We have seen two GMM approaches to estimation. Consistent with Bertschuk and Lechner's (1998) approach based on simple regression residuals, the GMM estimator would use $E\{\mathbf{z}_i \times [y_i - \Phi(\mathbf{x}_i{*}(\mathbf{D}, \rho)'\boldsymbol{\beta})]\} = \mathbf{0}$, where $\mathbf{z}_i$ is the set of instrumental variables. McMillen (1992) and Pinske and Slade (2006) use the generalized residuals, here $\lambda_i{*}(\mathbf{D}, \rho)$, defined in (17-20), instead,

$$E\left[\mathbf{z}_i \times \left\{\frac{(y_i - \Phi[\mathbf{x}_i^*(\mathbf{D}, \rho)'\boldsymbol{\beta}])\phi[\mathbf{x}_i^*(\mathbf{D}, \rho)'\boldsymbol{\beta}]}{\Phi[\mathbf{x}_i^*(\mathbf{D}, \rho)'\boldsymbol{\beta}](1 - \Phi[\mathbf{x}_i^*(\mathbf{D}, \rho)'\boldsymbol{\beta}])}\right\}\right] = E[\mathbf{z}_i \times \lambda(\mathbf{x}_i^*(\mathbf{D}, \rho)'\boldsymbol{\beta})] = \mathbf{0}.$$

Pinske and Slade (2006) used a probit model while Klier and McMillen proposed a logit model. The estimation method is largely the same in both cases.

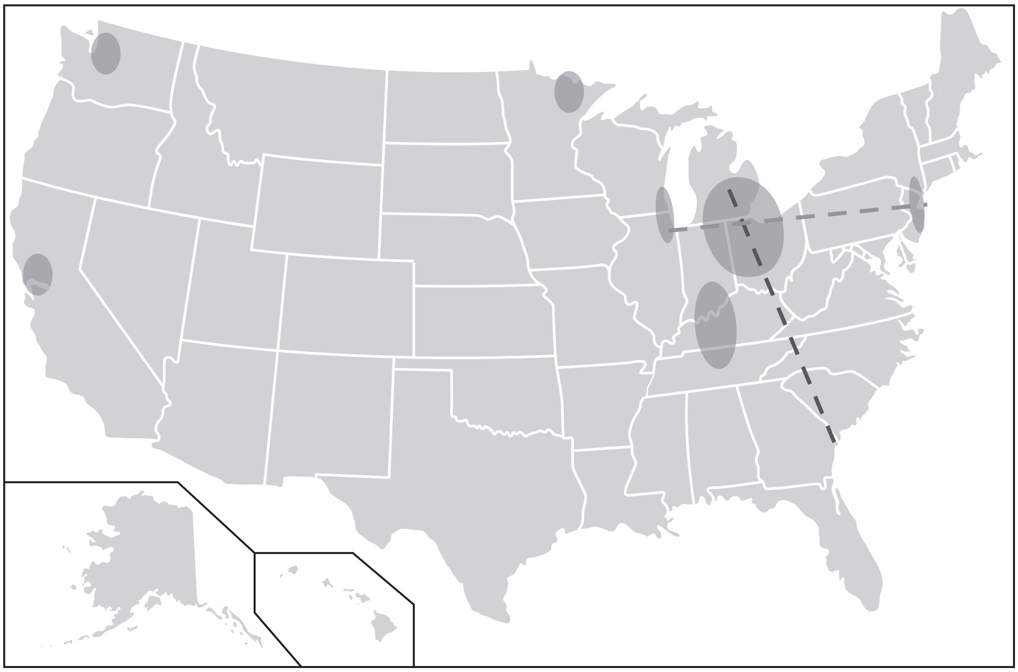
The preceding estimators use an approximation based on the marginal probability to form a feasible GMM estimator. Case (1992) suggests that if the contiguity pattern were compressed so that the data set consists of a finite number of neighborhoods, each with a small enough number of members, then the model could be handled directly by maximum likelihood. It would resemble a panel probit model in this case. Klier and McMillen used this approach to simplify their estimation procedure. Wang, Iglesias, and Wooldridge (2013) proposed a similar approach to an unrestricted model based on the principle of a partial likelihood. By using a spatial moving average for $\boldsymbol{\varepsilon}$, they show how to use pairs of observations to formulate a bivariate heteroscedastic probit model that identifies the spatial parameters.

