21
MODELS FOR DISCRETE CHOICE

21.1 INTRODUCTION

There are many settings in which the economic outcome we seek to model is a discrete choice among a set of alternatives, rather than a continuous measure of some activity. Consider, for example, modeling labor force participation, the decision of whether or not to make a major purchase, or the decision of which candidate to vote for in an election. For the first of these examples, intuition would suggest that factors such as age, education, marital status, number of children, and some economic data would be relevant in explaining whether an individual chooses to seek work or not in a given period. But something is obviously lacking if this example is treated as the same sort of regression model we used to analyze consumption or the costs of production or the movements of exchange rates. In this chapter, we shall examine a variety of what have come to be known as **qualitative response (QR)** models. There are numerous different types that apply in different situations. What they have in common is that they are models in which the dependent variable is an indicator of a discrete choice, such as a "yes or no" decision. In general, conventional regression methods are inappropriate in these cases.

This chapter is a lengthy but far from complete survey of topics in estimating QR models. Almost none of these models can be consistently estimated with linear regression methods. Therefore, readers interested in the mechanics of estimation may want to review the material in Appendices D and E before continuing. In most cases, the method of estimation is **maximum likelihood.** The various properties of maximum likelihood estimators are discussed in Chapter 17. We shall assume throughout this chapter 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 Maddala and Flores-Lagunes (2001).

21.2 DISCRETE CHOICE MODELS

The general class of models we shall consider are those for which the dependent variable takes values $0, 1, 2, \ldots$. In a few cases, the values will themselves be meaningful, as in the following:

1. Number of patents: $y = 0, 1, 2, \ldots$ These are **count data.**

In most of the cases we shall study, the values taken by the dependent variables are merely a coding for some qualitative outcome. Some examples are as follows:

2. Labor force participation: We equate "no" with 0 and "yes" with 1. These decisions are **qualitative choices.** The 0/1 coding is a mere convenience.
3. Opinions of a certain type of legislation: Let 0 represent "strongly opposed," 1 "opposed," 2 "neutral," 3 "support," and 4 "strongly support." These numbers are **rankings,** and the values chosen are not quantitative but merely an ordering. The difference between the outcomes represented by 1 and 0 is not necessarily the same as that between 2 and 1.
4. The occupational field chosen by an individual: Let 0 be clerk, 1 engineer, 2 lawyer, 3 politician, and so on. These data are merely categories, giving neither a ranking nor a count.
5. Consumer choice among alternative shopping areas: This case has the same characteristics as example 4, but the appropriate model is a bit different. These two examples will differ in the extent to which the choice is based on characteristics of the individual, which are probably dominant in the occupational choice, as opposed to attributes of the choices, which is likely the more important consideration in the choice of shopping venue.

None of these situations lends themselves readily to our familiar type of regression analysis. Nonetheless, in each case, we can 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:

$$\text{Prob(event } j \text{ occurs)} = \text{Prob}(Y = j) = F[\text{relevant effects, parameters}]. \quad \textbf{(21-1)}$$

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 alternatives.

Example 21.1 Labor Force Participation Model

In Example 4.3 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 $y_i = 1$ if in the labor force, and 0 if not.

21.3 MODELS FOR BINARY CHOICE

Models for explaining a binary (0/1) dependent variable typically arise in two contexts. In many cases, the analyst is essentially interested in a regressionlike model of the sort considered in Chapters 2 to 9. With data on the variable of interest and a set of covariates, the analyst is interested in specifying a relationship between the former and the latter, more or less along the lines of the models we have already studied. The relationship between voting behavior and income is typical. In other cases, the **binary choice** model arises in the context of a model in which the nature of the observed data dictate the special treatment of a binary choice model. 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.

21.3.1 THE REGRESSION APPROACH

To focus ideas, consider the model of labor force participation suggested in Example 21.1.[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

$$\begin{aligned} \text{Prob}(Y = 1 \mid \mathbf{x}) &= F(\mathbf{x}, \boldsymbol{\beta}) \\ \text{Prob}(Y = 0 \mid \mathbf{x}) &= 1 - F(\mathbf{x}, \boldsymbol{\beta}). \end{aligned} \tag{21-2}$$

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,

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

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

$$y = E[y \mid \mathbf{x}] + (y - E[y \mid \mathbf{x}]) = \mathbf{x}'\boldsymbol{\beta} + \varepsilon. \tag{21-3}$$

The **linear probability model** has a number of shortcomings. A minor complication arises because ε is heteroscedastic in a way that depends on $\boldsymbol{\beta}$. Since $\mathbf{x}'\boldsymbol{\beta} + \varepsilon$ 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

$$\text{Var}[\varepsilon \mid \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}(1 - \mathbf{x}'\boldsymbol{\beta}). \tag{21-4}$$

We could manage this complication with an FGLS estimator in the fashion of Chapter 11. 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.

[1] Models for qualitative dependent variables can now be found in most disciplines in economics. A frequent use is in labor economics in the analysis of microlevel data sets.

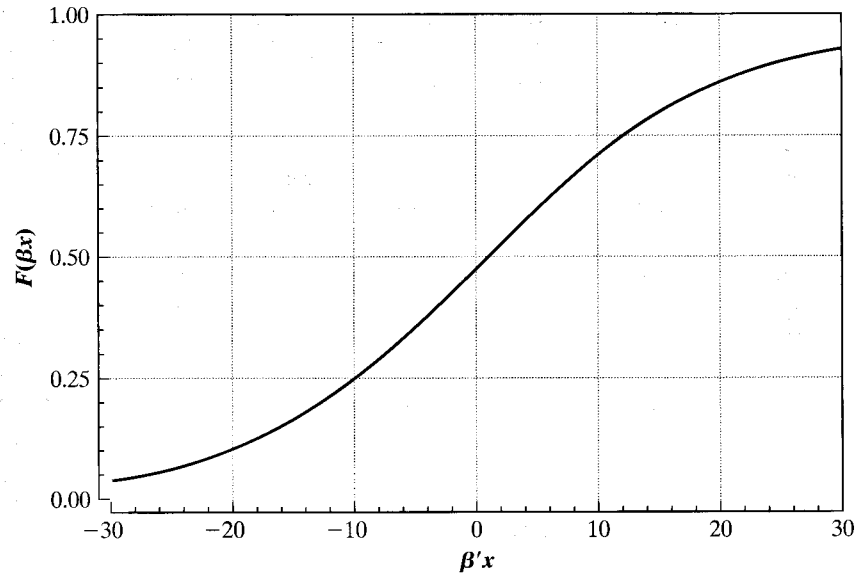


FIGURE 21.1 Model for a Probability.

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 model is becoming less frequently used except as a basis for comparison to some other more appropriate models.[2]

Our requirement, then, is a model that will produce predictions consistent with the underlying theory in (21-1). For a given regressor vector, we would expect

$$\begin{aligned} &\lim_{\mathbf{x}'\boldsymbol{\beta}\to+\infty} \text{Prob}(Y = 1 \mid \mathbf{x}) = 1 \\ &\lim_{\mathbf{x}'\boldsymbol{\beta}\to-\infty} \text{Prob}(Y = 1 \mid \mathbf{x}) = 0. \end{aligned} \tag{21-5}$$

See Figure 21.1. In principle, any proper, continuous probability distribution defined over the real line will suffice. The normal distribution has been used in many analyses. giving rise to the **probit** model,

$$\text{Prob}(Y = 1 \mid \mathbf{x}) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \phi(t)\, dt = \Phi(\mathbf{x}'\boldsymbol{\beta}). \tag{21-6}$$

The function $\Phi(.)$ is a commonly used notation for the standard normal distribution.

[2]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).

Partly because of its mathematical convenience, the logistic distribution,

$$\text{Prob}(Y = 1 \mid \mathbf{x}) = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}'\boldsymbol{\beta}}} = \Lambda(\mathbf{x}'\boldsymbol{\beta}), \tag{21-7}$$

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 **Weibull** 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,[3] 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 $y = 0$ when $\mathbf{x}'\boldsymbol{\beta}$ is extremely small (and smaller probabilities to $Y = 0$ when $\boldsymbol{\beta}'\mathbf{x}$ is very large) than the normal distribution. It is difficult to provide practical generalities on this basis, however, since 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 (Ys equal to 1) or very few nonresponses (Ys 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:

$$E[y \mid \mathbf{x}] = 0[1 - F(\mathbf{x}'\boldsymbol{\beta})] + 1[F(\mathbf{x}'\boldsymbol{\beta})] = F(\mathbf{x}'\boldsymbol{\beta}). \tag{21-8}$$

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,

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = \left\{ \frac{dF(\mathbf{x}'\boldsymbol{\beta})}{d(\mathbf{x}'\boldsymbol{\beta})} \right\} \boldsymbol{\beta} = f(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}, \tag{21-9}$$

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

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

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = \phi(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}, \tag{21-10}$$

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

$$\frac{d\Lambda(\mathbf{x}'\boldsymbol{\beta})}{d(\mathbf{x}'\boldsymbol{\beta})} = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{(1+e^{\mathbf{x}'\boldsymbol{\beta}})^2} = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1-\Lambda(\mathbf{x}'\boldsymbol{\beta})]. \tag{21-11}$$

Thus, in the logit model,

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = \Lambda(\mathbf{x}'\boldsymbol{\beta})[1-\Lambda(\mathbf{x}'\boldsymbol{\beta})]\boldsymbol{\beta}. \tag{21-12}$$

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. The functions are continuous with continuous first derivatives, so Theorem D.12 (the Slutsky theorem) and assuming that the data are "well behaved" a law of large numbers (Theorems D.4 and D.5) apply; in large samples these will give the same answer. 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. Since the derivative is with respect to a small change, it is not appropriate to apply (21-10) for the effect of a change in a dummy variable, or change of state. The appropriate marginal effect for a binary independent variable, say d, would be

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

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 21.3, the difference in the two probabilities for the probit model is $(0.5702 - 0.1057) = 0.4645$, whereas the derivative approximation reported below 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.

21.3.2 LATENT REGRESSION—INDEX FUNCTION MODELS

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

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

We assume that ε has mean zero and has either a standardized logistic with (known) variance $\pi^2/3$ [see (21-7)] or a standard normal distribution with variance one [see (21-6)]. We do not observe the net benefit of the purchase, only whether it is made or not. Therefore, our observation is

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

In this formulation, $\mathbf{x}'\boldsymbol{\beta}$ is called the index function.

Two aspects of this construction merit our attention. First, 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 data so it cannot be estimated. Second, the assumption of zero for the threshold 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 an 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

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

Since α is unknown, the difference $(\alpha - a)$ remains an unknown parameter. With the two normalizations,

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

If the distribution is symmetric, as are the normal and logistic, then

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

which provides an underlying structural model for the probability.

Example 21.2 Structural Equations for a Probit Model

Nakosteen and Zimmer (1980) analyze a model of migration based on the following structure:[5] For individual i, the market wage that can be earned at the present location is

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

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

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

[5] 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 21.5 is another example. 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.

would be

$$y_m^* = \mathbf{x}_m'\boldsymbol{\gamma} + \varepsilon_m.$$

Migration, however, 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

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

Since 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. If the disturbances are normally distributed, then the probit model we analyzed earlier is produced. Logistic disturbances produce the logit model instead.

21.3.3 RANDOM UTILITY MODELS

An alternative interpretation of data on individual choices is provided by the **random utility model.** Suppose that in the Nakosteen–Zimmer framework, y_m and y_p represent the individual's utility of two choices, which we might denote U^a and U^b. For another 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,

$$U^a = \mathbf{x}'\boldsymbol{\beta}_a + \varepsilon_a \quad \text{and} \quad U^b = \mathbf{x}'\boldsymbol{\beta}_b + \varepsilon_b. \tag{21-13}$$

Then, if we denote by $Y = 1$ the consumer's choice of alternative a, we have

$$\begin{aligned} \text{Prob}[Y = 1 \,|\, \mathbf{x}] &= \text{Prob}[U^a > U^b] \\ &= \text{Prob}[\mathbf{x}'\boldsymbol{\beta}_a + \varepsilon_a - \mathbf{x}'\boldsymbol{\beta}_b - \varepsilon_b > 0 \,|\, \mathbf{x}] \\ &= \text{Prob}[\mathbf{x}'(\boldsymbol{\beta}_a - \boldsymbol{\beta}_b) + \varepsilon_a - \varepsilon_b > 0 \,|\, \mathbf{x}] \\ &= \text{Prob}[\mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0 \,|\, \mathbf{x}] \end{aligned} \tag{21-14}$$

once again.

