Chs 17, 18 In 10/15

M. E. Williams, Inc.

Enclosure Memo

October 13, 2010

TO: Martha Wetherill

FROM: Martha Williams

RE: Greene 2140242

Dear Martha,

Please find herewith:

Copyedited manuscript for Chapters 17 and 18

Sincerely,

Martha

17
DISCRETE CHOICE

17.1 INTRODUCTION

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

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

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

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

Multinomial Choice: The individual chooses among more than two choices, once again, making the choice that provides the greatest utility. In the previous example, private travel might involve a choice of being a driver or passenger while public transport might involve a choice between bus and train. At one level, this is a minor variation of the binary choice case – the latter is, of course, a special case of the former. But, more elaborate models of multinomial choice allow a rich specification of consumer preferences. In the multinomial case, the observed response is simply a label for the selected choice; it might be a brand, the name of a place, or the type of travel mode. Numerical assignments are not meaningful in this setting.

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

Ordered Choice: The individual reveals the strength of their 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,...,*J* for some upper limit, *J*. For example, opinions might be labelled 0,1,2,3,4 to indicate the strength of preferences, for example, for a product, a movie, a candidate or a piece of legislation. But, in this context, the numerical values are only a ranking, not a quantitative measure. Thus a "1" is greater than a "0" in a qualitative sense, but not by one unit, and the difference between a "2" and a "1" is not the same as that between a "1" and a "0."

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

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

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

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

model for counts of events. We will then extend the model to demonstrate some recent applications and innovations.

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

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

17.2 Models for Binary Outcomes

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

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

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

Example 17.1 Labor Force Participation Model

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

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

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

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

17.2.1 RANDOM UTILITY MODELS FOR INDIVIDUAL CHOICE

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

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

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

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

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

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

Example 17.2 Structural Equations for a Binary Choice Model

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

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

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

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

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

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

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

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

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

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

17.2.2 A LATENT REGRESSION MODEL

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

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

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

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

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

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

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

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

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

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

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

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

17.2.3 FUNCTIONAL FORM AND REGRESSION

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

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

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

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

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

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

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

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

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

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

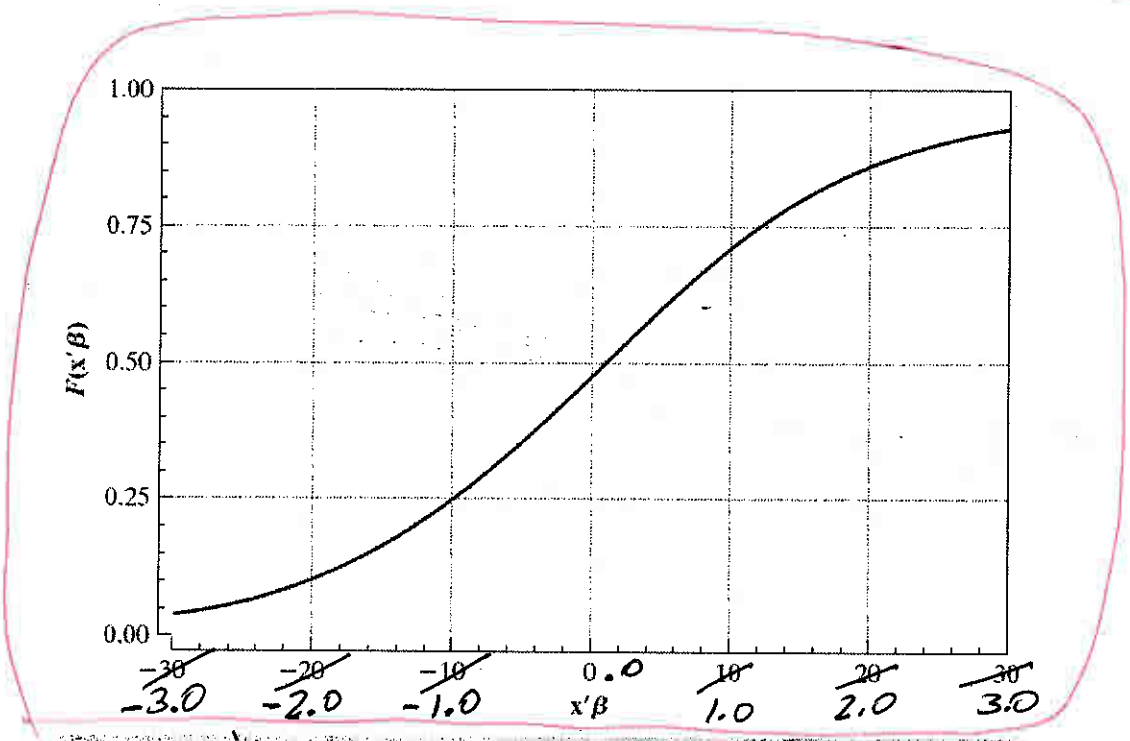

FIGURE 23.1 Model for a Probability.

17

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

$$\lim_{\mathbf{x}'\boldsymbol{\beta} \to +\infty} \text{Prob}(Y = 1 \mid \mathbf{x}) = 1$$

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

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

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

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

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

has also been used in many applications. We shall use the notation $\Lambda(.)$ to indicate the logistic cumulative distribution function. This model is called the **logit** model for reasons we shall discuss in the next section. Both of these distributions have the familiar bell shape of symmetric

distributions. Other models which do not assume symmetry, such as the **Gumbel model,**

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

and **complementary log log model,**

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

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

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

The probability model is a regression:

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

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

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

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

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

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

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

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

so, in the logit model,

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

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

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

Another complication for computing marginal effects in a binary choice model arises because $\mathbf{x}$ will often include dummy variables—for example, a labor force participation equation will often contain a dummy variable for marital status. Because the derivative is with respect to a small change, it is not appropriate to apply (17-11) for the effect of a change in a dummy variable, or a change of state. The appropriate marginal effect for a binary independent variable, say, d, would be

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

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

17.3 ESTIMATION AND INFERENCE IN BINARY CHOICE MODELS

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

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

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

$$L(\beta \mid \text{data}) = \prod_{i=1}^{n} [F(\mathbf{x}_i'\beta)]^{y_i} [1 - F(\mathbf{x}_i'\beta)]^{1-y_i}. \tag{23-16}$$

Taking logs, we obtain

$$\ln L = \sum_{i=1}^{n} \{y_i \ln F(\mathbf{x}_i'\beta) + (1 - y_i) \ln[1 - F(\mathbf{x}_i'\beta)]\}. \tag{23-17}$$

The **likelihood equations** are

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \left[\frac{y_i f_i}{F_i} + (1 - y_i)\frac{-f_i}{(1 - F_i)}\right]\mathbf{x}_i = \mathbf{0}, \tag{23-18}$$

where f_i is the density, $dF_i/d(\mathbf{x}_i'\beta)$. [In (23-18) and later, we will use the subscript i to indicate that the function has an argument $\mathbf{x}_i'\beta$.] 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 (23-18) will be nonlinear and require an iterative solution. All of the models we have seen thus far are relatively straightforward to analyze. For the logit model, by inserting (23-7) and (23-11) in (23-18), we get, after a bit of manipulation, the likelihood equations

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} (y_i - \Lambda_i)\mathbf{x}_i = \mathbf{0}. \tag{23-19}$$

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.[7] This implication also bears some similarity to the least squares normal equations if we view the term $y_i - \Lambda_i$ as a residual.[8] For the normal distribution, the log-likelihood is

$$\ln L = \sum_{y_i=0} \ln[1 - \Phi(\mathbf{x}_i'\beta)] + \sum_{y_i=1} \ln \Phi(\mathbf{x}_i'\beta). \tag{23-20}$$

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

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

Using the device suggested in footnote 6, we can reduce this to

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} \left[\frac{q_i \phi(q_i \mathbf{x}_i'\beta)}{\Phi(q_i \mathbf{x}_i'\beta)}\right]\mathbf{x}_i = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i = \mathbf{0}, \tag{23-21}$$

where $q_i = 2y_i - 1$.