Example 17.30 A Spatial Logit Model for Auto Supplier Locations

Klier and McMillen (2008) specified a binary logit model with spatial error correlation to model whether a county experienced a new auto supply location in 1991—2003. The data consist of 3,107 county observations. The weighting matrix is initially specified as $1/n_i$ where n_i = the number of counties that are contiguous to county *i*—share a common border. To speed up computation, the weighting matrix is further reduced so that counties are only contiguous if they are in the same census region. This produces a block diagonal **W** that greatly simplifies the estimation. Figure 17.7 [Based on Figure 2 from Klier and McMillen (2008)] illustrates clusters of U.S. counties that experienced entry of new auto suppliers. The east-west oriented line shows the existing focus of the industry. The north-south line (roughly oriented with historical U.S. Route 23) shows the focus of new plants in the years studied. Results for the spatial correlation model are compared to a pooled logit model. The estimated spatial autocorrelation coefficient, ρ, is moderately large (0.425 with a standard error of 0.180), however, the results are similar for the two specifications. For example, one of the central results, the coefficient on *Proportion Manufacturing Employment*, is 6.877 (1.039) in the pooled model and 5.307 (1.224) in the spatial model. The magnitudes of the coefficients are difficult to interpret and partial effects were not computed.[68] The signs are generally consistent with expectations.

[68] Wooldridge (2010) and Wang, Iglesias, and Wooldridge (2013) recommend analyzing Average Structural Functions (ASFs) for the heteroscedastic probit (logit) model considered here. Since the weighting matrix, W, does not involve any exogenous variables, the derivatives of the ASFs will be identical to the average partial effects. (See footnote 40 in Section 17.5.2.)

FIGURE 17.7 Counties with New Plants.



17.9 THE BIVARIATE PROBIT MODEL

In Chapter 10, we analyzed a number of different multiple-equation extensions of the linear 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 form as the seemingly unrelated regressions model. The general specification for a two-equation model would be

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

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 Section 17.6.1 (Examples 17.18 and 17.19), we extended (17-48) to

$$\begin{aligned} &T^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \qquad T = \mathbf{1}(T^* > 0),\\ &y^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \gamma T + \varepsilon_2, \quad y = \mathbf{1}(y^* > 0),\\ &\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix}\Big| \mathbf{x}_1, \mathbf{x}_2 \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right]. \end{aligned} \tag{17-49}$$

This model extends the case in Section 17.6.2, where T^* rather than T appears on the right-hand side of the second equation. In Example 17.35, T 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:

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

In Example 17.21, we studied 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 borrower 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 17.7.7.

In Section 17.10, 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.

17.9.1 MAXIMUM LIKELIHOOD ESTIMATION

The bivariate normal cdf is

$$\text{Prob}(X_1 < x_1, X_2 < x_2) = \int_{-\infty}^{x_2}\int_{-\infty}^{x_1} \phi_2(z_1, z_2, \rho)dz_1dz_2,$$

which we denote $\Phi_2(x_1, x_2, \rho)$. The density is[69]

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

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

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

and

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

[69] See Section B.9.

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,[70]

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

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 (17-20). 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

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

[70] 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

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

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 17.31 Tetrachoric Correlation

Returning once again to the health care application of Example 17.6 and several others, we now consider a second binary variable,

$$Hospital_{it} = \mathbf{1}(HospVis_{it} > 0).$$

Our previous analyses have focused on

$$Doctor_{it} = \mathbf{1}(DocVis_{it} > 0).$$

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 because it is used for continuous variables. Consider, instead, a bivariate probit model,

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

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.

17.9.2 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[71]

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

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 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% 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 17.32 demonstrates.

17.9.3 PARTIAL EFFECTS

There are several partial effects one might want to evaluate in a bivariate probit model.[72] A natural first step would be the derivatives of Prob$[y_1 = 1, y_2 = 1 | \mathbf{x}_1, \mathbf{x}_2]$. These can be deduced from (17-51) 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

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

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

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

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.