21.4 ESTIMATION AND INFERENCE IN BINARY CHOICE MODELS

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

or likelihood function,

$$\text{Prob}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \mathbf{X}) = \prod_{y_i=0}[1 - F(\mathbf{x}_i'\boldsymbol{\beta})] \prod_{y_i=1} F(\mathbf{x}_i'\boldsymbol{\beta}), \tag{21-15}$$

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

$$L(\boldsymbol{\beta} \mid \text{data}) = \prod_{i=1}^{n}[F(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i}[1 - F(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y_i}. \tag{21-16}$$

Taking logs, we obtain

$$\ln L = \sum_{i=1}^{n}\{y_i \ln F(\mathbf{x}_i'\boldsymbol{\beta}) + (1 - y_i)\ln[1 - F(\mathbf{x}_i'\boldsymbol{\beta})]\}.^{6} \tag{21-17}$$

The **likelihood equations** are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\left[\frac{y_i f_i}{F_i} + (1 - y_i)\frac{-f_i}{(1 - F_i)}\right]\mathbf{x}_i = \mathbf{0} \tag{21-18}$$

where f_i is the density, $dF_i/d(\mathbf{x}_i'\boldsymbol{\beta})$. [In (21-18) and later, we will use the subscript i to indicate that the function has an argument $\mathbf{x}_i'\boldsymbol{\beta}$.] The choice of a particular form for F_i leads to the empirical model.

Unless we are using the linear probability model, the likelihood equations in (21-18) 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 (21-7) and (21-11) in (21-18), we get, after a bit of manipulation, the likelihood equations

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}(y_i - \Lambda_i)\mathbf{x}_i = \mathbf{0}. \tag{21-19}$$

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

$$\ln L = \sum_{y_i=0}\ln[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})] + \sum_{y_i=1}\ln \Phi(\mathbf{x}_i'\boldsymbol{\beta}). \tag{21-20}$$

The first-order conditions for maximizing L are

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

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

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

[8] 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 20.3.5.

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

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

where $q_i = 2y_i - 1$.

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

$$\mathbf{H} = \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_i \Lambda_i (1 - \Lambda_i) \mathbf{x}_i \mathbf{x}_i'. \tag{21-22}$$

Since the second derivatives do not involve the random variable y_i, 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 (21-21). 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,

$$\mathbf{H} = \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} -\lambda_i (\lambda_i + \mathbf{x}_i' \boldsymbol{\beta}) \mathbf{x}_i \mathbf{x}_i'. \tag{21-23}$$

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. Since truncation always reduces variance—see Theorem 22.3—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 (17-18) and Example 17.4] 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 (21-19)] and $g_i = \lambda_i$ for the probit model [see (21-21)]. 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 (21-21)], this result is not true for the probit model. Amemiya (1981) showed that for the probit model,

$$E\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'} \right]_{\text{probit}} = \sum_{i=1}^{n} \lambda_{0i} \lambda_{i1} \mathbf{x}_i \mathbf{x}_i'. \tag{21-24}$$

Once again, the scalar part of the expression is always negative [see (21-23) and note that λ_{0i} is always negative and λ_{i1} is always positive]. The estimator of the asymptotic

[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 22.

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 below, though, for certain hypothesis tests, the BHHH estimator is a more convenient choice.

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,

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

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 the next section, $\mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}$ (with weighted $\mathbf{B}$ and $\mathbf{H}$), instead of $\mathbf{B}$ or $\mathbf{H}$ alone. (The weights are not squared in computing $\mathbf{B}$.)[11]

21.4.1 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 17.9 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

[11] 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.

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. 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.

21.4.2 MARGINAL EFFECTS

The predicted probabilities, $F(\mathbf{x}'\hat{\boldsymbol{\beta}}) = \hat{F}$ and the estimated marginal 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 5.2.4. 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 results is useful when a marginal effect is computed for a dummy variable. In that case, the estimated effect is

$$\Delta \hat{F} = \hat{F} \mid d = 1 - \hat{F} \mid d = 0. \tag{21-25}$$

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 (21-26)

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

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

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

TABLE 21.1 Estimated Probability Models

	Linear		*Logistic*		*Probit*		*Weibull*	
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	0.499	1.426	0.468	1.562	0.325
$f(\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}})$	1.000		0.189		0.328		0.208	

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

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

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

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

Collecting terms, we obtain

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

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

Example 21.3 Probability Models

The data listed in Appendix Table F21.1 were taken from a study by Spector and Mazzeo (1980), which examined whether a new method of teaching economics, the Personalized System of Instruction (PSI), significantly influenced performance in later economics courses. The "dependent variable" used in our application is GRADE, which indicates the whether a student's grade in an intermediate macroeconomics course was higher than that in the principles course. The other variables are GPA, their grade point average; TUCE, the score on a pretest that indicates entering knowledge of the material; and PSI, the binary variable indicator of whether the student was exposed to the new teaching method. (Spector and Mazzeo's specific equation was somewhat different from the one estimated here.)

Table 21.1 presents four sets of parameter estimates. The slope parameters and derivatives were computed for four probability models: linear, probit, logit, and Weibull. 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.

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.[12] One indicator is a comparison of the coefficients. In view of the different variances of the distributions, one for the normal and $\pi^2/3$ for the logistic, we might expect to obtain comparable estimates by multiplying the probit coefficients by $\pi/\sqrt{3} \approx 1.8$. Amemiya (1981) found, through trial and error, that scaling by 1.6 instead produced better results. This proportionality result is frequently cited. The result in (21-9) may help to explain the finding. The index $\mathbf{x}'\boldsymbol{\beta}$ is not the random variable. (See Section 21.3.2.) 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. Since 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 21.1 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 21.1, we have the following probabilities as a function of GPA, at the mean of TUCE:

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

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

Figure 21.2 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 GPAs than for those with low GPAs. At the sample mean of GPA of 3.117, the effect of PSI on the probability is 0.465. The simple derivative calculation of (21-9) is given in Table 21.1; the estimate is 0.468. But, of course, this calculation does not show the wide range of differences displayed in Figure 21.2.

Table 21.2 presents the estimated coefficients and marginal effects for the probit and logit models in Table 21.1. 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 (21-25) and (21-26) 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.

21.4.3 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

[12] One might be tempted in this case to suggest an asymmetric distribution for the model, such as the Weibull 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.

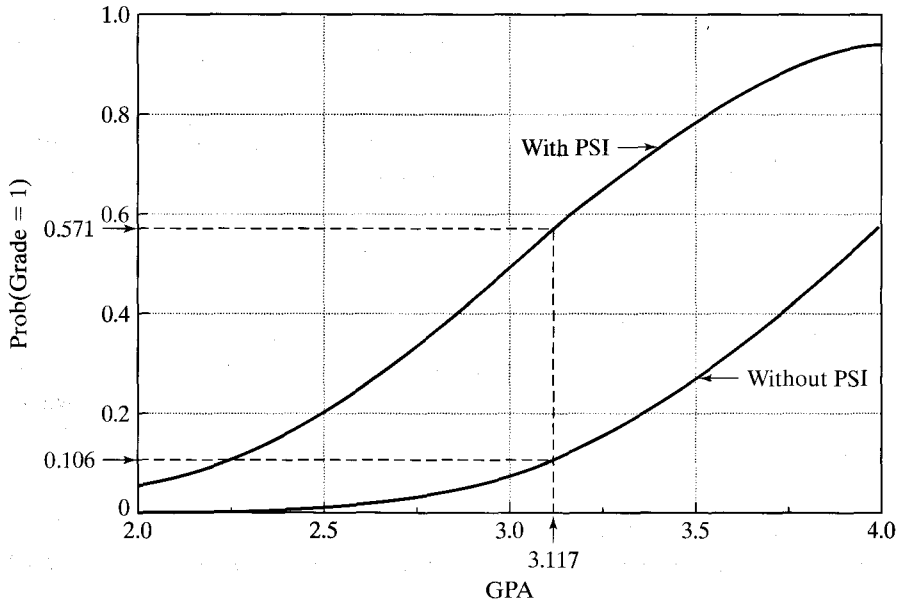


FIGURE 21.2 Effect of PSI on Predicted Probabilities.

TABLE 21.2 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		

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 $\mathbf{R} = [\mathbf{0} \mid \mathbf{I}_M]$ and $\mathbf{q} = \mathbf{0}$. Collecting terms, we find that the test statistic for this hypothesis is

$$W = \hat{\boldsymbol{\beta}}_M' \mathbf{V}_M^{-1} \hat{\boldsymbol{\beta}}_M, \tag{21-27}$$

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,

$$\ln L_0 = n[P \ln P + (1 - P) \ln(1 - P)], \tag{21-28}$$

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

$$\text{LM} = \left(\sum_{i=1}^{n} g_i \mathbf{x}_i\right)' \left[\sum_{i=1}^{n} E[-h_i]\mathbf{x}_i\mathbf{x}_i'\right]^{-1} \left(\sum_{i=1}^{n} g_i \mathbf{x}_i\right), \tag{21-29}$$

where $E[-h_i]$ is defined in (21-22) for the logit model and in (21-24) 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 17.5.3 is also convenient. For any of the models (probit, logit, Weibull, etc.), the first derivative vector can be written as

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

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

$$\text{LM} = n\left[\frac{1}{n}\mathbf{i}'(\mathbf{G}\mathbf{X})(\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{G}')\mathbf{i}\right] = nR_{\mathbf{i}}^2, \tag{21-30}$$

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 below.

21.4.4 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 above 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.**[13]

To reiterate, the Lagrange multiplier statistic is computed as follows. Let the null hypothesis, H_0, be a specification of the model, and let H_1 be the alternative. For example, H_0 might specify that only variables $\mathbf{x}_1$ appear in the model, whereas H_1 might specify that $\mathbf{x}_2$ appears in the model as well. The statistic is

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

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

[13] 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).

21.4.4.a Omitted Variables

The hypothesis to be tested is

$$\begin{aligned} H_0&: y^* = \boldsymbol{\beta}_1'\mathbf{x}_1 \qquad\quad +\varepsilon, \\ H_1&: y^* = \boldsymbol{\beta}_1'\mathbf{x}_1 + \boldsymbol{\beta}_2'\mathbf{x}_2 + \varepsilon, \end{aligned} \tag{21-31}$$

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

The statistic is then computed according to (21-29) or (21-30). It is noteworthy that in this case as in many others, the Lagrange multiplier is the coefficient of determination in a regression.

21.4.4.b Heteroscedasticity

We use the general formulation analyzed by Harvey (1976),[14]

$$\text{Var}[\varepsilon] = [\exp(\mathbf{z}'\boldsymbol{\gamma})]^2.$$[15]

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,

$$\begin{aligned} y^* &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \\ \text{Var}[\varepsilon \mid \mathbf{x}, \mathbf{z}] &= [\exp(\mathbf{z}'\boldsymbol{\gamma})]^2. \end{aligned} \tag{21-32}$$

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 example below.

The log-likelihood is

$$\ln L = \sum_{i=1}^{n}\left\{y_i \ln F\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\right) + (1 - y_i)\ln\left[1 - F\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}\right)\right]\right\}. \tag{21-33}$$

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

[15]See Section 11.7.1.

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

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\frac{f_i(y_i - F_i)}{F_i(1-F_i)}\right] \exp(-\mathbf{z}_i'\boldsymbol{\gamma})\mathbf{x}_i,$$
$$\frac{\partial \ln L}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} \left[\frac{f_i(y_i - F_i)}{F_i(1-F_i)}\right] \exp(-\mathbf{z}_i'\boldsymbol{\gamma})\mathbf{z}_i(-\mathbf{x}_i'\boldsymbol{\beta}), \tag{21-34}$$

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

$$\mathbf{w}_i = \begin{bmatrix} \mathbf{x}_i \\ (-\mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{z}_i \end{bmatrix} \tag{21-35}$$

computed at the maximum likelihood estimator, assuming that $\boldsymbol{\gamma} = \mathbf{0}$. Then (21-29) or (21-30) 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 21.4 Specification Tests in a Labor Force Participation Model

Using the data described in Example 21.1, we fit a probit model for labor force participation based on the specification

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

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 -490.84784. 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 7.4 and Example 7.6. 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 428 observations with *kids* = 1 and the 325 observations with *kids* = 0 are -347.87441 and -141.60501, respectively. The log-likelihood for the unrestricted model with separate coefficient vectors is thus the sum, -489.47942. The chi-squared statistic for testing the five restrictions of the pooled model is twice the difference, $\text{LR} = 2[-489.47942 - (-496.8663)] = 14.7738$. The 95 percent critical value from the chi-squared distribution with 5 degrees of freedom is 11.07 is 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% critical value is 15.09.)

TABLE 21.3 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)	—
Log L		−490.8478		−487.6356	
Correct Preds.		0s: 106, 1s: 357		0s: 115, 1s: 358	

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

Table 21.3 presents estimates of the probit model now 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.