[7] If the distribution is symmetric, as the normal and logistic are, then $1 - F(\mathbf{x}'\beta) = F(-\mathbf{x}'\beta)$. There is a further simplification. Let $q = 2y - 1$. Then $\ln L = \Sigma_i \ln F(q_i \mathbf{x}_i'\beta)$. See (23-21).

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

[9] This sort of construction arises in many models. The first derivative of the log-likelihood with respect to the constant term produces the **generalized residual** in many settings. See, for example, Chesher, Lancaster, and Irish (1985) and the equivalent result for the tobit model in Section 24.3.4.d.

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

$$\mathbf{H} = \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_i \Lambda_i (1 - \Lambda_i) \mathbf{x}_i \mathbf{x}_i'. \tag{23-22}$$

The second derivatives do not involve the random variable y_i, so Newton's method is also the **method of scoring** for the logit model. Note that the Hessian is always negative definite, so the log-likelihood is globally concave. Newton's method will usually converge to the maximum of the log-likelihood in just a few iterations unless the data are especially badly conditioned. The computation is slightly more involved for the probit model. A useful simplification is obtained by using the variable $\lambda(y_i, \boldsymbol{\beta}'\mathbf{x}_i) = \lambda_i$ that is defined in (23-21). The second derivatives can be obtained using the result that for any z, $d\phi(z)/dz = -z\phi(z)$. Then, for the probit model,

$$\mathbf{H} = \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} -\lambda_i (\lambda_i + \mathbf{x}_i'\boldsymbol{\beta}) \mathbf{x}_i \mathbf{x}_i'. \tag{23-23}$$

This matrix is also negative definite for all values of $\boldsymbol{\beta}$. The proof is less obvious than for the logit model.[9] It suffices to note that the scalar part in the summation is $\text{Var}[\varepsilon \mid \varepsilon \leq \boldsymbol{\beta}'\mathbf{x}] - 1$ when $y = 1$ and $\text{Var}[\varepsilon \mid \varepsilon \geq -\boldsymbol{\beta}'\mathbf{x}] - 1$ when $y = 0$. The unconditional variance is one. Because truncation always reduces variance—see Theorem 24.2—in both cases, the variance is between zero and one, so the value is negative.[10]

The asymptotic covariance matrix for the maximum likelihood estimator can be estimated by using the inverse of the Hessian evaluated at the maximum likelihood estimates. There are also two other estimators available. The Berndt, Hall, Hall, and Hausman estimator [see (16-18) and Example 16.4] would be

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

where $g_i = (y_i - \Lambda_i)$ for the logit model [see (23-19)] and $g_i = \lambda_i$ for the probit model [see (23-21)]. 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 (23-21)], this result is not true for the probit model. Amemiya (1981) showed that for the probit model,

$$E\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'}\right]_{\text{probit}} = \sum_{i=1}^{n} \lambda_{0i}\lambda_{1i} \mathbf{x}_i \mathbf{x}_i'. \tag{23-24}$$

Once again, the scalar part of the expression is always negative [see (23-21) and note that λ_{0i} is always negative and λ_{i1} is always positive]. The estimator of the asymptotic covariance matrix for the maximum likelihood estimator is then the negative inverse of whatever matrix is used to estimate the expected Hessian. Since the actual Hessian is generally used for the iterations, this option is the usual choice. As we shall see later, though, for certain hypothesis tests, the BHHH estimator is a more convenient choice.

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

[10] See Johnson and Kotz (1993) and Heckman (1979). We will make repeated use of this result in Chapter 24.

17.3.1

23.4.1 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 16.8 for discussion),

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

has been used in a number of recent studies based on the probit model [e.g., Fernandez and Rodriguez-Poo (1997), Horowitz (1993), and Blundell, Laisney, and Lechner (1993)]. If the probit model is correctly specified, then $\text{plim}(1/n)\hat{\mathbf{B}} = \text{plim}(1/n)(-\hat{\mathbf{H}})$ and either single matrix will suffice, so the robustness issue is moot (of course). On the other hand, the probit (Q-) maximum likelihood estimator is *not* consistent in the presence of any form of heteroscedasticity, unmeasured heterogeneity, omitted variables (even if they are orthogonal to the included ones), nonlinearity of the functional form of the index, or an error in the distributional assumption [with some narrow exceptions as described by Ruud (1986)]. Thus, in almost any case, the sandwich estimator provides an appropriate asymptotic covariance matrix for an estimator that is biased in an unknown direction. [See Section 16.8 and Freedman (2006).] White raises this issue explicitly, although it seems to receive little attention in the literature: "It is the consistency of the QMLE for the parameters of interest in a wide range of situations which insures its usefulness as the basis for robust estimation techniques" (1982a, p. 4). His very useful result is that if the quasi-maximum likelihood estimator converges to a probability limit, then the sandwich estimator can, under certain circumstances, be used to estimate the asymptotic covariance matrix of that estimator. But there is no guarantee that the QMLE *will* converge to anything interesting or useful. Simply computing a robust covariance matrix for an otherwise inconsistent estimator does not give it redemption. Consequently, the virtue of a robust covariance matrix in this setting is unclear.

17.3.2

23.4.2 MARGINAL EFFECTS AND AVERAGE PARTIAL EFFECTS

The predicted probabilities, $F(\mathbf{x}'\hat{\boldsymbol{\beta}}) = \hat{F}$ and the estimated ~~marginal~~ partial effects $f(\mathbf{x}'\hat{\boldsymbol{\beta}})\times\hat{\boldsymbol{\beta}} = \hat{f}\hat{\boldsymbol{\beta}}$ are nonlinear functions of the parameter estimates. To compute standard errors, we can use the linear approximation approach (delta method) discussed in Section 4.9.4. For the predicted probabilities,

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

where

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

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

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

Combining terms gives

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

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

$$\Delta \hat{F} = \hat{F}[(d = 1) - \hat{F}[(d = 0). \tag{23-25}$$

17-24

The asymptotic variance would be

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

where (23-26)

17-25

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

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

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

The matrix of derivatives is

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

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

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

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

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

Collecting terms, we obtain

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

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

17 14.1

Example 23.3 Probability Models

The data listed in Appendix Table F16.1 were taken from a study by Spector and Mazzeo (1980), which examined whether a new method of teaching economics, the Personalized System of Instruction (*PSI*), significantly influenced performance in later economics courses. The "dependent variable" used in our application is *GRADE*, which indicates the whether a student's grade in an intermediate macroeconomics course was higher than that in the principles course. The other variables are *GPA*, their grade point average; *TUCE*, the score on a pretest that indicates entering knowledge of the material; and *PSI*, the binary variable indicator of whether the student was exposed to the new teaching method. (Spector and Mazzeo's specific equation was somewhat different from the one estimated here.)

TB 17.1

17

Inference

complementary log log

Table 23.1 presents four sets of parameter estimates. The slope parameters and derivatives were computed for four probability models: linear, probit, logit, and Gumbel. The last three sets of estimates are computed by maximizing the appropriate log-likelihood function. Estimation is discussed in the next section, so standard errors are not presented here. The scale factor given in the last row is the density function evaluated at the means of the variables. Also, note that the slope given for *PSI* is the derivative, not the change in the function with *PSI* changed from zero to one with other variables held constant.

TABLE 23.1 Estimated Probability Models

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

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

The data used in this example are only moderately unbalanced between 0s and 1s for the dependent variable (21 and 11). As such, we might expect similar results for the probit and logit models.[21] 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 (23-9) may help to explain the finding. The index $\mathbf{x}'\boldsymbol{\beta}$ is not the random variable. (See Section 23.3.2.) The marginal effect in the probit model for, say, x_k is $\phi(\mathbf{x}'\boldsymbol{\beta}_p)\beta_{pk}$, whereas that for the logit is $\Lambda(1-\Lambda)\beta_{lk}$. (The subscripts p and l are for probit and logit.) Amemiya suggests that his approximation works best at the center of the distribution, where $F = 0.5$, or $\mathbf{x}'\boldsymbol{\beta} = 0$ for either distribution. Suppose it is. Then $\phi(0) = 0.3989$ and $\Lambda(0)[1-\Lambda(0)] = 0.25$. If the marginal effects are to be the same, then $0.3989\,\beta_{pk} = 0.25\beta_{lk}$, or $\beta_{lk} = 1.6\beta_{pk}$, which is the regularity observed by Amemiya. Note, though, that as we depart from the center of the distribution, the relationship will move away from 1.6. Because the logistic density descends more slowly than the normal, for unbalanced samples such as ours, the ratio of the logit coefficients to the probit coefficients will tend to be larger than 1.6. The ratios for the ones in Table 23.1 are closer to 1.7 than 1.6.