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 partial 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 (17-52). 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 probabilities to consider. The marginal probabilities are given by the univariate probabilities,

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

so the analysis of (17-11) and (17-12) applies. One pair of probabilities that might be of interest are

$$\begin{aligned}\text{Prob}[y_1 = 1 \mid y_2 = 1, \mathbf{x}] &= \frac{\text{Prob}[y_1 = 1, y_2 = 1 \mid \mathbf{x}]}{\text{Prob}[y_2 = 1 \mid \mathbf{x}]}\\ &= \frac{\Phi_2(\mathbf{x}'\boldsymbol{\gamma}_1, \mathbf{x}'\boldsymbol{\gamma}_2, \rho)}{\Phi(\mathbf{x}'\gamma_2)}\end{aligned}$$

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

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

Finally, one might construct the probability function,

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

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 17.32 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 17.24 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

$$\lambda_{LR} = 2\{-25{,}285.07 - [-17{,}422.72 + (-8{,}073.604)]\} = 422.508.$$

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 Prob$[y_1 = 1 | y_2 = 1]$. The direct effect is $[g_1/\Phi(\mathbf{x}'\boldsymbol{\gamma}_2)]\boldsymbol{\gamma}_1$ in the definition given earlier. The mean estimate of Prob$[y_1 = 1 | y_2 = 1]$ is 0.821285. In the table in Example 17.31, this would correspond to the raw proportion $P(D = 1, H = 1)/P(H = 1) = (1{,}975/27{,}326)/(2{,}395/27{,}326) = 0.8246$.

TABLE 17.24 Estimated Bivariate Probit Model[a]

	Doctor					*Hospital*	
	Model Estimates		*Partial Effects*			*Model Estimates*	
Variable	*Univariate*	*Bivariate*	*Direct*	*Indirect*	*Total*	*Univariate*	*Bivariate*
Constant	−0.1243	−0.1243				−1.3328	−1.3385
	(0.05815)	(0.05814)				(0.08320)	(0.07957)
Female	0.3559	0.3551	0.09650	−0.00724	0.08926	0.1023	0.1050
	(0.01602)	(0.01604)	(0.00500)	(0.00152)	(0.00513)	(0.02195)	(0.02174)
Age	0.01189	0.01188	0.00323	0.00032	0.00291	0.00461	0.00461
	(0.00080)	(0.00080)	(0.00023)	(0.00007)	(0.00024)	(0.00108)	(0.00106)
Income	−0.1324	−0.1337	−0.03632	−0.00306	−0.03939	0.03739	0.04441
	(0.04655)	(0.04628)	(0.01260)	(0.00411)	0.01254)	(0.06329)	(0.05946)
Kids	−0.1521	−0.1523	−0.04140	0.00105	−0.04036	−0.01714	−0.01517
	(0.01833)	(0.01825)	(0.00505)	(0.00177)	(0.00517)	(0.02562)	(0.02570)
Education	−0.01497	−0.01484	−0.00403	0.00151	−0.00252	−0.02196	−0.02191
	(0.00358)	(0.00358)	(0.00010)	(0.00035)	(0.00100)	(0.00522)	(0.00511)
Married	0.07352	0.07351	0.01998	0.00330	0.02328	−0.04824	−0.04789
	(0.02064)	(0.02063)	0.00563)	(0.00192)	(0.00574)	(0.02788)	(0.02777)
ln *L*	−17422.72	−25285.07				−8073.604	−25285.07

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

17.9.4 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

$$y^*_{1,it} = \mathbf{x}'_{1,it}\boldsymbol{\beta}_1 + \varepsilon_{1,it} + \alpha_{1,i} \quad y_{1,it} = 1 \text{ if } y^*_{1,it} > 0, 0 \text{ otherwise},$$

$$y^*_{2,it} = \mathbf{x}'_{2,it}\boldsymbol{\beta}_2 + \varepsilon_{2,it} + \alpha_{2,i} \quad y_{2,it} = 1 \text{ if } y^*_{2,it} > 0, 0 \text{ otherwise},$$

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \Big| \mathbf{x}_1, \mathbf{x}_2 \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right].$$

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 variable 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 17.7.3.b) offers greater promise. If $(\alpha_1, \alpha_2) = (u_1, u_2)$ are normally distributed random effects, with

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

then the unconditional log likelihood for the bivariate probit model,

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

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 Section 17.7.4.