21.4.4.c A Specification Test for Nonnested Models—Testing for the Distribution

Whether the logit or probit form, or some third alternative, is the best specification for a discrete choice model is a perennial question. Since the distributions are not nested within some higher level model, testing for an answer is always problematic. Building on the logic of the P_E test discussed in Section 9.4.3, Silva (2001) has suggested a score test which may be useful in this regard. The statistic is intended for a variety of discrete choice models, but is especially convenient for binary choice models which are based on a common single index formulation—the probability model is $\text{Prob}(y_i = 1 \mid \mathbf{x}_i) = F(\mathbf{x}_i'\boldsymbol{\beta})$. Let "1" denote Model 1 based on parameter vector $\boldsymbol{\beta}$ and "2" denote Model 2 with parameter vector $\boldsymbol{\gamma}$ and let Model 1 be the null specification while Model 2 is the alternative. A "super-model" which combines two alternatives would have likelihood function

$$L_\rho = \frac{[(1-\alpha)L_1(y \mid X, \beta)^\rho + \alpha L_2(y \mid X, \gamma)^\rho]^{1/\rho}}{\int_z [(1-\alpha)L_1(z \mid X, \beta)^\rho + \alpha L_2(z \mid X, \gamma)^\rho]^{1/\rho} dz}$$

(Note that integration is used generically here, since y is discrete.) The two mixing parameters are ρ and α. Silva derives an LM test in this context for the hypothesis $\alpha = 0$ for any particular value of ρ. The case when $\rho = 0$ is of particular interest. As he notes, it is the nonlinear counterpart to the Cox test we examined in Section 8.3.4. [For related results, see Pesaran and Pesaran (1993), Davidson and MacKinnon (1984, 1993).

Orme (1994), and Weeks (1996).] For binary choice models, Silva suggests the following procedure (as one of three computational strategies): Compute the parameters of the competing models by maximum likelihood and obtain predicted probabilities for $y_i = 1$, $\hat{P}_i^m$ where "i" denotes the observation and "m" = 1 or 2 for the two models.[15] The individual observations on the density for the null model, $\hat{f}_i^m$, are also required. The new variable

$$z_i(0) = \frac{\hat{P}_i^1(1-\hat{P}_i^1)}{\hat{f}_i^1} \ln\left[\frac{\hat{P}_i^1(1-\hat{P}_i^2)}{\hat{P}_i^2(1-\hat{P}_i^1)}\right]$$

is then computed. Finally, Model 1 is then reestimated with $z_i(0)$ added as an additional independent variable. A test of the hypothesis that its coefficient is zero is equivalent to a test of the null hypothesis that $\alpha = 1$, which favors Model 1. Rejection of the hypothesis favors Model 2. Silva's preferred procedure is the same as this based on

$$z_i(1) = \frac{\hat{P}_i^2 - \hat{P}_i^1}{\hat{f}_i^1}.$$

As suggested by the citations above, tests of this sort have a long history in this literature. Silva's simulation study for the Cox test ($\rho = 0$) and his score test ($\rho = 1$) suggest that the power of the test is quite erratic.

21.4.5 MEASURING GOODNESS OF FIT

There have been many fit measures suggested for QR models.[16] At a minimum, one should report the maximized value of the log-likelihood function, $\ln L$. Since 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 (21-28)], 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. 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 21.6.6). Unfortunately, the values between zero and one have no natural interpretation. If $F(\mathbf{x}_i'\boldsymbol{\beta})$ is a proper pdf, then even with many regressors the model cannot fit perfectly unless $\mathbf{x}_i'\boldsymbol{\beta}$ goes to $+\infty$ or $-\infty$. As a practical matter, it does happen. But when it does, it indicates a flaw in the model, not a good fit. If the range of one of the independent variables contains a value, say x^*, such that the sign of $(x - x^*)$ predicts y perfectly

[15] His conjecture about the computational burden is probably overstated given that modern software offers a variety of binary choice models essentially in push-button fashion.

[16] 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).

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 $\boldsymbol{\beta}$.[17] For example, one might mistakenly include as a regressor a dummy variables 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,

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

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 vary badly by the standard procedure, and this measure does not pick that point up. 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)

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

and Zavoina and McKelvey's (1975)

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

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 Example 21.4 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

$$\hat{y} = 1 \quad \text{if } \hat{F} > F^* \text{ and 0 otherwise.} \qquad \textbf{(21-36)}$$

[17] 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

$$\hat{y} = 1 \quad \text{if } P > 0.5 \text{ and 0 otherwise,} \tag{21-37}$$

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.[18] The important element to bear in mind is that the coefficients of the estimated model are not chosen so as to maximize this (or any other) fit measure, as they are in the linear regression model where **b** maximizes R^2. (The **maximum score** estimator discussed below 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 1000 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 $y = 0$.[19] In general, any prediction rule of the form in (21-36) 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 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 above are oriented more toward the relationship between the fitted probabilities and the actual values. Efron's and Cramer's statistics are 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 $y^* = \boldsymbol{\beta}'\mathbf{x} + \varepsilon$. Whether these have a close relationship to any type of fit in the familiar sense is a question that needs to be studied. In some cases,

[18] See Amemiya (1981).

[19] 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.

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 21.5 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 below 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.[20] 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.[21]

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

21.4.6 ANALYSIS OF PROPORTIONS DATA

Data for the analysis of binary responses will be in one of two forms. The data we have considered thus far are **individual;** each observation consists of $[y_i, \mathbf{x}_i]$, the actual response of an individual and associated regressor vector. **Grouped data** usually consist of counts or proportions. Grouped data are obtained by observing the response of n_i individuals, all of whom have the same $\mathbf{x}_i$. The observed dependent variable will consist of the proportion P_i of the n_i individuals ij who respond with $y_{ij} = 1$. An observation is thus $[n_i, P_i, \mathbf{x}_i]$, $i = 1, \ldots, N$. Election data are typical.[22] In the grouped data setting, it is possible to use regression methods as well as maximum likelihood procedures to analyze the relationship between P_i and $\mathbf{x}_i$. The observed P_i is an estimate of the population quantity, $\pi_i = F(\mathbf{x}_i'\boldsymbol{\beta})$. If we treat this problem as a simple one of sampling from a Bernoulli population, then, from basic statistics, we have

$$P_i = F(\boldsymbol{\beta}'\mathbf{x}_i) + \varepsilon_i = \pi_i + \varepsilon_i,$$

[20] 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.

[21] 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.

[22] The earliest work on probit modeling involved applications of grouped data in laboratory experiments. Each observation consisted of n_i subjects receiving dosage x_i of some treatment, such as an insecticide, and a proportion P_i "responding" to the treatment, usually by dying. Finney (1971) and Cox (1970) are useful and early surveys of this literature.

where

$$E[\varepsilon_i] = 0, \qquad \text{Var}[\varepsilon_i] = \frac{\pi_i(1-\pi_i)}{n_i}. \tag{21-38}$$

This heteroscedastic regression format suggests that the parameters could be estimated by a nonlinear weighted least squares regression. But there is a simpler way to proceed. Since the function $F(\mathbf{x}_i'\boldsymbol{\beta})$ is strictly monotonic, it has an inverse. (See Figure 21.1.) Consider, then, a Taylor series approximation to this function around the point $\varepsilon_i = 0$, that is, around the point $P_i = \pi_i$,

$$F^{-1}(P_i) = F^{-1}(\pi_i + \varepsilon_i) \approx F^{-1}(\pi_i) + \left[\frac{dF^{-1}(\pi_i)}{d\pi_i}\right](P_i - \pi_i).$$

But $F^{-1}(\pi_i) = \mathbf{x}_i'\boldsymbol{\beta}$ and

$$\frac{dF^{-1}(\pi_i)}{d\pi_i} = \frac{1}{F'(F^{-1}(\pi_i))} = \frac{1}{f(\pi_i)},$$

so

$$F^{-1}(P_i) \approx \mathbf{x}_i'\boldsymbol{\beta} + \frac{\varepsilon_i}{f(\pi_i)}.$$

This equation produces a heteroscedastic linear regression,

$$F^{-1}(P_i) = z_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i,$$

where

$$E[u_i \mid \mathbf{x}_i] = 0 \quad \text{and} \quad \text{Var}[u_i \mid \mathbf{x}_i] = \frac{F(\pi_i)[(1 - F(\pi_i)]}{n_i[f(\pi_i)]^2}. \tag{21-39}$$

The inverse function for the logistic model is particularly easy to obtain. If

$$\pi_i = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})},$$

then

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}.$$

This function is called the **logit** of π_i, hence the name "logit" model. For the normal distribution, the inverse function $\Phi^{-1}(\pi_i)$, called the **normit** of π_i, must be approximated. The usual approach is a ratio of polynomials.[23]

Weighted least squares regression based on (21-39) produces the **minimum chi-squared estimator (MCSE)** of $\boldsymbol{\beta}$. Since the weights are functions of the unknown parameters, a two-step procedure is called for. As always, simple least squares at the first step produces consistent but inefficient estimates. Then the estimated variances

$$w_i = \frac{\hat{\Phi}_i(1 - \hat{\Phi}_i)}{n_i\hat{\phi}_i^2}$$

[23] See Abramovitz and Stegun (1971) and Section E.5.2. The function normit +5 is called the **probit** of P_i. The term dates from the early days of this analysis, when the avoidance of negative numbers was a simplification with considerable payoff.

for the probit model or

$$w_i = \frac{1}{n_i \hat{\Lambda}_i (1 - \hat{\Lambda}_i)}$$

for the logit model based on the first-step estimates can be used for weighted least squares.[24] An iteration can then be set up,

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \left[\sum_{i=1}^{n} \frac{1}{\hat{w}_i^{(k)}} \mathbf{x}_i \mathbf{x}_i'\right]^{-1} \left[\sum_{i=1}^{n} \frac{1}{\hat{w}_i^{(k)}} \mathbf{x}_i F^{-1}\left(\hat{\pi}_i^{(k)}\right)\right]$$

where "(k)" indicates the kth iteration and "^" indicates computation of the quantity at the current (kth) estimate of $\boldsymbol{\beta}$. The MCSE has the same asymptotic properties as the maximum likelihood estimator at every step after the first, so, in fact, iteration is not necessary. Although they have the same probability limit, the MCSE is not algebraically the same as the MLE, and in a finite sample, they will differ numerically.

The log-likelihood function for a binary choice model with grouped data is

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

The likelihood equation that defines the maximum likelihood estimator is

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} n_i \left[P_i \frac{f(\mathbf{x}_i'\boldsymbol{\beta})}{F(\mathbf{x}_i'\boldsymbol{\beta})} - (1 - P_i) \frac{f(\mathbf{x}_i'\boldsymbol{\beta})}{1 - F(\mathbf{x}_i'\boldsymbol{\beta})}\right] \mathbf{x}_i = \mathbf{0}.$$

This equation closely resembles the solution for the individual data case, which makes sense if we view the grouped observation as n_i replications of an individual observation. On the other hand, it is clear on inspection that the solution to this set of equations will not be the same as the generalized (weighted) least squares estimator suggested in the previous paragraph. For convenience, define $F_i = F(\mathbf{x}_i'\boldsymbol{\beta})$, $f_i = f(\mathbf{x}_i'\boldsymbol{\beta})$, and $f_i' = [f'(z) \mid z = \mathbf{x}_i'\boldsymbol{\beta}] = [df(z)/dz] \mid z = \mathbf{x}_i'\boldsymbol{\beta}$. The Hessian of the log-likelihood is

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} n_i \left\{P_i \left[\left(\frac{f_i'}{F_i}\right) - \left(\frac{f_i}{F_i}\right)^2\right] - (1 - P_i)\left[\left(\frac{f_i'}{1 - F_i}\right) + \left(\frac{f_i}{(1 - F_i)}\right)^2\right]\right\} \mathbf{x}_i \mathbf{x}_i'.$$

To evaluate the expectation of the Hessian, we need only insert the expectation of the only stochastic element, P_i, which is $E[P_i \mid \mathbf{x}_i] = F_i$. Then

$$E\left[\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'}\right] = \sum_{i=1}^{n} n_i \left[f_i' - \frac{f_i^2}{F_i} - f_i' - \frac{f_i^2}{1 - F_i}\right] \mathbf{x}_i \mathbf{x}_i' = -\sum_{i=1}^{n} \left[\frac{n_i f_i^2}{F_i (1 - F_i)}\right] \mathbf{x}_i \mathbf{x}_i'.$$

The asymptotic covariance matrix for the maximum likelihood estimator is the negative inverse of this matrix. From (21-39), we see that it is exactly equal to

$$\text{Asy. Var[minimum } \chi^2 \text{ estimator]} = [\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}$$

[24] Simply using p_i and $f[F^{-1}(P_i)]$ might seem to be a simple expedient in computing the weights. But this method would be analogous to using y_i^2 instead of an estimate of σ_i^2 in a heteroscedastic regression. Fitted probabilities and, for the probit model, densities should be based on a consistent estimator of the parameters.

since the diagonal elements of $\mathbf{\Omega}^{-1}$ are precisely the values in brackets in the expression for the expected Hessian above. We conclude that although the MCSE and the MLE for this model are numerically different, they have the same asymptotic properties, consistent and asymptotically normal (the MCS estimator by virtue of the results of Chapter 10, the MLE by those in Chapter 17), and with asymptotic covariance matrix as previously given.