The computation of the derivatives of the conditional mean function is useful when the variable in question is continuous and often produces a reasonable approximation for a dummy variable. Another way to analyze the effect of a dummy variable on the whole distribution is to compute Prob$(Y = 1)$ over the range of $\mathbf{x}'\boldsymbol{\beta}$ (using the sample estimates) and with the two values of the binary variable. Using the coefficients from the probit model in Table 23.1, we have the following probabilities as a function of *GPA*, at the mean of TUCE:

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

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

Figure 23.2 shows these two functions plotted over the range of *GRADE* observed in the sample, 2.0 to 4.0. The marginal effect of *PSI* is the difference between the two functions, which ranges from only about 0.06 at $GPA = 2$ to about 0.50 at *GPA* of 3.5. This effect shows that the probability that a student's grade will increase after exposure to *PSI* is far greater for students with high *GPA*s than for those with low *GPA*s. At the sample mean of *GPA* of 3.117, the effect of *PSI* on the probability is 0.465. The simple derivative calculation of (23-9) is given in Table 23.1; the estimate is 0.468. But, of course, this calculation does not show the wide range of differences displayed in Figure 23.2.

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

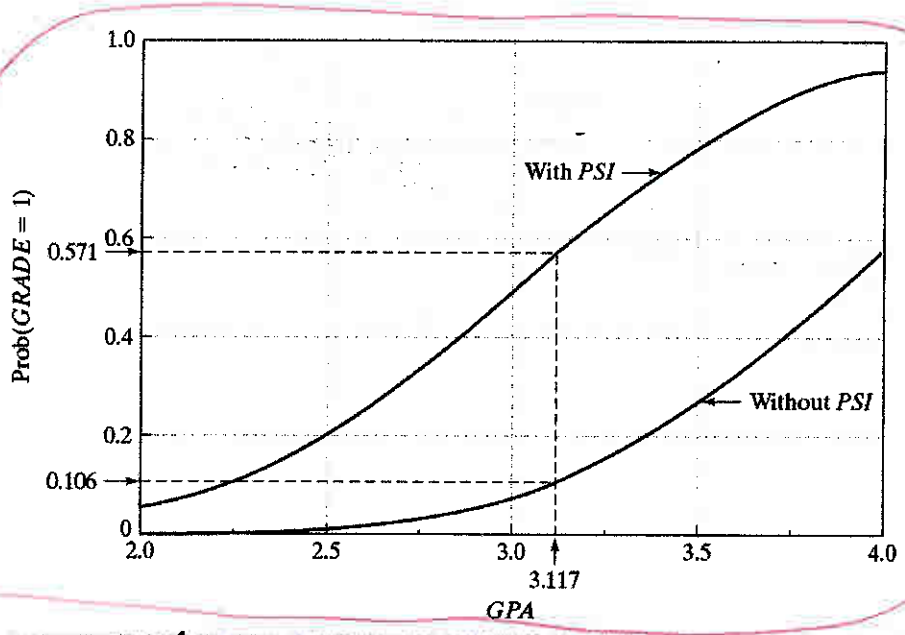

FIGURE 23.2 Effect of *PSI* on Predicted Probabilities.

Table 23.2 presents the estimated coefficients and marginal effects for the probit and logit models in Table 23.2. In both cases, the asymptotic covariance matrix is computed from the negative inverse of the actual Hessian of the log-likelihood. The standard errors for the estimated marginal effect of *PSI* are computed using (23-25) and (23-26) since *PSI* is a binary variable. In comparison, the simple derivatives produce estimates and standard errors of (0.449, 0.181) for the logit model and (0.464, 0.188) for the probit model. These differ only slightly from the results given in the table.

7.3.2.a Average Partial Effects

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

TABLE 23.2 Estimated Coefficients and Standard Errors (standard errors in parentheses)

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

of interest is

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

In practical terms, this suggests the computation

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

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

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

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

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

where Δ is the remaining higher-order terms. The first of the three terms is the marginal effect computed at the sample means. The second term is zero by construction. That leaves the remainder plus an average of a term that is a function of the variances and covariances of the data and the curvature of the probability function at the means. Little can be said to characterize these two terms in any particular sample, but one might guess they are likely to be small. We will examine an application in Example 23.4. 17.4

Computing the individual effects, then using the natural estimator to estimate the variance of the mean,

$$\text{Est. Var}[\bar{\hat{\gamma}}_k] = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{N}\left(\hat{\gamma}_k(\mathbf{x}_i) - \bar{\hat{\gamma}}_k\right)^2\right],$$

may badly estimate the asymptotic variance of the average partial effect. [See, e.g. Contoyannis et al. (2004, p. 498).] The reason is that the observations in the APE are highly correlated—they all use the same estimate of $\boldsymbol{\beta}$—but this variance computation treats them as a random sample. The following example shows the difference, which is substantial. To use the delta method to estimate asymptotic standard errors for the average partial effects, we would use

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

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

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

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

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

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

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

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

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

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

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

where

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

This treats the APE as a point estimator of a population parameter—one that converges in probability to what we assume is its population counterpart. But, it is conditioned on the sample data; convergence is with respect to $\hat{\boldsymbol{\beta}}$. This looks like a formidable amount of computation—Example 17.4 uses a sample of 27,326 observations, so it appears we need a double sum of roughly 750 million terms. However, the computation is actually linear in n, not quadratic, because the same matrix is used in the center of each product. The estimator of the asymptotic covariance matrix for the *APE* is simply

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

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

Example 17.4 Average Partial Effects

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

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

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

TABLE 17.3 Estimated Parameters and Partial Effects

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

17.3.2.b Interaction Effects

Models with **interaction effects**, such as

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

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

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

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

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

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

$$\partial\text{Prob}(DocVis > 0|\mathbf{x})/\partial Age = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})](\beta_2 + \beta_7\, Education),$$
$$\partial\text{Prob}(DocVis > 0|\mathbf{x})/\partial Education = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})](\beta_5 + \beta_7\, Age),$$

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

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

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

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

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

Not only is β_7 not the interesting effect, but there is a complicated additional term. Loosely, we can associate the first term as a "direct" effect – note that it is the naive term PE_7 from earlier. The second part can be attributed to the fact that we are differentiating a nonlinear model – essentially, the second part of the partial effect results from the nonlinearity of the function. The existence of an "interaction effect" in this model is inescapable – notice that the second part is

[13] See, e.g., Ai and Norton (2004) and Greene (2010).

nonzero (generally) even if β_7 does equal zero. Whether this is intended to represent an 'interaction' in some economic sense is unclear. In the absence of the product term in the model, probably not. We can see an implication of this in Figure 17.1. At the point where $\mathbf{x}'\boldsymbol{\beta} = 0$, where the probability equals one half, the probability function is linear. At that point, $(1-2\Lambda)$ will equal zero and the functional form effect will be zero as well. When $\mathbf{x}'\boldsymbol{\beta}$ departs from zero, the probability becomes nonlinear. (These same effects can be shown for the probit model at $\mathbf{x}'\boldsymbol{\beta} = 0$, the second derivative of the probit probability is $-\mathbf{x}'\boldsymbol{\beta}\phi(\mathbf{x}'\boldsymbol{\beta}) = 0$.)

We developed an extensive application of interaction effects in a nonlinear model in Example 7.6. In that application, using the same data for the numerical exercise, we analyzed a nonlinear regression $E[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$. 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 17.5 Interaction Effect

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

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

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

17-24

TABLE 23.4 Estimated Coefficients

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

*Marginal effect and estimated standard error include both mean (β) and variance (γ) effects.