Example 17.33 Bivariate Random Effects Model for Doctor and Hospital Visits

We will extend the pooled bivariate probit model presented in Example 17.32 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 17.25 presents the pooled and random effects estimates. The log-likelihood functions for the pooled and random effects models are −25,285.07 and

TABLE 17.25 Estimated Random Effects Bivariate Probit Model

	Doctor		*Hospital*	
	Pooled	*Random Effects*	*Pooled*	*Random Effects*
Constant	−0.1243	−0.2976	−1.3385	−1.5855
	(0.0581)	(0.0965)	(0.0796)	(0.1085)
Female	0.3551	0.4548	0.1050	0.1280
	(0.0160)	(0.0286)	(0.0217)	(0.0295)
Age	0.0119	0.0199	0.0046	0.0050
	(0.0008)	(0.0013)	(0.0011)	(0.0014)
Income	−0.1337	−0.0106	0.0444	0.1336
	(0.0463)	(0.0640)	(0.0595)	(0.0773)
Kids	−0.1523	−0.1544	−0.0152	0.0216
	(0.0183)	(0.0269)	(0.0257)	(0.0321)
Education	−0.0148	−0.0257	−0.0219	−0.0244
	(0.0036)	(0.0061)	(0.0051)	(0.0068)
Married	0.0735	0.0288	−0.0479	−0.1050
	(0.0206)	(0.0317)	(0.0278)	(0.0355)
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

−23,769.67, respectively. Two times the difference is 3,030.76. 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% critical value is 7.81, so the pooling hypothesis would be rejected. The change in the correlation coefficient from 0.2981 to 0.1501 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.

17.9.5 A RECURSIVE BIVARIATE PROBIT MODEL

Section 17.6.2 examines a case in which there is an endogenous continuous variable in a binary choice (probit) model. The model is

$$T = \mathbf{x}_T'\boldsymbol{\beta}_T + \varepsilon_T,$$

$$y^* = \mathbf{x}_y'\boldsymbol{\beta}_y + \gamma T + \varepsilon_y, \quad y = \mathbf{1}(y^* > 0),$$

$$\begin{pmatrix} \varepsilon_T \\ \varepsilon_y \end{pmatrix} | \mathbf{x}_T, \mathbf{x}_y \Bigg) \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right].$$

The application examined there involved a labor force participation model that was conditioned on an endogenous variable, the non-wife part of family income. In many cases, the endogenous variable in the equation is also binary. In the application we will examine below, 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

$$\begin{aligned}
&T^* = \mathbf{x}_T'\boldsymbol{\beta}_T + \varepsilon_T, \qquad T = \mathbf{1}(T^* > 0),\\
&y^* = \mathbf{x}_y'\boldsymbol{\beta}_y + \gamma T + \varepsilon_y, \qquad y = \mathbf{1}(y^* > 0),\\
&\begin{pmatrix}\varepsilon_T\\ \varepsilon_y\end{pmatrix}\!\Big|\,\mathbf{x}_T, \mathbf{x}_y \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right].
\end{aligned}$$