There is a complication in using the MCS estimator. The FGLS estimator breaks down if any of the sample proportions equals one or zero. A number of ad hoc patches have been suggested; the one that seems to be most widely used is to add or subtract a small constant, say 0.001, to or from the observed proportion when it is zero or one. The familiar results in (21-38) also suggest that when the proportion is based on a large population, the variance of the estimator can be exceedingly low. This issue will resurface in surprisingly low standard errors and high t ratios in the weighted regression. Unfortunately, that is a consequence of the model.[25] The same result will emerge in maximum likelihood estimation with grouped data.

21.5 EXTENSIONS OF THE BINARY CHOICE MODEL

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.

21.5.1 RANDOM AND FIXED EFFECTS MODELS FOR PANEL DATA

The availability of high quality panel data sets on microeconomic behavior has maintained an interest in extending the models of Chapter 13 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

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

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 which 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

[25] Whether the proportion should, in fact, be considered as a single observation from a distribution of proportions is a question that arises in all these cases. It is unambiguous in the bioassay cases noted earlier. But the issue is less clear with election data, especially since in these cases, the n_i will represent most of if not all the potential respondents in location i rather than a random sample of respondents.

joint probabilities from a T_i variate distribution, which is generally problematic.[26] (We will return to this issue below.) A more promising approach is an effects model,

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

where, as before (see Section 13.4), u_i is the unobserved, individual specific heterogeneity. Once again, we distinguish between "random" and "fixed" effects models by the relationship between u_i and $\mathbf{x}_{it}$. The assumption that u_i is unrelated to $\mathbf{x}_{it}$, so that the conditional distribution $f(u_i \mid \mathbf{x}_{it})$ is not dependent on $\mathbf{x}_{it}$, produces the **random effects model.** Note that this places a restriction on the distribution of the heterogeneity. If that distribution is unrestricted, so that u_i and $\mathbf{x}_{it}$ may be correlated, then we have what is called the **fixed effects model.** The distinction does not relate to any intrinsic characteristic of the effect, itself.

As we shall see shortly, this is a modeling framework that is fraught with difficulties and unconventional estimation problems. Among them are: 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.

We begin with the random effects specification, then consider fixed effects and some **semiparametric** approaches that do not require the distinction. We conclude with a brief look at dynamic models of **state dependence.**[27]

21.5.1.a Random Effects Models

A specification which has the same structure as the random effects model of Section 13.4, 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

$$E[v_{it} \mid \mathbf{X}] = 0; \text{Cov}[v_{it}, v_{js} \mid \mathbf{X}] = \text{Var}[v_{it} \mid \mathbf{X}] = 1 \quad \text{if } i = j \text{ and } t = s; \text{ 0 otherwise}$$
$$E[u_i \mid \mathbf{X}] = 0; \text{Cov}[u_i, u_j \mid \mathbf{X}] = \text{Var}[u_i \mid \mathbf{X}] = \sigma_u^2 \quad \text{if } i = j; \text{ 0 otherwise}$$
$$\text{Cov}[v_{it}, u_j \mid \mathbf{X}] = 0 \text{ for all } i, t, j$$

[26] 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 E.5.6), 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 (1980) and Stata (2001)] 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 which takes only two values. We should not be too optimistic about this if T_i is even moderately large.

[27] A survey of some of these results is given by Hsiao (1996). Most of Hsiao (1996) 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).

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

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

and

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

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

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

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

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

$$L_i = P(y_{i1}, \ldots, y_{iT_i} \mid \mathbf{X}) = \int_{L_{iT_i}}^{U_{iT_i}} \cdots \int_{L_{i1}}^{U_{i1}} f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i}) d\varepsilon_{i1} d\varepsilon_{i2} \ldots d\varepsilon_{iT_i}. \quad \textbf{(21-40)}$$

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

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

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_i's are independent, so

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

Inserting this result in (21-40) produces

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

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

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

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

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

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

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

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

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, Weibull, and so on. The intricate part remaining is to determine how to do the outer integration. **Butler and Moffitt's method** assuming that u_i is normally distributed is fairly straightforward, so we will consider it first. We will then consider some other possibilities. For the probit model, the individual probabilities inside the product would be $\Phi[q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)]$, where $\Phi[.]$ is the standard normal CDF and $q_{it} = 2y_{it} - 1$. For the logit model, $\Phi[.]$ would be replaced with the logistic probability, $\Lambda[.]$. For the present, treat the entire function as a function of u_i, $g(u_i)$. The integral is, then

$$L_i = \int_{-\infty}^{\infty} \frac{1}{\sigma_u\sqrt{2\pi}} e^{-\frac{u_i^2}{2\sigma_u^2}} g(u_i)\, du_i.$$

Let $r_i = u_i/(\sigma_u\sqrt{2})$. Then, $u_i = (\sigma_u\sqrt{2})r_i = \theta r_i$ and $du_i = \theta dr_i$. Making the change of variable produces

$$L_i = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-r_i^2} g(\theta r_i)\, dr_i.$$

(Several constants cancel out of the fractions.) Returning to our probit (or logit model), we now have

$$L_i = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-r_i^2} \left[\prod_{t=1}^{T_i} \Phi(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \theta r_i)) \right] dr_i.$$

The payoff to all this manipulation is that this likelihood function involves only one-dimensional integrals. The inner integrals are the CDF of the standard normal distribution or the logistic or extreme value distributions, which are simple to obtain. The function is amenable to Gauss–Hermite **quadrature** for computation. (Gauss–Hermite quadrature is discussed in Section E.5.4.) Assembling all the pieces, we obtain the approximation to the log-likelihood;

$$\ln L_H = \sum_{i=1}^{n} \left\{ \ln \left[\frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} \prod_{t=1}^{T_i} w_h \Phi(q_{it}(\mathbf{x}_{it}'\boldsymbol{\beta} + \theta z_h)) \right] \right\}$$

where H is the number of points for the quadrature, and w_h and z_h are the weights and nodes for the quadrature. Maximizing this function remains a complex problem. But, it is made quite feasible by the transformations which reduce the integration to one dimension. This technique for the probit model has been incorporated in most contemporary econometric software and can be easily extended to other models.

The first and second derivatives are likewise complex but still computable by quadrature. An estimate of σ_u is obtained from the result $\sigma_u = \theta/\sqrt{2}$ and a standard error can be obtained by dividing that for $\hat{\theta}$ by $\sqrt{2}$. The model may be adapted to the logit or any other formulation just by changing the CDF in the preceding equation from $\Phi[.]$ to the logistic CDF, $\Lambda[.]$ or the other appropriate CDF.

The hypothesis of no cross-period correlation can be tested, in principle, using any of the three classical testing procedures we have discussed to examine the statistical significance of the estimated σ_u.

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. A recent 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 (1995a, pp. 425–431).]

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 which allows some flexibility is the method of **maximum simulated likelihood** (MSL) which was discussed in Section 17.8. The transformed likelihood we derived above is an expectation;

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

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

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

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

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

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

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 applications in the following Examples 16.5, 17.10, and 21.6 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;

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

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 $\mathbf{u}$. Since the draws are uncorrelated—$\boldsymbol{\Gamma}$ is diagonal—this is essentially the same estimation problem as the random effects model considered previously. This model is estimated in Example 17.10. Example 16.5 presents a similar model that assumes that the distribution of $\boldsymbol{\beta}_i$ is discrete rather than continuous.

21.5.1.b Fixed Effects Models

The fixed effects model is

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

where d_{it} is a dummy variable which 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

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

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 to symmetric distributions such as the normal and logistic, so that the probability can be conveniently written as $\text{Prob}(Y_{it} = y_{it} \mid \mathbf{x}_{it}) = P[q_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]$. It will be convenient to let $z_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}$ so $\text{Prob}(Y_{it} = y_{it} \mid \mathbf{x}_{it}) = P(q_{it} z_{it})$.

In our previous application of this model, in the linear regression case, we found that estimation of the parameters was made possible by a transformation of the data to deviations from group means which eliminated the person specific constants from the estimator. (See Section 13.3.2.) Save for the special case discussed below, that will not be possible here, so that if one desires to estimate the parameters of this model, it will be necessary actually to compute the possibly huge number of constant terms at the same time. This has been widely viewed as a practical obstacle to estimation of this model because of the need to invert a potentially large second derivatives matrix, but this is a misconception. [See, e.g., Maddala (1987), p. 317.] The likelihood equations for this model are

$$\frac{\partial \ln L}{\partial \alpha_i} = \sum_{t=1}^{T_i} \frac{q_{it} f(q_{it} z_{it})}{P(q_{it} z_{it})} = \sum_{t=1}^{T_i} g_{it} = g_{ii} = 0$$

and

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \frac{q_{it} f(q_{it} z_{it})}{P(q_{it} z_{it})} \mathbf{x}_{it} = \sum_{t=1}^{T_i} g_{it} \mathbf{x}_{it} = \mathbf{0}$$

where $f(.)$ is the density that corresponds to $P(.)$. For our two familiar models, $g_{it} = q_{it}\phi(q_{it} z_{it})/\Phi(q_{it} z_{it})$ for the normal and $q_{it}[1 - \Lambda(q_{it} z_{it})]$ for the logistic. Note that for these distributions, g_{it} is always negative when y_{it} is zero and always positive when y_{it} equals one. (The use of q_{it} as in the preceding assumes the distribution is symmetric. For asymmetric distributions such as the Weibull, g_{it} and h_{it} would be more complicated,

but the central results would be the same.) The second derivatives matrix is

$$\frac{\partial^2 \ln L}{\partial \alpha_i^2} = \sum_{t=1}^{T_i} \left[\frac{f'(q_{it} z_{it})}{P(q_{it} z_{it})} - \left(\frac{f(q_{it} z_{it})}{P(q_{it} z_{it})} \right)^2 \right] = \sum_{t=1}^{T_i} h_{it} = h_{ii} < 0,$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \alpha_i} = \sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it}$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} \sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it} \mathbf{x}_{it}' = \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}'}, \text{ a negative semidefinite matrix.}$$

Note that the leading q_{it} falls out of the second derivatives since in each appearance, since $q_{it}^2 = 1$. The derivatives of the densities with respect to their arguments are $-(q_{it} z_{it}) \phi(q_{it} z_{it})$ for the normal distribution and $[1 - 2\Lambda(q_{it} z_{it})] f(q_{it} z_{it})$ for the logistic. In both cases, h_{it} is negative for all values of $q_{it} z_{it}$. The likelihood equations are a large system, but the solution turns out to be surprisingly straightforward. [See Greene (2001).]

By using the formula for the partitioned inverse, we find that the $K \times K$ submatrix of the inverse of the Hessian that corresponds to $\boldsymbol{\beta}$, which would provide the asymptotic covariance matrix for the MLE, is

$$\mathbf{H}^{\boldsymbol{\beta}\boldsymbol{\beta}'} = \left\{ \sum_{i=1}^{n} \left[\sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it} \mathbf{x}_{it}' - \frac{1}{h_{ii}} \left(\sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it} \right) \left(\sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it}' \right) \right] \right\}^{-1}$$

$$= \left\{ \sum_{i=1}^{n} \left[\sum_{t=1}^{T_i} h_{it} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \right] \right\}^{-1} \quad \text{where } \bar{\mathbf{x}}_i = \frac{\sum_{t=1}^{T_i} h_{it} \mathbf{x}_{it}}{h_{ii}}.$$

Note the striking similarity to the result we had for the fixed effects model in the linear case. By assembling the Hessian as a partitioned matrix for $\boldsymbol{\beta}$ and the full vector of constant terms, then using (A-66b) and the definitions above to isolate one diagonal element, we find

$$\mathbf{H}^{\alpha_i \alpha_i} = \frac{1}{h_{ii}} + \bar{\mathbf{x}}_i' \mathbf{H}^{\boldsymbol{\beta}\boldsymbol{\beta}'} \bar{\mathbf{x}}_i$$

Once again, the result has the same format as its counterpart in the linear model. In principle, the negatives of these would be the estimators of the asymptotic variances of the maximum likelihood estimators. (Asymptotic properties in this model are problematic, as we consider below.)

All of these can be computed quite easily once the parameter estimates are in hand, so that in fact, practical estimation of the model is not really the obstacle. (This must be qualified, however. Looking at the likelihood equation for a constant term, it is clear that if y_{it} is the same in every period then there is no solution. For example, if $y_{it} = 1$ in every period, then $\partial \ln L / \partial \alpha_i$ must be positive, so it cannot be equated to zero with finite coefficients. Such groups would have to be removed from the sample in order to fit this model.) It is shown in Greene (2001) in spite of the potentially large number of parameters in the model, Newton's method can be used with the following iteration

which uses only the $K \times K$ matrix computed above and a few $K \times 1$ vectors:

$$\hat{\boldsymbol{\beta}}^{(s+1)} = \hat{\boldsymbol{\beta}}^{(s)} - \left\{ \sum_{i=1}^{n} \left[\sum_{t=1}^{T_i} h_{it} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \right] \right\}^{-1} \left\{ \sum_{i=1}^{n} \left[\sum_{t=1}^{T_i} g_{it} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i) \right] \right\}$$

$$= \hat{\boldsymbol{\beta}}^{(s)} + \boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(s)}$$

and

$$\hat{\alpha}_i^{(s+1)} = \hat{\alpha}_i^{(s)} - \left[(g_{ii}/h_{ii}) + \bar{\mathbf{x}}_i' \boldsymbol{\Delta}_{\boldsymbol{\beta}}^{(s)} \right].^{29}$$

This is a large amount of computation involving many summations, but it is linear in the number of parameters and does not involve any $n \times n$ matrices.