Table 23.4 presents estimates of the probit model with a correction for heteroscedasticity of the form

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

The three tests for homoscedasticity give

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

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

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

The 99 percent critical value for two restrictions is 5.99, so the LM statistic conflicts with the other two.

17.3.3 ~~20.4.5~~ MEASURING GOODNESS OF FIT

There have been many fit measures suggested for QR models.[14] At a minimum, one should report the maximized value of the log-likelihood function, ln L. Because the hypothesis that all the slopes in the model are zero is often interesting, the log-likelihood computed with only a constant term, ln L_0 [see (~~23.28~~)], should also be reported. An analog to the R^2 in a conventional regression is McFadden's (1974) likelihood ratio index,

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

This measure has an intuitive appeal in that it is bounded by zero and one. (See Section ~~16~~.6.5.) If all the slope coefficients are zero, then it equals zero. There is no way to make LRI equal 1, although one can come close. If F_i is always one when y equals one and zero when y equals zero, then ln L equals zero (the log of one) and LRI equals one. It has been suggested that this finding is indicative of a "perfect fit" and that LRI increases as the fit of the model improves. To a degree, this point is true (~~see the analysis in Section 23.8.4~~). Unfortunately, the values between zero and one have no natural interpretation. If $F(\mathbf{x}_i'\boldsymbol{\beta})$ is a proper pdf, then even with many regressors the model cannot fit perfectly unless $\mathbf{x}_i'\boldsymbol{\beta}$ goes to $+\infty$ or $-\infty$. As a practical matter, it does happen. But when it does, it

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

indicates a flaw in the model, not a good fit. If the range of one of the independent variables contains a value, say, x^*, such that the sign of $(x - x^*)$ predicts y perfectly and vice versa, then the model will become a perfect predictor. This result also holds in general if the sign of $\mathbf{x}'\boldsymbol{\beta}$ gives a perfect predictor for some vector $\boldsymbol{\beta}$.[15] For example, one might mistakenly include as a regressor a dummy variable that is identical, or nearly so, to the dependent variable. In this case, the maximization procedure will break down precisely because $\mathbf{x}'\boldsymbol{\beta}$ is diverging during the iterations. [See McKenzie (1998) for an application and discussion.] Of course, this situation is not at all what we had in mind for a good fit.

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

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

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

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

Cramer's measure heavily penalizes the incorrect predictions, and because each proportion is taken within the subsample, it is not unduly influenced by the large proportionate size of the group of more frequent outcomes. Some of the other proposed fit measures are Efron's (1978)

$$R^2_{\text{Ef}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{p}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

Veall and Zimmermann's (1992)

$$R^2_{\text{VZ}} = \left(\frac{\delta - 1}{\delta - \text{LRI}}\right)\text{LRI}, \quad \delta = \frac{n}{2\log L_0},$$

and Zavoina and McKelvey's (1975)

$$R^2_{\text{MZ}} = \frac{\sum_{i=1}^{n}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}} - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}\right)^2}{n + \sum_{i=1}^{n}\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}} - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}\right)^2}.$$

The last of these measures corresponds to the regression variation divided by the total variation in the latent index function model, where the disturbance variance is $\sigma^2 = 1$. The values of several of these statistics are given with the model results in Table 23.15 with the application in Section 23.8.4 for illustration.

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

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

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

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

$$\hat{y} = 1 \text{ if } P > 0.5 \text{ and } 0 \text{ otherwise,} \qquad (23\text{-}37)$$

where P is the simple proportion of ones in the sample. This rule will always predict correctly $100P$ percent of the observations, which means that the naive model does not have zero fit. In fact, if the proportion of ones in the sample is very high, it is possible to construct examples in which the second model will generate more correct predictions than the first! Once again, this flaw is not in the model; it is a flaw in the fit measure.[16] The important element to bear in mind is that the coefficients of the estimated model are not chosen so as to maximize this (or any other) fit measure, as they are in the linear regression model where $\mathbf{b}$ maximizes R^2. ~~(The maximum score estimator discussed in Example 23.12 addresses this issue directly.)~~

Another consideration is that 0.5, although the usual choice, may not be a very good value to use for the threshold. If the sample is **unbalanced**—that is, has many more ones than zeros, or vice versa—then by this prediction rule it might never predict a one (or zero). To consider an example, suppose that in a sample of 10,000 observations, only 1,000 have $Y = 1$. We know that the average predicted probability in the sample will be 0.10. As such, it may require an extreme configuration of regressors even to produce an F of 0.2, to say nothing of 0.5. In such a setting, the prediction rule may fail every time to predict when $Y = 1$. The obvious adjustment is to reduce F^*. Of course, this adjustment comes at a cost. If we reduce the threshold F^* so as to predict $y = 1$ more often, then we will increase the number of correct classifications of observations that do have $y = 1$, but we will also increase the number of times that we *incorrectly* classify as ones observations that have $y = 0$.[17] In general, any prediction rule of the form in (~~23-36~~) will make two types of errors: It will incorrectly classify zeros as ones and ones as zeros. In practice, these errors need not be symmetric in the costs that result. For example, in a credit scoring model [see Boyes, Hoffman, and Low (1989)], incorrectly classifying an applicant as a bad risk is not the same as incorrectly classifying a bad risk as a good one. Changing F^* will always reduce the probability of one type of error while increasing the probability of the other. There is no correct answer as to the best value to choose. It depends on the setting and on the criterion function upon which the prediction rule depends.

The likelihood ratio index and ~~Veall and Zimmermann's~~ various modification of it are obviously related to the likelihood ratio statistic for testing the hypothesis that the coefficient vector is zero. ~~Efron's and~~ Cramer's measures ~~listed previously are~~ is oriented more toward the relationship between the fitted probabilities and the actual values. ~~Efron's and Cramer's statistics are~~ It is usefully tied to the standard prediction rule $\hat{y} = \mathbf{1}[\hat{F} > 0.5]$. ~~The McKelvey and Zavoina measure is an analog to the regression coefficient of determination, based on the underlying regression $y^* = \mathbf{x}'\beta + \varepsilon$.~~ Whether these ~~have~~ it has a close relationship to any type of fit in the familiar sense is a question that needs to be

[16] See Amemiya (1981).

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

studied. In some cases, it appears so. But the maximum likelihood estimator, on which all the fit measures are based, is not chosen so as to maximize a fitting criterion based on prediction of y as it is in the classical regression (which maximizes R^2). It is chosen to maximize the joint density of the observed dependent variables. It remains an interesting question for research whether fitting y well or obtaining good parameter estimates is a preferable estimation criterion. Evidently, they need not be the same thing.

Example 23.6 Prediction with a Probit Model

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

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

23.4.6 CHOICE-BASED SAMPLING

In some studies [e.g., Boyes, Hoffman, and Low (1989), Greene (1992)], the mix of ones and zeros in the observed sample of the dependent variable is deliberately skewed in favor of one outcome or the other to achieve a more balanced sample than random sampling would produce. The sampling is said to be **choice based**. In the studies noted, the dependent variable measured the occurrence of loan default, which is a relatively uncommon occurrence. To enrich the sample, observations with $y = 1$ (default) were oversampled. Intuition should suggest (correctly) that the bias in the sample should be transmitted to the parameter estimates, which will be estimated so as to mimic the sample, not the population, which is known to be different. Manski and Lerman (1977) derived the weighted endogenous sampling maximum likelihood (WESML) estimator for this situation. The estimator requires that the true population proportions, ω_1 and ω_0, be known. Let p_1 and p_0 be the sample proportions of ones and zeros. Then the estimator is obtained by maximizing a weighted log-likelihood,

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

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

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

where $\mathbf{G}_i(\hat{\boldsymbol{\beta}}) = \partial F(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$ and $\hat{\mathbf{V}}$ is the estimated asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$. (The terms with equal subscripts are the same computation we did earlier with the sample means.) This looks like a formidable amount of computation—Example 23.4 uses a sample of 27,326 observations, so at first blush, it appears we need a double sum of roughly 750 million terms. The computation is actually linear in n, not quadratic, however, because the same matrix is used in the center of each product. Moving the first derivative matrix outside the inner summation and using the $1/n$ twice, we find that the estimator of the asymptotic covariance matrix for the *APE* is simply

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

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

Example 23.4 Average Partial Effects

We estimated a binary logit model for $y = 1(DocVis > 0)$ using the German health care utilization data examined in Example 11.10 (and in Examples 11.11, 16.10, 16.12, 16.13, 16.16, and 18.6). The model is

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

No account of the panel nature of the data set was taken for this exercise. The sample contains 27,326 observations, which should be large enough to reveal the large sample behavior of the computations. Table 23.3 presents the parameter estimates for the logit probability model and both the marginal effects and the average partial effects, each with standard errors computed using the results given earlier. The results do suggest the similarity of the computations. The values in parentheses are based on the naïve estimator that ignores the covariances.