This model is qualitatively different from the bivariate probit model in (17-48); the first dependent variable, T, appears on the right-hand side of the second equation.[73] 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 does not need special consideration in formulating the log likelihood.[74] We can establish this fact with the following (admittedly trivial) argument: The term that enters the log likelihood is $P(y = 1, T = 1) = P(y = 1 \mid T = 1)P(T = 1)$. Given the model as stated, the marginal probability for $T = 1$ is just $\Phi(\mathbf{x}_T'\boldsymbol{\beta}_T)$, whereas the conditional probability is $\Phi_2(\cdots)/\Phi(\mathbf{x}_T'\boldsymbol{\beta}_T)$. The product returns the bivariate normal probability we had earlier. The other three terms in the log likelihood are derived similarly, which produces:

$$\begin{aligned}
P(y = 1, T = 1) &= \Phi(\mathbf{x}_y'\boldsymbol{\beta}_y + \gamma, \mathbf{x}_T'\boldsymbol{\beta}_T, \rho),\\
P(y = 1, T = 0) &= \Phi(\mathbf{x}_y'\boldsymbol{\beta}_y, -\mathbf{x}_T'\boldsymbol{\beta}_T, -\rho),\\
P(y = 0, T = 1) &= \Phi[-(\mathbf{x}_y'\boldsymbol{\beta}_y + \gamma), \mathbf{x}_T'\boldsymbol{\beta}_T, -\rho],\\
P(y = 0, T = 0) &= \Phi(-\mathbf{x}_y'\boldsymbol{\beta}_y, -\mathbf{x}_T'\boldsymbol{\beta}_T, \rho).
\end{aligned}$$

These terms are exactly those of (17-48) that we obtain just by carrying T 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. In this instance, we are maximizing the full 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. The log likelihood for this model is

$$\ln L = \sum_{i=1}^{n} \ln \Phi[q_{y,i}(\mathbf{x}_{yi}'\boldsymbol{\beta}_y + \gamma T_i), q_{T,i}(\mathbf{x}_{T,i}'\boldsymbol{\beta}_T), q_{y,i}q_{\mathrm{T},i}\rho],$$

where $q_{y,i} = (2y_i - 1)$ and $q_{T,i} = 2(T_i - 1)$.[75]

[73] Eisenberg and Rowe (2006) is another application of this model. In their study, they analyzed the joint (recursive) effect of $T =$ veteran status on y, smoking behavior. The estimator they used was two-stage least squares and GMM. Evans and Schwab (1995), examined below, fit their model by MLE and by 2SLS for comparison.

[74] The model appears in Maddala (1983, p. 123).

[75] If one were armed with only a univariate probit estimator, it might be tempting to mimic 2SLS to estimate this model using a two-step procedure: (1) estimate $\boldsymbol{\beta}_T$ by a probit regression of T on $\mathbf{x}_T$, then (2) estimate $(\boldsymbol{\beta}_y, \gamma)$ by probit regression of y on $[\mathbf{x}_y, \Phi(\mathbf{x}_T{}'\hat{\boldsymbol{\beta}}_T)]$. This would be an example of a forbidden regression. [See Wooldridge (2010, pp. 267, 594).] The first step works, but the second does not produce consistent estimators of the parameters of interest. The estimating equation at the second is improper—the conditional probability is conditioned on T, not on the probability that T equals one. The temptation should be easy to resist; the recursive bivariate probit model is a built-in procedure in contemporary software.

Example 17.34 The Impact of Catholic School Attendance on High School Performance

Evans and Schwab (1995) considered the effect of Catholic school attendance on two success measures, graduation from high school and entrance to college. Their model is

$$C^* = \mathbf{x}'\boldsymbol{\beta}_C + \varepsilon_C, \qquad C = \mathbf{1}(C^* > 0),$$

$$G^* = \mathbf{x}'\boldsymbol{\beta}_G + \delta R + \gamma C + \varepsilon_G, \quad G = \mathbf{1}(G^* > 0),$$

$$\left(\begin{matrix}\varepsilon_C \\ \varepsilon_G\end{matrix}\middle| \mathbf{x}_C, \mathbf{x}_G\right) \sim N\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix}\right].$$

The binary variables are $C = \mathbf{1}$(Attended Catholic School) and $G = \mathbf{1}$(Graduated from high school). In a second specification of the model, $G = \mathbf{1}$(Entered a four-year college after graduation). Covariates included race, gender, family income, parents' education, family structure, religiosity, and a tenth-grade test score. The parameters of the model are all identified (estimable) whether or not there are variables in the G equation that are not in the C equation (i.e., whether or not there are exclusion restrictions) by dint of the nonlinearity of the structure. However, mindful of the dubiousness of a model that is identified *only* by the nonlinearity, the authors included $\text{R} = \mathbf{1}$ (Student is Catholic) in the equation, to aid identification. That would seem important here, as of more than 30 variables in the equations, only two, the test score and a "% Catholic in County of Residence," were not also dummy variables. (Income was categorized.)