The problems with the fixed effects estimator are statistical, not practical.[30] The estimator relies on T_i increasing for the constant terms to be consistent—in essence, each α_i is estimated with T_i observations. But, in this setting, not only is T_i fixed, it is likely to be quite small. As such, the estimators of the constant terms are not consistent (not because they converge to something other than what they are trying to estimate, but because they do not converge at all). The estimator of $\boldsymbol{\beta}$ is a function of the estimators of α, which means that the MLE of $\boldsymbol{\beta}$ is not consistent either. This is the **incidental parameters problem.** [See Neyman and Scott (1948) and Lancaster (2000).] There is, as well, a small sample (small T_i) bias in the estimators. How serious this bias is remains a question in the literature. Two pieces of received wisdom are Hsiao's (1986) results for a binary logit model and Heckman and MaCurdy's (1980) results for the probit model. Hsiao found that for $T_i = 2$, the bias in the MLE of $\boldsymbol{\beta}$ is 100 percent, which is extremely pessimistic. Heckman and MaCurdy found in a Monte Carlo study that in samples of $n = 100$ and $T = 8$, the bias appeared to be on the order of 10 percent, which is substantive, but certainly less severe than Hsiao's results suggest. The fixed effects approach does have some appeal in that it does not require an assumption of orthogonality of the independent variables and the heterogeneity. An ongoing pursuit in the literature is concerned with the severity of the tradeoff of this virtue against the incidental parameters problem. Some commentary on this issue appears in Arellano (2001).

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

[29]Similar results appear in Prentice and Gloeckler (1978) who attribute it to Rao (1973), and Chamberlain (1983).

[30]See Vytlacil, Aakvik and Heckman (2002), Chamberlain (1980, 1984), Newey (1994), Bover and Arellano (1997) and Chen (1998) for some extensions of parametric forms of the binary choice models with fixed effects.

A fixed effects binary logit model is

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

The unconditional likelihood for the nT independent observations is

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

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

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

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

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

$$= \frac{\exp\left(\sum_{t=1}^{T_i} y_{it} \mathbf{x}_{it}'\boldsymbol{\beta}\right)}{\sum_{\Sigma_t d_{it} = S_i} \exp\left(\sum_{t=1}^{T_i} d_{it} \mathbf{x}_{it}'\boldsymbol{\beta}\right)}.$$

The function in the denominator is summed over the set of all $\binom{T_i}{S_i}$ different sequences of T_i zeros and ones that have the same sum as $S_i = \sum_{t=1}^{T_i} y_{it}$.[31]

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

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

For each pair of observations, we have these possibilities:

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

The ith term in L^c for either of these is just one, so they contribute nothing to the conditional likelihood function.[32] When we take logs, these terms (and these observations) will drop out. But suppose that $y_{i1} = 0$ and $y_{i2} = 1$. Then

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

[31] The enumeration of all these computations stands to be quite a burden—see Arellano (2000, p. 47) or Baltagi (1995, p. 180) who [citing Greene (1993)] suggests that $T_i > 10$ would be excessive. In fact, using a recursion suggested by Krailo and Pike (1984), the computation even with T_i up to 100 is routine.

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

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

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

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

As in the linear regression model, it is of some interest to test whether there is indeed heterogeneity. With homogeneity ($\alpha_i = \alpha$), there is no unusual problem, and the model can be estimated, as usual, as a logit model. It is not possible to test the hypothesis using the likelihood ratio test, however, because the two likelihoods are not comparable. (The conditional likelihood is based on a restricted data set.) None of the usual tests of restrictions can be used because the individual effects are never actually estimated.[33] Hausman's (1978) specification test is a natural one to use here, however. Under the null hypothesis of homogeneity, both Chamberlain's conditional maximum likelihood estimator (CMLE) and the usual maximum likelihood estimator are consistent, but Chamberlain's is inefficient. (It fails to use the information that $\alpha_i = \alpha$, and it may not use all the data.) Under the alternative hypothesis, the unconditional maximum likelihood estimator is inconsistent,[34] whereas Chamberlain's estimator is consistent and efficient. The Hausman test can be based on the chi-squared statistic

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

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

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

[34] Hsaio (1996) derives the result explicitly for some particular cases.

Example 21.6 Individual Effects in a Binary Choice Model

To illustrate the fixed and random effects estimators, we continue the analyses of Examples 16.5 and 17.10.[35] The binary dependent variable is

$$y_{it} = 1 \text{ if firm } i \text{ realized a product innovation in year } t \text{ and 0 if not.}$$

The sample consists of 1,270 German firms observed for 5 years, 1984–1988. Independent variables in the model that we formulated were

x_{it1} = constant,

x_{it2} = log of sales,

x_{it3} = relative size = ratio of employment in business unit to employment in the industry,

x_{it4} = ratio of industry imports to (industry sales + imports),

x_{it5} = ratio of industry foreign direct investment to (industry sales + imports),

x_{it6} = productivity = ratio of industry value added to industry industry employment,

Latent class and **random parameters models** were fit to these data in Examples 16.5 and 17.10. (For this example, we have dropped the two sector dummy variables as they are constant across periods. This precludes estimation of the fixed effects models.) Table 21.4 presents estimates of the probit and logit models with individual effects. The differences across the models are quite large. Note, for example, that the signs of the sales and FDI variables, both of which are highly significant in the base case, change sign in the fixed effects model. (The random effects logit model is estimated by appending a normally distributed individual effect to the model and using the Butler and Moffitt method described earlier.)

The evidence of heterogeneity in the data is quite substantial. The simple likelihood ratio tests of either panel data form against the base case leads to rejection of the restricted model. (The fixed effects logit model cannot be used for this test because it is based on the conditional log likelihood whereas the other two forms are based on unconditional likelihoods. It was not possible to fit the logit model with the full set of fixed effects. The relative size variable has some, but not enough within group variation, and the model became unstable after only a few iterations.) The Hausman statistic based on the logit estimates equals 19.59. The 95 percent critical value from the chi-squared distribution with 5 degrees of freedom is 11.07, so based on the logit estimates, we would reject the homogeneity restriction. In this setting, unlike in the linear model (see Section 13.4.4), neither the probit nor the logit model provides a means of testing for whether the random or fixed effects model is preferred.

21.5.2 SEMIPARAMETRIC 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,

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

$$\hat{\boldsymbol{\beta}}_{\text{LP}} \approx 0.25\boldsymbol{\beta}_{\text{logit}} \text{ for the slopes.}^{36}$$

[35] The data are from by Bertschek and Lechner (1998). Description of the data appears in Example 16.5 and in the original paper.

[36] An additional 0.5 is added for the constant term in both models.

TABLE 21.4 Estimated Panel Data Models. (Standard Errors in Parentheses; Marginal Effects in Brackets.)

	Probit			*Logit*		
	Base	***Random***	***Fixed***	***Base***	***Random***	***Fixed***
Constant	−2.35 (0.214)	−3.51 (0.502)	—	−3.83 (0.351)	−0.751 (0.611)	—
InSales	0.243 (0.194) [0.094]	0.353 (0.448) [0.088]	−0.650 (0.355) [−0.255]	0.408 (0.0323) [0.097]	0.429 (0.547) [0.103]	−0.863 (0.530)
RelSize	1.17 (0.141) [0.450]	1.59 (0.241) [0.398]	0.278 (0.734) [0.110]	2.16 (0.272) [0.517]	1.36 (0.296) [0.328]	0.340 (1.06)
Imports	0.909 (0.143) [0.350]	1.40 (0.343) [0.351]	3.50 (2.92) [1.38]	1.49 (0.232) [0.356]	0.858 (0.418) [0.207]	4.69 (4.34)
FDI	3.39 (0.394) [1.31]	4.55 (0.828) [1.14]	−8.13 (3.38) [−3.20]	5.75 (0.705) [1.37]	1.98 (1.01) [0.477]	−10.44 (5.01)
Prod	−4.71 (0.553) [−1.82]	−5.62 (0.753) [−1.41]	5.30 (4.03) [2.09]	−9.33 (1.13) [−2.29]	−1.76 (0.927) [−0.424]	6.64 (5.93)
ρ	—	0.582 (0.019)	—		0.252 (0.081)	—
Ln L	−4134.86	−3546.01	−2086.26	−4128.98	−3545.84	−1388.51

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.[37] 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.[38] The approximation would seem to suggest that with this type of misspecification, we would once again obtain a scaled version of the correct coefficient vector. (Amemiya also reports the widely observed relationship $\hat{\boldsymbol{\beta}}_{\text{logit}} = 1.6\hat{\boldsymbol{\beta}}_{\text{probit}}$, which follows from the results above.)

Greene (1983), building on Goldberger (1981), finds that if the probit model is correctly specified and if the regressors are themselves joint normally distributed, then the probability limit of the least squares estimator is a multiple of the true coefficient

[37]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.

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

vector.[39] Greene's result is useful only for the same purpose as Amemiya's quick correction of OLS. Multivariate normality is obviously inconsistent with most applications. For example, nearly all applications include at least one dummy variable. Ruud (1982) and Cheung and Goldberger (1984), however, have shown that much weaker conditions than joint normality will produce the same proportionality result. For a probit model, Cheung and Goldberger require only that $E[\mathbf{x} \mid y^*]$ be linear in y^*. Several authors have built on these observations to pursue the issue of what circumstances will lead to proportionality results such as these. Ruud (1986) and Stoker (1986) have extended them to a very wide class of models that goes well beyond those of Cheung and Goldberger. Curiously enough, Stoker's results rule out dummy variables, but it is those for which the proportionality result seems to be most robust.[40]

21.5.3 THE MAXIMUM SCORE ESTIMATOR (MSCORE)

In Section 21.4.5, we discussed the issue of prediction rules for the probit and logit models. In contrast to the linear regression model, estimation of these binary choice models is not based on a fitting rule, such as the sum of squared residuals, which is related to the fit of the model to the data. The maximum score estimator *is* based on a fitting rule,

$$\text{Maximize}_{\boldsymbol{\beta}} S_{n\alpha}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}[z_i - (1-2\alpha)]\text{sgn}(\mathbf{x}_i'\boldsymbol{\beta}).^{41}$$

The parameter α is a preset quantile, and $z_i = 2y_i - 1$. (So $z = -1$ if $y = 0$.) If α is set to $\frac{1}{2}$, then the maximum score estimator chooses the $\boldsymbol{\beta}$ to maximize the number of times that the prediction has the same sign as z. This result matches our prediction rule in (21-36) with $F^* = 0.5$. So for $\alpha = 0.5$, maximum score attempts to maximize the number of correct predictions. Since the sign of $\mathbf{x}'\boldsymbol{\beta}$ is the same for all positive multiples of $\boldsymbol{\beta}$, the estimator is computed subject to the constraint that $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$.

Since there is no log-likelihood function underlying the fitting criterion, there is no information matrix to provide a method of obtaining standard errors for the estimates. **Bootstrapping** can used to provide at least some idea of the sampling variability of the estimator. (See Section E.4.) The method proceeds as follows. After the set of coefficients $\mathbf{b}_n$ is computed, R randomly drawn samples of m observations are drawn from the original data set *with replacement*. The bootstrap sample size m may be less than or equal to n, the sample size. With each such sample, the maximum score estimator is recomputed, giving $\mathbf{b}_m(r)$. Then the **mean-squared deviation matrix**

$$\mathbf{MSD}(\mathbf{b}) = \frac{1}{R}\sum_{b=1}^{R}[\mathbf{b}_m(r) - \mathbf{b}_n][\mathbf{b}_m(r) - \mathbf{b}_n]'$$

[39] The scale factor is estimable with the sample data, so under these assumptions, a method of moments estimator is available.

[40] See Greene (1983).

[41] See Manski (1975, 1985, 1986) and Manski and Thompson (1986). For extensions of this model, see Horowitz (1992), Charlier, Melenberg and van Soest (1995), Kyriazidou (1997) and Lee (1996).