17.3.4 ~~23.4.3~~ HYPOTHESIS TESTS

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

TABLE 23.3 Estimated Parameters and Partial Effects

	Parameter Estimates		*Marginal Effects*		*Average Partial Effects*	
Variable	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*
Constant	0.25111	0.091135				
Age	0.020709	0.0012852	0.0048133	0.00029819	0.0047109	0.00028727 (0.00042971)
Income	−0.18592	0.075064	−0.043213	0.017446	−0.042294	0.017069 (0.0038579)
Kids	−0.22947	0.029537	−0.053335	0.0068626	−0.052201	0.0066921 (0.0047615)
Education	−0.045588	0.0056465	−0.010596	0.0013122	−0.010370	0.0012787 (0.00094595)
Married	0.085293	0.033286	0.019824	0.0077362	0.019403	0.0075686 (0.0017698)

restrictions $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, the statistic is

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

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

$$W = \hat{\boldsymbol{\beta}}_M'\mathbf{V}_M^{-1}\hat{\boldsymbol{\beta}}_M, \qquad (\text{23-27})$$

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

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

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

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

$$\ln L_0 = n[P \ln P + (1 - P)\ln(1 - P)], \qquad (\text{23-28})$$

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

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

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

$$\text{LM} = \left(\sum_{i=1}^{n} g_i\mathbf{x}_i\right)'\left[\sum_{i=1}^{n} E[-h_i]\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\left(\sum_{i=1}^{n} g_i\mathbf{x}_i\right), \qquad (\text{23-29})$$

where $E[-h_i]$ is defined in (23-22) for the logit model and in (23-24) for the probit model.

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

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

where R^2 is the uncentered coefficient of determination in the regression of $(y_i - \bar{y})$ on $\mathbf{x}_i$ and $\bar{y}$ is the proportion of 1s in the sample. An alternative formulation based on the BHHH estimator, which we developed in Section 16.6.3 is also convenient. For any of the models (probit, logit, Gumbel, etc.), the first derivative vector can be written as

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

AQ: Confirm x-ref to Sec. 16.6.3 is correct

17-30

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

$$\text{LM} = n\left[\frac{1}{n}\mathbf{i}'(\mathbf{G}\mathbf{X})(\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{G}')\mathbf{i}\right] = nR_{\mathbf{i}}^2, \qquad (23\text{-}30)$$

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

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

23.4.4 SPECIFICATION TESTS FOR BINARY CHOICE MODELS

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

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

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

1. If x_2 is omitted from a model containing x_1 and x_2, (i.e. $\beta_2 \neq 0$) then
$$\text{plim}\,\hat{\beta}_1 = c_1\beta_1 + c_2\beta_2,$$
where c_1 and c_2 are complicated functions of the unknown parameters. The implication is that even if the omitted variable is uncorrelated with the included one, the coefficient on the included variable will be inconsistent.
2. If the disturbances in the underlying regression are heteroscedastic, then the maximum likelihood estimators are inconsistent and the covariance matrix is inappropriate.

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

Any of the three methods of hypothesis testing discussed here can be used to analyze these specification problems. The Lagrange multiplier test has the advantage that it can be carried out using the estimates from the restricted model, which sometimes brings a large saving in computational effort. This situation is especially true for the test for **heteroscedasticity.**[12]

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,

[12] 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).

Example 17.7 Testing for Structural Break in a Logit Model

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

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

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

$$LR = -2[-17673.09788 - (-9541.77802 - 7855.96999)] = 550.69744.$$

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

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

Another way to obtain the same result is to add to the pooled model the original six variables now multiplied by the *Female* dummy variable. We use the augmented $\mathbf{X}$ matrix $\mathbf{X}^* = [\mathbf{X}, female \times \mathbf{X}]$. The model with 12 variables is now estimated, and a test of the pooling hypothesis is done by testing the joint hypothesis that the coefficients on these six additional variables are zero. The Lagrange multiplier test is carried out by using this augmented model as well. To apply (17-32), the necessary derivatives are in (17-19). For the logit model, the derivative matrix is simply $\mathbf{G}^* = \text{diag}[y_i - \Lambda(\mathbf{x}_i^{*\prime}\beta)]$. For the LM test, the vector β that is used is the one for the restricted model. Thus, $\hat{\beta}^* = \left(\hat{\beta}'_{Pooled}, 0, 0, 0, 0, 0, 0\right)'$. The estimated probabilities that appear in $\mathbf{G}^*$ are simply those obtained from the pooled model. Then,

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

The pooling hypothesis is rejected by all three procedures.

TABLE 17.4 Estimated Models for Pooling Hypothesis

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

For a particular value z^*, we compute a set of n weights using the **kernel function,**

$$w_i(z^*) = K[(z^* - z_i)/(\lambda s)],$$

where

$$K(r_i) = P(r_i)[1 - P(r_i)],$$

and

$$P(r_i) = [1 + \exp(-cr_i)]^{-1}.$$

The constant $c = (\pi/\sqrt{3})^{-1} \approx 0.55133$ is used to standardize the logistic distribution that is used for the kernel function. (See Section 14.4.1.) The parameter λ is the smoothing (bandwidth) parameter. Large values will flatten the estimated function through $\bar{y}$, whereas values close to zero will allow greater variation in the function but might cause it to be unstable. There is no good theory for the choice, but some suggestions have been made based on descriptive statistics. [See Wong (1983) and Manski (1986).] Finally, the function value is estimated with

$$F(z^*) \approx \frac{\sum_{i=1}^{n} w_i(z^*) y_i}{\sum_{i=1}^{n} w_i(z^*)}.$$

The nonparametric estimator displays a relationship between $\mathbf{x}'\boldsymbol{\beta}$ and $E[y_i]$. At first blush, this relationship might suggest that we could deduce the marginal effects, but unfortunately, that is not the case. The coefficients in this setting are not meaningful, so all we can deduce is an estimate of the density, $f(z)$, by using first differences of the estimated regression function. It might seem, therefore, that the analysis has produced relatively little payoff for the effort. But that should come as no surprise if we reconsider the assumptions we have made to reach this point. The only assumptions made thus far are that for a given vector of covariates $\mathbf{x}_i$ and coefficient vector $\boldsymbol{\beta}$ (i.e., *any* $\boldsymbol{\beta}$), there exists a smooth function $F(\mathbf{x}'\boldsymbol{\beta}) = E[y_i \mid z_i]$. We have also assumed, at least implicitly, that the coefficients carry some information about the covariation of $\mathbf{x}'\boldsymbol{\beta}$ and the response variable. The technique will approximate any such function [see Manski (1986)].

There is a large and burgeoning literature on kernel estimation and nonparametric estimation in econometrics. [A recent application is Melenberg and van Soest (1996).] As this simple example suggests, with the radically different forms of the specified model, the information that is culled from the data changes radically as well. The general principle now made evident is that the fewer assumptions one makes about the population, the less precise the information that can be deduced by statistical techniques. That tradeoff is inherent in the methodology.