Example 17.35 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:

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

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 programs, and 29 have both. (See Appendix Table F17.1.) The estimated parameters are given in Table 17.26. Both bivariate probit and 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 17.26 Estimates of a Recursive Simultaneous Bivariate Probit Model (estimated standard errors in parentheses)

	Single Equation		*Bivariate Probit*	
Variable	*Coefficient*	*Std. Err.*	*Coefficient*	*Std. Err.*
Gender Economics Equation				
Constant	−1.4176	(0.8768)	−1.1911	(2.2155)
AcRep	−0.0114	(0.0036)	−0.0123	(0.0079)
WomStud	1.1095	(0.4699)	0.8835	(2.2603)
EconFac	0.0673	(0.0569)	0.0677	(0.0695)
PctWEcon	2.5391	(0.8997)	2.5636	(1.0144)
Relig	−0.3482	(0.4212)	−0.3741	(0.5264)
Women's Studies Equation				
AcRep	−0.0196	(0.0042)	−0.0194	(0.0057)
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	

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 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 partial 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 G. The total effect of academic reputation and, likewise, religious affiliation is the sum of these two parts. Consider first the gender economics variable, G. The conditional probability is

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

TABLE 17.27 Partial Effects in Gender Economics Model

	Direct	*Indirect*	*Total*	*(Type of Variable, Mean)*
AcRep	−0.0017	−0.0005	−0.0022	(Continuous, 119.242)
PctWEcon	0.3602		0.3602	(Continuous, 0.24787)
EconFac	0.0095		0.0095	(Continuous, 6.74242)
Relig			−0.0716[a]	(Binary, 0.57576)
PctWFac		0.0508	0.0508	(Continuous, 0.35772)

[a] Direct and indirect effects for binary variables are the same.

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 probability function may not produce an accurate result. Instead, we would compute the probability 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 first compute the average treatment effect (see Section 17.6.1) by averaging

$$\text{TE} = \Phi(\mathbf{x}_G{}'\boldsymbol{\beta}_G + \gamma) - \Phi(\mathbf{x}_G{}'\boldsymbol{\beta}_G)$$

over the full sample of schools. The average treatment effect for the schools that actually do have a women's studies program would be

$$\text{TET} = \Phi\left[\frac{(\mathbf{x}_G'\boldsymbol{\beta}_G + \gamma) - \rho(\mathbf{x}_W'\boldsymbol{\beta}_W)}{\sqrt{1 - \rho^2}}\right] - \Phi\left[\frac{(\mathbf{x}_G'\boldsymbol{\beta}_G) - \rho(\mathbf{x}_W'\boldsymbol{\beta}_W)}{\sqrt{1 - \rho^2}}\right]$$

and averaging over the schools that have a women's studies program ($W = 1$).

Table 17.27 presents the estimates of the partial effects and some descriptive statistics for the data. Numerically, the strongest effect appears to be exerted by the representation of women on the faculty; its coefficient of 0.3602 is by far the largest. However, this variable cannot change by a full unit because it is a proportion. An increase of 1% in the presence of women on the economics faculty raises the probability by only 0.0036, which is comparable in scale to the effect of academic reputation. The effect of women on the faculty is likewise fairly small, only 0.000508 per 1% change. As might have been expected, the single most important influence is the presence of a women's studies program. The estimated average treatment effect is 0.1452 (0.3891). The average treatment effect on the schools that have women's studies programs (ATET) is 0.2293 (0.5165). 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.