TABLE 21.5 Maximum Score Estimator

	Maximum Score		*Probit*	
	Estimate	*Mean Square Dev.*	*Estimate*	*Standard Error*
Constant β_1	−0.9317	0.1066	−7.4522	2.5420
GPA β_2	0.3582	0.2152	1.6260	0.6939
TUCE β_3	−0.01513	0.02800	0.05173	0.08389
PSI β_4	0.05902	0.2749	1.4264	0.5950
		Fitted		Fitted
		0 1		0 1
	Actual	0 21 0	Actual	0 18 3
		1 4 7		1 3 8

is computed. The authors of the technique emphasize that this matrix is not a covariance matrix.[42]

Example 21.7 The Maximum Score Estimator

Table 21.5 presents maximum score estimates for Spector and Mazzeo's GRADE model using $\alpha = 0.5$. Note that they are quite far removed from the probit estimates. (The estimates are extremely sensitive to the choice of α.) Of course, there is no meaningful comparison of the coefficients, since the maximum score estimates are not the slopes of a conditional mean function. The prediction performance of the model is also quite sensitive to α, but that is to be expected.[43] As expected, the maximum score estimator performs better than the probit estimator. The score is precisely the number of correct predictions in the 2×2 table, so the best that the probit model could possibly do is obtain the "maximum score." In this example, it does not quite attain that maximum. [The literature awaits a comparison of the prediction performance of the probit/logit (parametric) approaches and this semiparametric model.] The relevant scores for the two estimators are also given in the table.

Semiparametric approaches such as this one have the virtue that they do not make a possibly erroneous assumption about the underlying distribution. On the other hand, as seen in the example, there is no guarantee that the estimator will outperform the fully parametric estimator. One additional practical consideration is that semiparametric estimators such as this one are very computation intensive. At present, the maximum score estimator is not usable for more than roughly 15 coefficients and perhaps 1,500 to 2,000 observations.[44] A third shortcoming of the approach is, unfortunately, inherent in

[42]Note that we are not yet agreed that $\mathbf{b}_n$ even converges to a meaningful vector, since no underlying probability distribution as such has been assumed. Once it is agreed that there is an underlying regression function at work, then a meaningful set of asymptotic results, including consistency, can be developed. Manski and Thompson (1986) and Kim and Pollard (1990) present a number of results. Even so, it has been shown that the bootstrap MSD matrix is useful for little more than descriptive purposes. Horowitz's (1993) smoothed maximum score estimator replaces the discontinuous sgn $(\boldsymbol{\beta}'\mathbf{x}_i)$ in the MSCORE criterion with a continuous weighting function, $\Phi(\boldsymbol{\beta}'\mathbf{x}_i/h)$, where h is a bandwidth proportional to $n^{-1/5}$. He argues that this estimator is an improvement over Manski's MSCORE estimator. ("Its asymptotic distribution is very complicated and not useful for making inferences in applications." Later in the same paragraph he argues, "There has been no theoretical investigation of the properties of the bootstrap in maximum score estimation.")

[43]The criterion function for choosing $\mathbf{b}$ is not continuous, and it has more than one optimum. M. E. Bissey reported finding that the score function varies significantly between the local optima as well. [Personal correspondence to the author, University of York (1995).]

[44]Communication from C. Manski to the author. The maximum score estimator has been implemented by Manski and Thompson (1986) and Greene (1995a).

its design. The parametric assumptions of the probit or logit produce a large amount of information about the relationship between the response variable and the covariates. In the final analysis, the marginal effects discussed earlier might well have been the primary objective of the study. That information is lost here.

21.5.4 SEMIPARAMETRIC ESTIMATION

The fully parametric probit and logit models remain by far the mainstays of empirical research on binary choice. Fully nonparametric discrete choice models are fairly exotic and have made only limited inroads in the literature, and much of that literature is theoretical [e.g., Matzkin (1993)]. The primary obstacle to application is their paucity of interpretable results. (See Example 21.9.) Of course, one could argue on this basis that the firm results produced by the fully parametric models are merely fragile artifacts of the detailed specification, not genuine reflections of some underlying truth. [In this connection, see Manski (1995).] But that orthodox view raises the question of what motivates the study to begin with and what one hopes to learn by embarking upon it. The intent of model building to approximate reality so as to draw useful conclusions is hardly limited to the analysis of binary choices. Semiparametric estimators represent a middle ground between these extreme views.[45] The single index model of Klein and Spady (1993) has been used in several applications, including Gerfin (1996), Horowitz (1993), and Fernandez and Rodriguez-Poo (1997).[46]

The single index formulation departs from a linear "regression" formulation,

$$E[y_i \mid \mathbf{x}_i] = E[y_i \mid \mathbf{x}_i'\boldsymbol{\beta}].$$

Then

$$\text{Prob}(y_i = 1 \mid \mathbf{x}_i) = F(\mathbf{x}_i'\boldsymbol{\beta} \mid \mathbf{x}_i) = G(\mathbf{x}_i'\boldsymbol{\beta}),$$

where G is an unknown continuous distribution function whose range is [0, 1]. The function G is not specified a priori; it is estimated along with the parameters. (Since G as well as $\boldsymbol{\beta}$ is to be estimated, a constant term is not identified; essentially, G provides the location for the index that would otherwise be provided by a constant.) The criterion function for estimation, in which subscripts n denote estimators of their unsubscripted counterparts, is

$$\ln L_n = \frac{1}{n}\sum_{i=1}^{n}\{y_i \ln G_n(\mathbf{x}_i'\boldsymbol{\beta}_n) + (1 - y_i)\ln[1 - G_n(\mathbf{x}_i'\boldsymbol{\beta}_n)]\}.$$

The estimator of the probability function, G_n, is computed at each iteration using a nonparametric kernel estimator of the density of $\mathbf{x}'\boldsymbol{\beta}_n$; we did this calculation in Section 16.4. For the Klein and Spady estimator, the nonparametric regression

[45] Recent proposals for semiparametric estimators in addition to the one developed here include Lewbel (1997, 2000), Lewbel and Honore (2001), and Altonji and Matzkin (2001). In spite of nearly 10 years of development, this is a nascent literature. The theoretical development tends to focus on root-n consistent coefficient estimation in models which provide no means of computation of probabilities or marginal effects.

[46] A symposium on the subject is Hardle and Manski (1993).

estimator is

$$G_n(z_i) = \frac{\bar{y} g_n(z_i \mid y_i = 1)}{\bar{y} g_n(z_i \mid y_i = 1) + (1 - \bar{y}) g_n(z_i \mid y_i = 0)},$$

where $g_n(z_i \mid y_i)$ is the **kernel estimate of the density** of $z_i = \boldsymbol{\beta}_n' \mathbf{x}_i$. This result is

$$g_n(z_i \mid y_i = 1) = \frac{1}{n \bar{y} h_n} \sum_{j=1}^{n} y_j K\left(\frac{z_i - \boldsymbol{\beta}_n' \mathbf{x}_j}{h_n}\right);$$

$g_n(z_i \mid y_i = 0)$ is obtained by replacing $\bar{y}$ with $1 - \bar{y}$ in the leading scalar and y_j with $1 - y_j$ in the summation. As before, h_n is the bandwidth. There is no firm theory for choosing the kernel function or the bandwidth. Both Horowitz and Gerfin used the standard normal density. Two different methods for choosing the bandwidth are suggested by them.[47] Klein and Spady provide theoretical background for computing asymptotic standard errors.

Example 21.8 A Comparison of Binary Choice Estimators

Gerfin (1996) did an extensive analysis of several binary choice estimators, the probit model, Klein and Spady's single index model, and Horowitz's smoothed maximum score estimator. (A fourth "seminonparametric" estimator was also examined, but in the interest of brevity, we confine our attention to the three more widely used procedures.) The several models were all fit to two data sets on labor force participation of married women, one from Switzerland and one from Germany. Variables included in the equation were (our notation), $x_1 =$ a constant, $x_2 =$ age, $x_3 =$ age^2, $x_4 =$ education, $x_5 =$ number of young children, $x_6 =$ number of older children, $x_7 =$ log of yearly nonlabor income, and $x_8 =$ a dummy variable for permanent foreign resident (Swiss data only). Coefficient estimates for the models are not directly comparable. We suggested in Example 21.3 that they could be made comparable by transforming them to marginal effects. Neither MSCORE nor the single index model, however, produces a marginal effect (which does suggest a question of interpretation). The author obtained comparability by dividing all coefficients by the absolute value of the coefficient on x_7. The set of normalized coefficients estimated for the Swiss data appears in Table 21.6, with estimated standard errors (from Gerfin's Table III) shown in parentheses.

Given the very large differences in the models, the agreement of the estimates is impressive. [A similar comparison of the same estimators with comparable concordance may be found in Horowitz (1993, p. 56).] In every case, the standard error of the probit estimator is smaller than that of the others. It is tempting to conclude that it is a more efficient estimator, but that is true only if the normal distribution assumed for the model is correct. In any event, the smaller standard error is the payoff to the sharper specification of the distribution. This payoff could be viewed in much the same way that parametric restrictions in the classical regression make the asymptotic covariance matrix of the restricted least squares estimator smaller than its unrestricted counterpart, even if the restrictions are incorrect.

Gerfin then produced plots of $F(z)$ for z in the range of the sample values of $\mathbf{b}'\mathbf{x}$. Once again, the functions are surprisingly close. In the German data, however, the Klein–Spady estimator is nonmonotonic over a sizeable range, which would cause some difficult problems of interpretation. The maximum score estimator does not produce an estimate of the probability, so it is excluded from this comparison. Another comparison is based on the predictions of the observed response. Two approaches are tried, first counting the number of cases in which the predicted probability exceeds 0.5. ($\mathbf{b}'\mathbf{x} > 0$ for MSCORE) and second by summing the sample values of $F(\mathbf{b}'\mathbf{x})$. (Once again, MSCORE is excluded.) By the second approach,

[47] The function $G_n(z)$ involves an enormous amount of computation, on the order of n^2, in principle. As Gerfin (1996) observes, however, computation of the kernel estimator can be cast as a Fourier transform, for which the fast Fourier transform reduces the amount of computation to the order of $n \log_2 n$. This value is only slightly larger than linear in n. See Press et al. (1986) and Gerfin (1996).

TABLE 21.6 Estimated Parameters for Semiparametric Models

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	h
Probit	5.62	3.11	−0.44	0.03	−1.07	−0.22	−1.00	1.07	—
	(1.35)	(0.77)	(0.10)	(0.03)	(0.26)	(0.09)	—	(0.29)	
Single	—	2.98	−0.44	0.02	−1.32	−0.25	−1.00	1.06	0.40
index	—	(0.90)	(0.12)	(0.03)	(0.33)	(0.11)	—	(0.32)	
MSCORE	5.83	2.84	−0.40	0.03	−0.80	−0.16	−1.00	0.91	0.70
	(1.78)	(0.98)	(0.13)	(0.05)	(0.43)	(0.20)	—	(0.57)	

the estimators are almost indistinguishable, but the results for the first differ widely. Of 401 ones (out of 873 observations), the counts of predicted ones are 389 for probit, 382 for Klein/Spady, and 355 for MSCORE. (The results do not indicate how many of these counts are correct predictions.)

21.5.5 A KERNEL ESTIMATOR FOR A NONPARAMETRIC REGRESSION FUNCTION

As noted, one unsatisfactory aspect of semiparametric formulations such as MSCORE is that the amount of information that the procedure provides about the population is limited; this aspect is, after all, the purpose of dispensing with the firm (parametric) assumptions of the probit and logit models. Thus, in the preceding example, there is little that one can say about the population that generated the data based on the MSCORE "estimates" in the table. The estimates do allow predictions of the response variable. But there is little information about any relationship between the response and the independent variables based on the "estimation" results. Even the mean-squared deviation matrix is suspect as an estimator of the asymptotic covariance matrix of the MSCORE coefficients.

The authors of the technique have proposed a secondary analysis of the results. Let

$$F_{\beta}(z_i) = E[y_i \mid \mathbf{x}_i'\boldsymbol{\beta} = z_i]$$

denote a smooth regression function for the response variable. Based on a parameter vector $\boldsymbol{\beta}$, the authors propose to estimate the regression by the **method of kernels** as follows. For the n observations in the sample and for the given $\boldsymbol{\beta}$ (e.g., $\mathbf{b}_n$ from MSCORE), let

$$z_i = \mathbf{x}_i'\boldsymbol{\beta},$$

$$s = \left[\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2\right]^{1/2}.$$

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 16.4.1.) 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

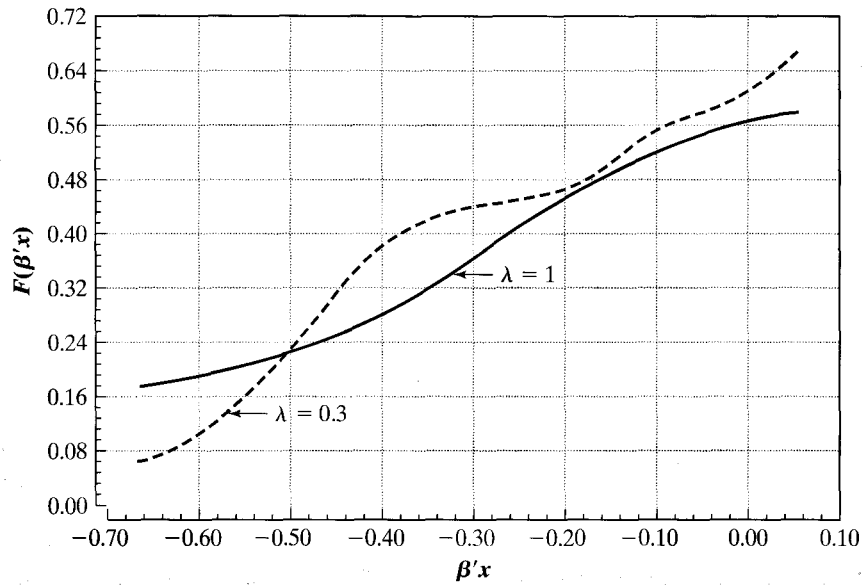
$$F(z^*) \approx \frac{\sum_{i=1}^{n} w_i(z^*) y_i}{\sum_{i=1}^{n} w_i(z^*)}.$$

Example 21.9 Nonparametric Regression

Figure 21.3 shows a plot of two estimates of the regression function for $E[\text{GRADE} \mid z]$. The coefficients are the MSCORE estimates given in Table 21.5. The plot is produced by computing fitted values for 100 equally spaced points in the range of $\mathbf{x}'\mathbf{b}_n$, which for these data and coefficients is [−0.66229, 0.05505]. The function is estimated with two values of the smoothing parameter, 1.0 and 0.3. As expected, the function based on $\lambda = 1.0$ is much flatter than that based on $\lambda = 0.3$. Clearly, the results of the analysis are crucially dependent on the value assumed.