Aw! x-ref to Exm 17.12 ok? It's on msp 17-61

17.3.5 ~~23.7~~ **ENDOGENOUS RIGHT-HAND-SIDE VARIABLES IN BINARY CHOICE MODELS**

17.12 The analysis in Example ~~23.10~~ (Magazine Prices Revisited) suggests that the presence of endogenous right-hand-side variables in a binary choice model presents familiar problems for estimation. The problem is made worse in nonlinear models because even

if one has an instrumental variable readily at hand, it may not be immediately clear what is to be done with it. The instrumental variable estimator described in Chapter 12 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. (This will figure prominently in the analysis in Section 23.9.) Consider the model

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

Thus, w_i is endogenous in this model. The maximum likelihood estimators considered earlier will not consistently estimate $(\boldsymbol{\beta}, \gamma)$. [Without an additional specification that allows us to formalize $\text{Prob}(y_i = 1 \mid \mathbf{x}_i, w_i)$, we cannot state what the MLE will, in fact, estimate.] Suppose that we have a "relevant" (see Section 12.7) instrumental variable, z_i such that

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

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

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

However, y_i^* is not observed, y_i is. But the "residual," $y_i - \mathbf{x}_i'\boldsymbol{\beta} - \gamma w_i$, would have no meaning even if the true parameters were known.[37] One approach that was used in Avery et al. (1983), Butler and Chatterjee (1997), and Bertschek and Lechner (1998) is to assume that the instrumental variable is orthogonal to the residual $[y - \Phi(\mathbf{x}_i'\boldsymbol{\beta} + \gamma w_i)]$; that is,

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

This form of the moment equation, based on observables, can form the basis of a straightforward two-step GMM estimator. (See Chapter 15 for details.)

The GMM estimator is not less parametric than the full information maximum likelihood estimator described following because the probit model based on the normal distribution is still invoked to specify the moment equation.[38] Nothing is gained in simplicity or robustness of this approach to full information maximum likelihood estimation, which we now consider. (As Bertschek and Lechner argue, however, the gains might come in terms of practical implementation and computation time. The same considerations motivated Avery et al.)

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

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

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

a simultaneous equations model. The model equations are

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

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

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

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

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

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

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

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

Therefore,

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

17-33 (23.43)

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

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

17.8

Example 23.13 Labor Supply Model

5.2 and 17.1

In Examples 4.3 and 23.1, we examined a labor suppy model for married women using Mroz's (1987) data on labor supply. The wife's labor force participation equation suggested in Example 23.1 is

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

17.1

17-35

17.5

TABLE ~~23.11~~ Estimated Labor Supply Model

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

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

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

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

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

5.2 17.5

TB 17.5

As before, we use the Mroz (1987) labor supply data described in Example ~~4.3~~. Table ~~23.11~~ reports the single-equation and maximum likelihood estimates of the parameters of the two equations. Comparing the two sets of probit estimates, it appears that the (assumed) endogeneity of the husband's hours is not substantially affecting the estimates. There are two simple ways to test the hypothesis that ρ equals zero. The FIML estimator produces an estimated asymptotic standard error with the estimate of ρ, so a Wald test can be carried out. For the preceding results, the Wald statistic would be $(-0.4221/0.26921)^2 = 2.458$. The critical value from the chi-squared table for one degree of freedom would be 3.84, so we would not reject the hypothesis. The second approach would use the likelihood ratio test. Under the null hypothesis of exogeneity, the probit model and the regression equation can be estimated independently. The log-likelihood for the full model would be the sum of the two log-likelihoods, which would be −6357.508 based on the following results. Without the restriction $\rho = 0$, the combined log likelihood is −6357.093. Twice the difference is 0.831, which is also well under the 3.84 critical value, so on this basis as well, we would not reject the null hypothesis that $\rho = 0$.

KT

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

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

32 17-33

Note that (~~23-43~~) satisfies the assumption; however, they reach this point without assuming either joint or marginal normality. The authors propose a three-step, semiparametric

approach to estimating the structural parameters. In an application somewhat similar to Example 17.8, they apply the technique to a labor force participation model for British men in which a variable of interest is a dummy variable for education greater than 16 years; the endogenous variable in the participation equation, also of interest, is earned income of the spouse, and an instrumental variable is a welfare benefit entitlement. Their findings are rather more substantial than ours; they find that when the endogeneity of other family income is accommodated in the equation, the education coefficient increases by 40 percent and remains significant, but the coefficient on other income increases by more than tenfold.

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

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

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

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

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

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

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

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

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

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

studied. In some cases, it appears so. But the maximum likelihood estimator, on which all the fit measures are based, is not chosen so as to maximize a fitting criterion based on prediction of y as it is in the classical regression (which maximizes R^2). It is chosen to maximize the joint density of the observed dependent variables. It remains an interesting question for research whether fitting y well or obtaining good parameter estimates is a preferable estimation criterion. Evidently, they need not be the same thing.

Example 23.6 Prediction with a Probit Model

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

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

17.3.6 ~~23.4.6~~ CHOICE-BASED SAMPLING — ENDOGENOUS

In some studies [e.g., Boyes, Hoffman, and Low (1989), Greene (1992)], the mix of ones and zeros in the observed sample of the dependent variable is deliberately skewed in favor of one outcome or the other to achieve a more balanced sample than random sampling would produce. The sampling is said to be **choice based.** In the studies noted, the dependent variable measured the occurrence of loan default, which is a relatively uncommon occurrence. To enrich the sample, observations with $y = 1$ (default) were oversampled. Intuition should suggest (correctly) that the bias in the sample should be transmitted to the parameter estimates, which will be estimated so as to mimic the sample, not the population, which is known to be different. Manski and Lerman (1977) derived the weighted endogenous sampling maximum likelihood (WESML) estimator for this situation. The estimator requires that the true population proportions, ω_1 and ω_0, be known. Let p_1 and p_0 be the sample proportions of ones and zeros. Then the estimator is obtained by maximizing a weighted log-likelihood,

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

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

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

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 23.4.1, $\mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}$ (with weighted $\mathbf{B}$ and $\mathbf{H}$), instead of $\mathbf{B}$ or $\mathbf{H}$ alone. (The weights are not squared in computing $\mathbf{B}$.)[20]

23.4.7 DYNAMIC BINARY CHOICE MODELS

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

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

Lagged effects, or **persistence,** in a binary choice setting can arise from three sources, serial correlation in ε_{it}, the **heterogeneity,** α_i, or true **state dependence** through the term $\gamma y_{i,t-1}$. Chiappori (1998) [and see Arellano (2001)] suggests an application to the French automobile insurance market in which the incentives built into the pricing system are such that having an accident in one period should lower the probability of having one in the next (state dependence), but, some drivers remain more likely to have accidents than others in every period, which would reflect the heterogeneity instead. State dependence is likely to be particularly important in the typical panel which has only a few observations for each individual. Heckman (1981a) examined this issue at length. Among his findings were that the somewhat muted small sample bias in fixed effects models with $T = 8$ was made much worse when there was state dependence. A related problem is that with a relatively short panel, the **initial conditions,** y_{i0}, have a crucial impact on the entire path of outcomes. Modeling dynamic effects and initial conditions in binary choice models is more complex than in the linear model, and by comparison there are relatively fewer firm results in the applied literature.[21]

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 is usefulness for practical application. Lewbel (2000) has extended his fixed effects estimator to dynamic models as well. In this framework, the narrow assumptions about the independent variables somewhat

[20] 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.

[21] 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.

Example 17.9 Credit Scoring

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

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

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 17.6 presents the unweighted and weighted estimates for this application. The change in the estimates produced by the weighting is quite modest, save for the constant term. The results are consistent with the conventional wisdom that *Income* and *OwnRent* are two important variables in a credit application and self employment receives a substantial negative weight. But, as might be expected, the single most significant influence on cardholder status is major derogatory reports. Since lenders are strongly focused on default probability, past evidence of default behavior will be a major consideration.