17.10 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

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

The joint probabilities of the observed events, $[y_{i1}, y_{i2} \ldots, y_{iM}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iM}]$, $i = 1, \ldots, n$ 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

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

The practical obstacle to this extension is the evaluation of the M-variate normal integrals and their derivatives. Simulation-based integration using the GHK simulator or simulated likelihood methods (see Chapter 15) allow for estimation of relatively large models. We consider an application in Example 17.36.[76]

The **multivariate probit model** in another form presents a useful extension of the random effects probit model for panel data (Section 17.7.2). If the parameter vectors in all equations are constrained to be equal, we obtain what Bertschek and Lechner (1998) call the "panel probit model",

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

The Butler and Moffitt (1982) approach for this model (see Section 17.4.2) has proved useful in many applications. But the underlying assumption that $\text{Cov}[\varepsilon_{it}, \varepsilon_{is}] = \rho$ 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.[77] Hyslop (1999), Bertschek and Lechner (1998), Greene (2004 and Example 17.26), and Cappellari and Jenkins (2006) are applications.

Example 17.36 A Multivariate Probit Model for Product Innovations

Bertschek and Lechner applied the panel probit model to an analysis of the innovation activity of 1,270 German firms observed in five years, 1984–1988, in response to imports and foreign direct investment.[78] The probit model to be estimated is based on the latent regression

[76] 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 17.26), which applies the technique to a panel data application with $T = 7$. Example 17.23 develops a five-variate application.

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

[78] See Bertschek (1995).

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

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 17.28 presents the single-equation, pooled probit model estimates.[79] 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 will produce a consistent estimator of the parameter vector, this is a case in which the cluster estimator (see Section 14.8.2) provides an appropriate asymptotic covariance matrix. Note

TABLE 17.28 Estimated Pooled Probit Model

		Estimated Standard Errors				*Partial 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	—	—	—
ln 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.

[79] 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 17.29 Estimated Constrained Multivariate Probit Model (Estimated standard errors in parentheses)

Coefficients	*Full Maximum Likelihood Using GHK Simulator*	*Random Effects* ρ **= 0.578 (0.0189)**
Constant	−1.797** (0.341)	−2.839 (0.534)
ln Sales	0.154** (0.0334)	0.245 (0.052)
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 (0.095)
log likelihood	−3,522.85	−3,535.55
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% level,

**Indicates significant at 99% level based on a two-tailed test.

that the standard errors in column SE(4) of the table are considerably higher than the uncorrected ones in columns 1 and 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. The FIML estimates and the random effects model using the Butler and Moffitt (1982) quadrature method are reported in Table 17.29. The correlations reported are based on the FIML estimates. Also noteworthy in Table 17.30 is the divergence of the random effects estimates from the FIML estimates. The log-likelihood function is −3,535.55 for the random effects model and −3,522.85 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% and 21.7 for 99% significance, so the hypothesis of the random effects model would be rejected in favor of the more general panel probit model.

17.11 SUMMARY AND CONCLUSIONS

This chapter has surveyed a large range of techniques for modeling binary choice variables. The model for choice between two alternatives provides the framework for a 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 models with 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
- Average partial effect
- Binary choice model
- Bivariate probit
- Butler and Moffitt method
- Characteristics
- Choice-based sampling
- Complementary log log model
- Conditional likelihood function
- Control function
- Event count
- Fixed effects model
- Generalized residual
- Gumbel model
- Incidental parameters problem
- Index function model
- Initial conditions
- Interaction effect
- Inverse probability weighted (IPW)
- Latent regression
- Linear probability model (LPM)
- Logit
- Marginal effects
- Maximum simulated likelihood (MSL)
- Method of scoring
- Microeconometrics
- Minimal sufficient statistic
- Multinomial choice
- Multivariate probit model
- Nonresponse bias
- Ordered choice model
- Persistence
- Quadrature
- Qualitative response (QR)
- Random effects model
- Recursive model
- 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:

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

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 xs, where p is the sample proportion of 1s.

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% 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 (for example, 0.11 for observation 1) and $(1 - y_i)$ with (1 − the rate), in (17-16).

b. If trucks cost \$20,000 each, then is a goal of 90% 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 (17-26).
8. In the panel data models estimated in Section 17.7, 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.)

Application

1. Appendix Table F17.2 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 all of them necessarily belong in your model. Use these data to build a binary choice model for A. Report all computed results for the model. Compute the partial 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?