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

FIGURE 21.3 Nonparametric Regression.



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 $\boldsymbol{\beta}$ (that is, *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.

21.5.6 DYNAMIC BINARY CHOICE MODELS

A random or fixed effects model which 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.

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 Kyriadizou (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 noted earlier 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

limit its practical applicability. Honore and Kyriazidou (2000) have combined the logic of the conditional logit model and Manski's maximum score estimator. They specify

$$\text{Prob}(y_{i0} = 1 \mid \mathbf{x}_i, \alpha_i) = p_0(\mathbf{x}_i, \alpha_i) \quad \text{where } \mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT})$$

$$\text{Prob}(y_{it} = 1 \mid \mathbf{x}_i, \alpha_i, y_{i0}, y_{i1}, \ldots, y_{i,t-1}) = F(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \gamma y_{i,t-1}) \quad t = 1, \ldots, T$$

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

Semiparametric estimators for dynamic models at this point in the development are still primarily of theoretical interest. Models that extend the parametric formulations to include state dependence have a much longer history, including Heckman (1978, 1981a, 1981b), Heckman and MaCurdy (1980), Jakubson (1988), Keane (1993) and Beck et al. (2001) to name a few.[48] In general, even without heterogeneity, dynamic models ultimately involve modeling the joint outcome $(y_{i0}, \ldots, y_{iT})$ which necessitates some treatment involving multivariate integration. Example 21.10 describes a recent application.

Example 21.10 An Intertemporal Labor Force Participation Equation

Hyslop (1999) presents a model of the labor force participation of married women. The focus of the study is the high degree of persistence in the participation decision. Data used in the study were the years 1979–1985 of the Panel Study of Income Dynamics. A sample of 1812 continuously married couples were studied. Exogenous variables which appeared in the model were measures of permanent and transitory income and fertility captured in yearly counts of the number of children from 0–2, 3–5 and 6–17 years old. Hyslop's formulation, in general terms, is

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

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

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

(stochastic specification)

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

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

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

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

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

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

The presence of the autocorrelation and state dependence in the model invalidate the simple maximum likelihood procedures we have 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 E.4.2e. 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 17.8.

21.6 BIVARIATE AND MULTIVARIATE PROBIT MODELS

In Chapter 14, we analyzed a number of different multiple-equation extensions of the classical and generalized regression model. A natural extension of the probit model would be to allow more than one equation, with correlated disturbances, in the same spirit as the seemingly unrelated regressions model. The general specification for a two-equation model would be

$$\begin{aligned}
&y_1^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1, \quad y_1 = 1 \quad \text{if } y_1^* > 0, 0 \text{ otherwise,}\\
&y_2^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2, \quad y_2 = 1 \quad \text{if } y_2^* > 0, 0 \text{ otherwise,}\\
&E[\varepsilon_1 \mid \mathbf{x}_1, \mathbf{x}_2] = E[\varepsilon_2 \mid \mathbf{x}_1, \mathbf{x}_2] = 0,\\
&\text{Var}[\varepsilon_1 \mid \mathbf{x}_1, \mathbf{x}_2] = \text{Var}[\varepsilon_2 \mid \mathbf{x}_1, \mathbf{x}_2] = 1,\\
&\text{Cov}[\varepsilon_1, \varepsilon_2 \mid \mathbf{x}_1, \mathbf{x}_2] = \rho.
\end{aligned} \tag{21-41}$$

21.6.1 MAXIMUM LIKELIHOOD ESTIMATION

The bivariate normal cdf is

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

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

$$\phi_2(x_1, x_2, \rho) = \frac{e^{-(1/2)(x_1^2 + x_2^2 - 2\rho x_1 x_2)/(1-\rho^2)}}{2\pi(1-\rho^2)^{1/2}}.$$ [49]

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

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

and

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

Note the national convention. The subscript 2 is used to indicate the bivariate normal distribution in the density ϕ_2 and cdf Φ_2. In all other cases, the subscript 2 indicates

[49] See Section B.9.

the variables in the second equation above. As before, $\phi(.)$ and $\Phi(.)$ without subscripts denote the univariate standard normal density and cdf.

The probabilities that enter the likelihood function are

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

which accounts for all the necessary sign changes needed to compute probabilities for ys equal to zero and one. Thus,

$$\log L = \sum_{i=1}^{n} \ln \Phi_2(w_{i1}, w_{i2}, \rho_{i^*}).^{50}$$

The derivatives of the log-likelihood then reduce to

$$\begin{aligned} \frac{\partial \ln L}{\partial \boldsymbol{\beta}_j} &= \sum_{i=1}^{n} \left(\frac{q_{ij} g_{ij}}{\Phi_2} \right) \mathbf{x}_{ij}, \quad j = 1, 2, \\ \frac{\partial \ln L}{\partial \rho} &= \sum_{i=1}^{n} \frac{q_{i1} q_{i2} \phi_2}{\Phi_2}, \end{aligned} \tag{21-42}$$

where

$$g_{i1} = \phi(w_{i1}) \Phi \left[\frac{w_{i2} - \rho_{i^*} w_{i1}}{\sqrt{1 - \rho_{i^*}^2}} \right] \tag{21-43}$$

and the subscripts 1 and 2 in g_{i1} are reversed to obtain g_{i2}. Before considering the Hessian, it is useful to note what becomes of the preceding if $\rho = 0$. For $\partial \ln L / \partial \boldsymbol{\beta}_1$, if $\rho = \rho_{i^*} = 0$, then g_{i1} reduces to $\phi(w_{i1})\Phi(w_{i2})$, ϕ_2 is $\phi(w_{i1})\phi(w_{i2})$, and Φ_2 is $\Phi(w_{i1})\Phi(w_{i2})$. Inserting these results in (21-42) with q_{i1} and q_{i2} produces (21-21). Since both functions in $\partial \ln L / \partial \rho$ factor into the product of the univariate functions, $\partial \ln L / \partial \rho$ reduces to $\sum_{i=1}^{n} \lambda_{i1} \lambda_{i2}$ where λ_{ij}, $j = 1, 2$, is defined in (21-21). (This result will reappear in the LM statistic below.)

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

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

By multiplying it out, you can show that

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

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

Then

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1'} = \sum_{i=1}^{n} \mathbf{x}_{i1}\mathbf{x}_{i1}' \left[\frac{-w_{i1}g_{i1}}{\Phi_2} - \frac{\rho_{i^*}\phi_2}{\Phi_2} - \frac{g_{i1}^2}{\Phi_2^2} \right],$$

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_2'} = \sum_{i=1}^{n} q_{i1}q_{i2}\mathbf{x}_{i1}\mathbf{x}_{i2}' \left[\frac{\phi_2}{\Phi_2} - \frac{g_{i1}g_{i2}}{\Phi_2^2} \right],$$

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}_1 \partial \rho} = \sum_{i=1}^{n} q_{i2}\mathbf{x}_{i1} \frac{\phi_2}{\Phi_2} \left[\rho_{i^*}\delta_i v_{i1} - w_{i1} - \frac{g_{i1}}{\Phi_2} \right],$$

$$\frac{\partial^2 \log L}{\partial \rho^2} = \sum_{i=1}^{n} \frac{\phi_2}{\Phi_2} \left[\delta_i^2 \rho_{i^*}(1 - \mathbf{w}_i'\mathbf{R}_i^{-1}\mathbf{w}_i) + \delta_i^2 w_{i1}w_{i2} - \frac{\phi_2}{\Phi_2} \right],$$

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

21.6.2 TESTING FOR ZERO CORRELATION

The Lagrange multiplier statistic is a convenient device for testing for the absence of correlation in this model. Under the null hypothesis that ρ equals zero, the model consists of independent probit equations, which can be estimated separately. Moreover, in the multivariate model, all the bivariate (or multivariate) densities and probabilities factor into the products of the marginals if the correlations are zero, which makes construction of the test statistic a simple matter of manipulating the results of the independent probits. The Lagrange multiplier statistic for testing H_0: $\rho = 0$ in a bivariate probit model is[51]

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

As usual, the advantage of the LM statistic is that it obviates computing the bivariate probit model. But, the full unrestricted model is now fairly common in commercial software, so that advantage is minor. The likelihood ratio or Wald test can often be used with equal ease.

21.6.3 MARGINAL EFFECTS

There are several "marginal effects" one might want to evaluate in a bivariate probit model.[52] For convenience in evaluating them, we will define a vector $\mathbf{x} = \mathbf{x}_1 \cup \mathbf{x}_2$ and let

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

[52] See Greene (1996b).

$\mathbf{x}_1'\boldsymbol{\beta}_1 = \mathbf{x}'\boldsymbol{\gamma}_1$. Thus, $\boldsymbol{\gamma}_1$ contains all the nonzero elements of $\boldsymbol{\beta}_1$ and possibly some zeros in the positions of variables in $\mathbf{x}$ that appear only in the other equation; $\boldsymbol{\gamma}_2$ is defined likewise. The bivariate probability is

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

Signs are changed appropriately if the probability of the zero outcome is desired in either case. (See 21-41.) The marginal effects of changes in $\mathbf{x}$ on this probability are given by

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

where g_1 and g_2 are defined in (21-43). The familiar univariate cases will arise if $\rho = 0$, and effects specific to one equation or the other will be produced by zeros in the corresponding position in one or the other parameter vector. There are also some conditional mean functions to consider. The unconditional mean functions are given by the univariate probabilities:

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

so the analysis of (21-9) and (21-10) applies. One pair of conditional mean functions that might be of interest are

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

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

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

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

Finally, one might construct the nonlinear conditional mean function

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

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

21.6.4 SAMPLE SELECTION

There are situations in which the observed variables in the bivariate probit model are censored in one way or another. For example, in an evaluation of credit scoring models, Boyes, Hoffman, and Low (1989) analyzed data generated by the following rule:

$y_1 = 1$ if individual i defaults on a loan, 0 otherwise,

$y_2 = 2$ if the individual is granted a loan, 0 otherwise.