Table 17.6 Estimated Card Application Equation (t ratios in parentheses)

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

17.3.7 SPECIFICATION ANALYSIS

In his survey of qualitative response models, Amemiya (1981) reports the following widely cited approximations for the linear probability (LP) model: Over the range of probabilities of 30 to 70 percent,

$$\hat{\beta}_{LP} \approx 0.4\beta_{\text{probit}} \text{ for the slopes,}$$
$$\hat{\beta}_{LP} \approx 0.25\beta_{\text{logit}} \text{ for the slopes.}$$

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

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

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

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

$$\text{LM} = n\left[\frac{1}{n}\mathbf{i}'(\mathbf{G}\mathbf{X})(\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{G}')\mathbf{i}\right] = nR_{\mathbf{i}}^2, \tag{23-30}$$

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

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

23.6.3 SPECIFICATION TESTS FOR BINARY CHOICE MODELS

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

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

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

1. If x_2 is omitted from a model containing x_1 and x_2, (i.e., $\beta_2 \neq 0$) then
$$\text{plim}\,\hat{\beta}_1 = c_1\beta_1 + c_2\beta_2,$$
where c_1 and c_2 are complicated functions of the unknown parameters. The implication is that even if the omitted variable is uncorrelated with the included one, the coefficient on the included variable will be inconsistent.
2. If the disturbances in the underlying regression are heteroscedastic, then the maximum likelihood estimators are inconsistent and the covariance matrix is inappropriate.

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

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

To reiterate, the Lagrange multiplier statistic is computed as follows. Let the null hypothesis, H_0, be a specification of the model, and let H_1 be the alternative. For example,

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

H_0 might specify that only variables $\mathbf{x}_1$ appear in the model, whereas H_1 might specify that $\mathbf{x}_2$ appears in the model as well. The statistic is

$$\text{LM} = \mathbf{g}_0'\mathbf{V}_0^{-1}\mathbf{g}_0,$$

where $\mathbf{g}_0$ is the vector of derivatives of the log-likelihood as specified by H_1 but evaluated at the maximum likelihood estimator of the parameters assuming that H_0 is true, and $\mathbf{V}_0^{-1}$ is any of the three consistent estimators of the asymptotic variance matrix of the maximum likelihood estimator under H_1, also computed using the maximum likelihood estimators based on H_0. The statistic is asymptotically distributed as chi-squared with degrees of freedom equal to the number of restrictions.

17.3.7.a ~~23.4.4.a~~ **Omitted Variables**

The hypothesis to be tested is

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

33 17-34 (~~23-31~~)

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, $[\mathbf{x}_1, \mathbf{x}_2]$.

30 17-31 31 17-32

The statistic is then computed according to (~~23-29~~) or (~~23-30~~). It is noteworthy that in this case as in many others, the Lagrange multiplier is the coefficient of determination in a regression. The likelihood ratio test is equally straightforward. Using the estimates of the two models, the statistic is simply $2(\ln L_1 - \ln L_0)$.

17.3.7.b ~~23.4.4.b~~ **Heteroscedasticity**

14 26

We use the general formulation analyzed by Harvey (1976) (see Section ~~16~~9.2.a),[18]

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

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

34

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

17-35 (~~23-32~~)

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

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

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

FN 26

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

The log-likelihood is

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

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

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

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

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

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

computed at the maximum likelihood estimator, assuming that $\boldsymbol{\gamma} = \mathbf{0}$. Then (~~23-29~~) or (~~23-30~~) can be used as usual for the Lagrange multiplier statistic.

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

Example ~~23.6~~ Specification Tests in a Labor Force Participation Model

Using the data described in Example ~~23.1~~, we fit a probit model for labor force participation based on the specification

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

For these data, $P = 428/753 = 0.568393$. The restricted (all slopes equal zero, free constant term) log-likelihood is $325 \times \ln(325/753) + 428 \times \ln(428/753) = -514.8732$. The unrestricted log-likelihood for the probit model is -490.8478. The chi-squared statistic is, therefore, 48.05072. The critical value from the chi-squared distribution with 5 degrees of freedom is 11.07, so the joint hypothesis that the coefficients on *age*, age^2, *family income*, and *kids* are all zero is rejected.

Consider the alternative hypothesis, that the constant term and the coefficients on *age*, age^2, *family income*, and *education* are the same whether *kids* equals one or zero, against the alternative that an altogether different equation applies for the two groups of women, those with *kids* = 1 and those with *kids* = 0. To test this hypothesis, we would use a counterpart to the **Chow test** of Section 6.4 and Example 6.7. The restricted model in this instance would be based on the pooled data set of all 753 observations. The log-likelihood for the pooled model—which has a constant term, *age*, age^2, *family income*, and *education* is -496.8663. The log-likelihoods for this model based on the 524 observations with *kids* = 1 and the 229 observations with *kids* = 0 are -347.87441 and -141.60501, respectively. The log-likelihood for the unrestricted model with separate coefficient vectors is thus the sum, -489.47942. The chi-squared statistic for testing the five restrictions of the pooled model is twice the difference, $LR = 2[-489.47942 - (-496.8663)] = 14.7738$. The 95 percent critical value from the chi-squared distribution with 5 degrees of freedom is 11.07, so at this significance level, the hypothesis that the constant terms and the coefficients on *age*, age^2, *family income*, and *education* are the same is rejected. (The 99 percent critical value is 15.09.)

17-45

17.7

TABLE 23.4 Estimated Coefficients

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

*Marginal effect and estimated standard error include both mean (β) and variance (γ) effects.

17.7

Table 23.4 presents estimates of the probit model with a correction for heteroscedasticity of the form

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

The three tests for homoscedasticity give

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

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

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

95

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

23.4.5 MEASURING GOODNESS OF FIT

There have been many fit measures suggested for QR models.[14] At a minimum, one should report the maximized value of the log-likelihood function, ln L. Because the hypothesis that all the slopes in the model are zero is often interesting, the log-likelihood computed with only a constant term, ln L_0 [see (23-28)], should also be reported. An analog to the R^2 in a conventional regression is McFadden's (1974) likelihood ratio index,

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

This measure has an intuitive appeal in that it is bounded by zero and one. (See Section 16.6.5.) If all the slope coefficients are zero, then it equals zero. There is no way to make LRI equal 1, although one can come close. If F_i is always one when y equals one and zero when y equals zero, then ln L equals zero (the log of one) and LRI equals one. It has been suggested that this finding is indicative of a "perfect fit" and that LRI increases as the fit of the model improves. To a degree, this point is true (see the analysis in Section 23.8.4). Unfortunately, the values between zero and one have no natural interpretation. If $F(\mathbf{x}_i'\boldsymbol{\beta})$ is a proper pdf, then even with many regressors the model cannot fit perfectly unless $\mathbf{x}_i'\boldsymbol{\beta}$ goes to $+\infty$ or $-\infty$. As a practical matter, it does happen. But when it does, it

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

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

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

This still involves a $T = 7$ order normal integration, which is approximated in the study using a simulator similar to the GHK simulator discussed in 17.3.3. 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 17.5.1.

17.4 ~~23.5~~ BINARY CHOICE MODELS FOR PANEL DATA

Qualitative response models have been a growth industry in econometrics. The recent literature, particularly in the area of panel data analysis, has produced a number of new techniques. The availability of high-quality panel data sets on microeconomic behavior has maintained an interest in extending the models of Chapter 9 to binary (and other discrete choice) models. In this section, we will survey a few results from this rapidly growing literature.

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

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

The second line of this definition is often written

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

to indicate a variable that equals one when the condition in parentheses is true and zero when it is not. Ideally, we would like to specify that ε_{it} and ε_{is} are freely correlated within a group, but uncorrelated across groups. But doing so will involve computing joint probabilities from a T_i variate distribution, which is generally problematic.[23] (We will return to this issue later.) A more promising approach is an effects model,

$$\begin{aligned} y_{it}^* &= \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it} + u_i, \quad i = 1, \ldots, n, t = 1, \ldots, T_i, \\ y_{it} &= 1 \quad \text{if } y_{it}^* > 0, \text{ and 0 otherwise,} \end{aligned} \tag{17-40}$$

where, as before (see ~~Section 9.5~~), u_i is the unobserved, individual specific heterogeneity. Once again, we distinguish between "random" and "fixed" effects models by the relationship between u_i and $\mathbf{x}_{it}$. The assumption that u_i is unrelated to $\mathbf{x}_{it}$, so that the conditional distribution $f(u_i \mid \mathbf{x}_{it})$ is not dependent on $\mathbf{x}_{it}$, produces the **random effects model.** Note that this places a restriction on the distribution of the heterogeneity.

[23] A "limited information" approach based on the GMM estimation method has been suggested by Avery, Hansen, and Hotz (1983). With recent advances in simulation-based computation of multinormal integrals (see Section 17.5.1), some work on such a panel data estimator has appeared in the literature. See, for example, Geweke, Keane, and Runkle (1994, 1997). The GEE estimator of Diggle, Liang, and Zeger (1994) [see also, Liang and Zeger (1986) and Stata (2006)] seems to be another possibility. However, in all these cases, it must be remembered that the procedure specifies estimation of a correlation matrix for a T_i vector of unobserved variables based on a dependent variable that takes only two values. We should not be too optimistic about this if T_i is even moderately large.

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

As we shall see shortly, this is a modeling framework that is fraught with difficulties and unconventional estimation problems. Among them are the following: estimation of the random effects model requires very strong assumptions about the heterogeneity; the fixed effects model encounters an **incidental parameters problem** that renders the maximum likelihood estimator inconsistent.

17.4.1 The Pooled Estimator

To begin, it is useful to consider the pooled estimator that results if we simply ignore the heterogeneity, u_i in (17-40) and fit the model as if the cross section specification of Section 17.2.2 applies. In this instance, the adage that "ignoring the heterogeneity does not make it go away," applies even more forcefully than in the linear regression case.

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

The random effects case is more benign. From (17-40), the marginal probability implied by the model is

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

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

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

which is not computable. The useful result would be

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

Wooldridge (2002a) shows that the end result, assuming normality of both v_{it} and u_i is $E_u[PE(\mathbf{x}_{it},u_i)] = \delta\phi(\mathbf{x}_{it}'\boldsymbol{\delta})$. Thus far, surprisingly, it would seem that simply pooling the data and using the simple MLE "works." The estimated standard errors will be incorrect, so a correction such as the cluster estimator shown in Section 14.8.4 would be appropriate. Three considerations suggest that one might want to proceed to the full MLE in spite of these results: (1) The pooled estimator will be inefficient compared to the full MLE; (2) the pooled estimator does not produce an estimator of σ_u which might be of interest in its own right; (3) the FIML estimator is available in contemporary software and is no more difficult to estimate then the pooled estimator. Note that the pooled estimator is not justified (over the FIML approach) on robustness considerations because the same normality and random effects assumptions that are needed to obtain the FIML estimator will be needed to obtain the preceding results for the pooled estimator.

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

As we shall see shortly, this is a modeling framework that is fraught with difficulties and unconventional estimation problems. Among them are the following: estimation of the random effects model requires very strong assumptions about the heterogeneity; the fixed effects model encounters an **incidental parameters problem** that renders the maximum likelihood estimator inconsistent.

17.4.2 ~~23.5.1~~ Random Effects Models

A specification that has the same structure as the random effects model of Section 9.5 has been implemented by Butler and Moffitt (1982). We will sketch the derivation to suggest how random effects can be handled in discrete and limited dependent variable models such as this one. Full details on estimation and inference may be found in Butler and Moffitt (1982) and Greene (1995a). We will then examine some extensions of the Butler and Moffitt model.

The random effects model specifies

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

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

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

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

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

and $\mathbf{X}$ indicates all the exogenous data in the sample, $\mathbf{x}_{it}$ for all i and t.[24] Then,

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

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

and

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

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

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

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

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

$$L_i = P(y_{i1}, \ldots, y_{iT_i} \mid \mathbf{X}) = \int_{L_{iT_i}}^{U_{iT_i}} \cdots \int_{L_{i1}}^{U_{i1}} f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i}) d\varepsilon_{i1} d\varepsilon_{i2} \ldots d\varepsilon_{iT_i}. \quad \text{(23-38)}$$

17-41 40

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

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

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

So,

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

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

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

17-40

Inserting this result in ~~(23-38)~~ produces

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

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

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

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

41

$$L_i = P[y_{i1}, \ldots, y_{iT_i} \mid \mathbf{X}] = \int_{-\infty}^{+\infty} \left[\prod_{t=1}^{T_i} \left(\int_{L_{it}}^{U_{it}} f(\varepsilon_{it} \mid u_i)\, d\varepsilon_{it} \right) \right] f(u_i)\, du_i. \quad \text{17-42 (23-39)}$$

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

42

$$L_i = P[y_{i1}, \ldots, y_{iT_i} \mid \mathbf{X}] = \int_{-\infty}^{+\infty} \left[\prod_{t=1}^{T_i} \text{Prob}(Y_{it} = y_{it} \mid \mathbf{x}_{it}'\boldsymbol{\beta} + u_i) \right] f(u_i)\, du_i. \quad \text{17-43 (23-40)}$$

It remains to specify the distributions, but the important result thus far is that the entire computation requires only one-dimensional integration. The inner probabilities may be any of the models we have considered so far, such as probit, logit, Gumbel, and so on. The intricate part that remains is to determine how to do the outer integration. **Butler and Moffitt's method** assuming that u_i is normally distributed is detailed in Section ~~16.9.6.b.~~

KT

14.9.6.c

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

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

A remaining problem with the Butler and Moffitt specification is its assumption of normality. In general, other distributions are problematic because of the difficulty of finding either a closed form for the integral or a satisfactory method of approximating the integral. An alternative approach that allows some flexibility is the method of **maximum simulated likelihood** (MSL), which was discussed in Section 15.6 and Chapter 17. The transformed likelihood we derived in (23-40) is an expectation:

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

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

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

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

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

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

This suggests, based on the results in Chapter 17, 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

$$\ln L_{Simulated} = \sum_{i=1}^{n} \ln \left\{ \frac{1}{R} \sum_{r=1}^{R} \left[\prod_{t=1}^{T_i} F[q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma_u u_{ir})] \right] \right\}. \qquad (23\text{-}41)$$

This function is maximized with respect $\boldsymbol{\beta}$ and σ_u. Note that in the preceding, as in the quadrature approximated log-likelihood, the model can be based on a probit, logit, or any other functional form desired. There is an additional degree of flexibility in this approach. The Hermite quadrature approach is essentially limited by its functional form to the normal distribution. But, in the simulation approach, u_{ir} can come from some other distribution. For example, it might be believed that the dispersion of the heterogeneity is greater than implied by a normal distribution. The logistic distribution might be preferable. A random sample from the logistic distribution can be created by sampling $(w_{i1}, \ldots, w_{iR})$ from the standard uniform $[0, 1]$ distribution; then $u_{ir} = \ln[w_{ir}/(1 - w_{ir})]$. Other distributions, such as the uniform itself, are also possible.

We have examined two approaches to estimation of a probit model with random effects. GMM estimation is another possibility. Avery, Hansen, and Hotz (1983), Bertschek and Lechner (1998), and Inkmann (2000) examine this approach; the latter two offer some comparison with the quadrature and simulation-based estimators considered here. (Our application in Example 23.16 will use the Bertschek and Lechner data.)

The preceding opens another possibility. The random effects model can be cast as a model with a random constant term;

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

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

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

where $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Gamma}\mathbf{u}_i$ where $\boldsymbol{\Gamma}$ is a nonnegative definite diagonal matrix—some of its diagonal elements could be zero for nonrandom parameters. The method of estimation is essentially the same as before. 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 23.11. Example 23.11 also presents a similar model that assumes that the distribution of $\boldsymbol{\beta}_i$ is discrete rather than continuous.

17.4.3 ~~23.5.2~~ Fixed Effects Models

The fixed effects model is

$$y_{it}^* = \alpha_i d_{it} + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \ldots, n, \; t = 1, \ldots, T_i, \qquad (17\text{-}45)$$
$$y_{it} = 1 \quad \text{if } y_{it}^* > 0, \text{ and 0 otherwise,}$$

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

$$\ln L = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \ln P(y_{it} \mid \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}), \qquad (17\text{-}46)$$

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. What follows can be extended to any index function model, but for the present, we'll confine our attention

where $q_{it} = 2y_{it} - 1$