Greene (1992) applied the same model to y_1 = default on credit card loans, in which y_2 denotes whether an application for the card was accepted or not. For a given individual,

y_1 is not observed unless y_2 equals one. Thus, there are three types of observations in the sample, with unconditional probabilities:[53]

$$\begin{aligned}
y_2 = 0: &\quad \text{Prob}(y_2 = 0 \quad\quad | \mathbf{x}_1, \mathbf{x}_2) = 1 - \Phi(\mathbf{x}_2'\boldsymbol{\beta}_2), \\
y_1 = 0, y_2 = 1: &\quad \text{Prob}(y_1 = 0,\ y_2 = 1 | \mathbf{x}_1, \mathbf{x}_2) = \Phi_2[-\mathbf{x}_1'\boldsymbol{\beta}_1, \mathbf{x}_2'\boldsymbol{\beta}_2, -\rho], \\
y_1 = 1, y_2 = 1: &\quad \text{Prob}(y_1 = 1,\ y_2 = 1 | \mathbf{x}_1, \mathbf{x}_2) = \Phi_2[\mathbf{x}_1'\boldsymbol{\beta}_1, \mathbf{x}_2'\boldsymbol{\beta}_2, \rho].
\end{aligned}$$

The log-likelihood function is based on these probabilities.[54]

21.6.5 A MULTIVARIATE PROBIT MODEL

In principle, a multivariate model would extend (21-41) to more than two outcome variables just by adding equations. The practical obstacle to such an extension is primarily the evaluation of higher-order multivariate normal integrals. Some progress has been made on using quadrature for trivariate integration, but existing results are not sufficient to allow accurate and efficient evaluation for more than two variables in a sample of even moderate size. An altogether different approach has been used in recent applications. Lerman and Manski (1981) suggested that one might approximate multivariate normal probabilities by random sampling. For example, to approximate $\text{Prob}(y_1 > 1, y_2 < 3, y_3 < -1) | \mathbf{x}_1, \mathbf{x}_2, \rho_{12}, \rho_{13}, \rho_{23})$, we would simply draw random observations from this trivariate normal distribution (see Section E.5.6.) and count the number of observations that satisfy the inequality. To obtain an accurate estimate of the probability, quite a large number of draws is required. Also, the substantive possibility of getting zero such draws in a finite number of draws is problematic. Nonetheless, the logic of the Lerman–Manski approach is sound. As discussed in Section E.5.6 recent developments have produced methods of producing quite accurate estimates of multivariate normal integrals based on this principle. The evaluation of multivariate normal integral is generally a much less formidable obstacle to the estimation of models based on the multivariate normal distribution.[55]

McFadden (1989) pointed out that for purposes of maximum likelihood estimation, accurate evaluation of probabilities is not necessarily the problem that needs to be solved. One can view the computation of the log-likelihood and its derivatives as a problem of estimating a mean. That is, in (21-41) and (21-42), the same problem arises if we divide by n. The idea is that even though the individual terms in the average might be in error, if the error has mean zero, then it will average out in the summation. The important insight, then, is that if we can obtain probability estimates that only err randomly both positively and negatively, then it may be possible to obtain an estimate of the log-likelihood and its derivatives that is reasonably close to the one that would

[53] The model was first proposed by Wynand and van Praag (1981).

[54] Extensions of the bivariate probit model to other types of censoring are discussed in Poirier (1980) and Abowd and Farber (1982).

[55] Papers that propose improved methods of simulating probabilities include Pakes and Pollard (1989) and especially Börsch-Supan and Hajivassilou (1990), Geweke (1989), and Keane (1994). A symposium in the November 1994 issue of *Review of Economics and Statistics* presents discussion of numerous issues in specification and estimation of models based on simulation of probabilities. Applications that employ simulation techniques for evaluation of multivariate normal integrals are now fairly numerous. See, for example, Hyslop (1999) (Example 21.10) who applies the technique to a panel data application with $T = 7$.

result from actually computing the integral. From a practical standpoint, it does not take inordinately large numbers of random draws to achieve this result, which with the progress that has been made on Monte Carlo integration, has made feasible multivariate models that previously were intractable.

The multivariate probit model in another form presents a useful extension of the probit model to panel data. The structural equation for the model would be

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

The Butler and Moffitt approach for this model has proved useful in numerous applications. But, the underlying assumption that $\text{Cov}[\varepsilon_{it}, \varepsilon_{is}] = \rho$ is a substantive restriction. By treating this structure as a multivariate probit model with a restriction that the coefficient vector be the same in every period, one can obtain a model with free correlations across periods. Hyslop (1999) and Greene (2002) are two applications.

21.6.6 APPLICATION: GENDER ECONOMICS COURSES IN LIBERAL ARTS COLLEGES

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

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

The dependent variables in the model are

y_1 = presence of a gender economics course,

y_2 = presence of a women's studies program on the campus.

The independent variables in the model are

z_1 = constant term;

z_2 = academic reputation of the college, coded 1 (best), 2, . . . to 141;

z_3 = size of the full time economics faculty, a count;

z_4 = percentage of the economics faculty that are women, proportion (0 to 1);

z_5 = religious affiliation of the college, 0 = no, 1 = yes;

z_6 = percentage of the college faculty that are women, proportion (0 to 1);

z_7–z_{10} = regional dummy variables, south, midwest, northeast, west.

The regressor vectors are

$$\mathbf{x}_1 = z_1, z_2, z_3, z_4, z_5, \quad \mathbf{x}_2 = z_2, z_6, z_5, z_7\text{–}z_{10}.$$

Burnett's model illustrates a number of interesting aspects of the bivariate probit model. Note that this model is qualitatively different from the bivariate probit model in (21-41); the second dependent variable, y_2, appears on the right-hand side of the first equation. This model is a **recursive,** simultaneous-equations model. Surprisingly, the endogenous nature of one of the variables on the right-hand side of the first equation can be ignored in formulating the log-likelihood. [The model appears in Maddala (1983, p. 123).] We can establish this fact with the following (admittedly trivial) argument: The term that

enters the log-likelihood is $P(y_1 = 1, y_2 = 1) = P(y_1 = 1 \mid y_2 = 1)P(y_2 = 1)$. Given the model as stated, the marginal probability for y_2 is just $\Phi(\mathbf{x}_2'\boldsymbol{\beta}_2)$, whereas the conditional probability is $\Phi_2(\ldots)/\Phi(\mathbf{x}_2'\boldsymbol{\beta}_2)$. The product returns the probability we had earlier. The other three terms in the log-likelihood are derived similarly, which produces (Maddala's results with some sign changes):

$$P_{11} = \Phi_2(\mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma y_2, \mathbf{x}_2'\boldsymbol{\beta}_2, \rho), \qquad P_{10} = \Phi_2(\mathbf{x}_1'\boldsymbol{\beta}_1, -\mathbf{x}_2'\boldsymbol{\beta}_2, -\rho)$$
$$P_{01} = \Phi_2[-(\mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma y_2), \boldsymbol{\beta}_2'\mathbf{x}_2, -\rho], \qquad P_{00} = \Phi_2(-\mathbf{x}_1'\boldsymbol{\beta}_1, -\mathbf{x}_2'\boldsymbol{\beta}_2, \rho).$$

These terms are exactly those of (21-41) that we obtain just by carrying y_2 in the first equation with no special attention to its endogenous nature. We can ignore the simultaneity in this model and we cannot in the linear regression model because, in this instance, we are maximizing the log-likelihood, whereas in the linear regression case, we are manipulating certain sample moments that do not converge to the necessary population parameters in the presence of simultaneity. Note that the same result is at work in Section 15.6.2, where the FIML estimator of the simultaneous equations model is obtained with the endogenous variables on the right-hand sides of the equations, *but not by using ordinary least squares.*

The marginal effects in this model are fairly involved, and as before, we can consider several different types. Consider, for example, z_2, academic reputation. There is a direct effect produced by its presence in the first equation, but there is also an indirect effect. Academic reputation enters the women's studies equation and, therefore, influences the probability that y_2 equals one. Since y_2 appears in the first equation, this effect is transmitted back to y_1. The total effect of academic reputation and, likewise, religious affiliation is the sum of these two parts. Consider first the gender economics variable, y_1. The conditional mean is

$$\begin{aligned} E[y_1 \mid \mathbf{x}_1, \mathbf{x}_2] &= \text{Prob}[y_2 = 1]E[y_1 \mid y_2 = 1, \mathbf{x}_1, \mathbf{x}_2] + \text{Prob}[y_2 = 0]E[y_1 \mid y_2 = 0, \mathbf{x}_1, \mathbf{x}_2] \\ &= \Phi_2(\mathbf{x}_1'\beta_1 + \gamma y_2, \mathbf{x}_2'\boldsymbol{\beta}_2, \rho) + \Phi_2(\mathbf{x}_1'\boldsymbol{\beta}_1, -\mathbf{x}_2'\boldsymbol{\beta}_2, -\rho). \end{aligned}$$

Derivatives can be computed using our earlier results. We are also interested in the effect of religious affiliation. Since this variable is binary, simply differentiating the conditional mean function may not produce an accurate result. Instead, we would compute the conditional mean function with this variable set to one and then zero, and take the difference. Finally, what is the effect of the presence of a women's studies program on the probability that the college will offer a gender economics course? To compute this effect, we would compute $\text{Prob}[y_1 = 1 \mid y_2 = 1, \mathbf{x}_1, \mathbf{x}_2] - \text{Prob}[y_1 = 1 \mid y_2 = 0, \mathbf{x}_1, \mathbf{x}_2]$. In all cases, standard errors for the estimated marginal effects can be computed using the delta method.

Maximum likelihood estimates of the parameters of Burnett's model were computed by Greene (1998) using her sample of 132 liberal arts colleges; 31 of the schools offer gender economics, 58 have women's studies, and 29 have both. The estimated parameters are given in Table 21.7. Both bivariate probit and the single-equation estimates are given. The estimate of ρ is only 0.1359, with a standard error of 1.2359. The Wald statistic for the test of the hypothesis that ρ equals zero is $(0.1359/1.2539)^2 = 0.011753$. For a single restriction, the critical value from the chi-squared table is 3.84, so the hypothesis cannot be rejected. The likelihood ratio statistic for the same hypothesis is

TABLE 21.7 Estimates of a Recursive Simultaneous Bivariate Probit Model (Estimated Standard Errors in Parentheses)

	Single Equation		*Bivariate Probit*	
Variable	*Coefficient*	*Standard Error*	*Coefficient*	*Standard Error*
Gender Economics Equation				
Constant	−1.4176	(0.8069)	−1.1911	(2.2155)
AcRep	−0.01143	(0.004081)	−0.01233	(0.007937)
WomStud	1.1095	(0.5674)	0.8835	(2.2603)
EconFac	0.06730	(0.06874)	0.06769	(0.06952)
PctWecon	2.5391	(0.9869)	2.5636	(1.0144)
Relig	−0.3482	(0.4984)	−0.3741	(0.5265)
Women's Studies Equation				
AcRep	−0.01957	(0.005524)	−0.01939	(0.005704)
PctWfac	1.9429	(0.8435)	1.8914	(0.8714)
Relig	−0.4494	(0.3331)	−0.4584	(0.3403)
South	1.3597	(0.6594)	1.3471	(0.6897)
West	2.3386	(0.8104)	2.3376	(0.8611)
North	1.8867	(0.8204)	1.9009	(0.8495)
Midwest	1.8248	(0.8723)	1.8070	(0.8952)
ρ	0.0000	(0.0000)	0.1359	(1.2539)
Log L	−85.6458		−85.6317	

$2[-85.6317 - (-85.6458)] = 0.0282$, which leads to the same conclusion. The Lagrange multiplier statistic is 0.003807, which is consistent. This result might seem counterintuitive, given the setting. Surely "gender economics" and "women's studies" are highly correlated, but this finding does not contradict that proposition. The correlation coefficient measures the correlation between the disturbances in the equations, the omitted factors. That is, ρ measures (roughly) the correlation between the outcomes after the influence of the included factors is accounted for. Thus, the value 0.13 measures the effect after the influence of women's studies is already accounted for. As discussed in the next paragraph, the proposition turns out to be right. The single most important determinant (at least within this model) of whether a gender economics course will be offered is indeed whether the college offers a women's studies program.

Table 21.8 presents the estimates of the marginal effects and some descriptive statistics for the data. The calculations were simplified slightly by using the restricted model with $\rho = 0$. Computations of the marginal effects still require the decomposition above, but they are simplified slightly by the result that if ρ equals zero, then the bivariate probabilities factor into the products of the marginals. Numerically, the strongest effect appears to be exerted by the representation of women on the faculty; its coefficient of +0.4491 is by far the largest. This variable, however, cannot change by a full unit because it is a proportion. An increase of 1 percent in the presence of women on the faculty raises the probability by only +0.004, which is comparable in scale to the effect of academic reputation. The effect of women on the faculty is likewise fairly small, only 0.0013 per 1 percent change. As might have been expected, the single most important influence is the presence of a women's studies program, which increases the likelihood of a gender economics course by a full 0.1863. Of course, the raw data would have anticipated this result; of the 31 schools that offer a gender economics course, 29 also

TABLE 21.8 Marginal Effects in Gender Economics Model

	Direct	*Indirect*	*Total*	*(Std. Error)*	*(Type of Variable, Mean)*
Gender Economics Equation					
AcRep	−0.002022	−0.001453	−0.003476	(0.00126)	(Continuous, 119.242)
PctWecon	+0.4491		+0.4491	(0.1568)	(Continuous, 0.24787)
EconFac	+0.01190		+0.1190	(0.01292)	(Continuous, 6.74242)
Relig	−0.07049	−0.03227	−0.1028	(0.1055)	(Binary, 0.57576)
WomStud	+0.1863		+0.1863	(0.0868)	(Endogenous, 0.43939)
PctWfac		+0.13951	+0.13951	(0.08916)	(Continuous, 0.35772)
Women's Studies Equation					
AcRep	−0.00754		−0.00754	(0.002187)	(Continuous, 119.242)
PctWfac	+0.13789		+0.13789	(0.01002)	(Continuous, 0.35772)
Relig	−0.13265		−0.13266	(0.18803)	(Binary, 0.57576)

have a women's studies program and only two do not. Note finally that the effect of religious affiliation (whatever it is) is mostly direct.

Before closing this application, we can use this opportunity to examine the fit measures listed in Section 21.4.5. We computed the various fit measures using seven different specifications of the gender economics equation:

1. Single-equation probit estimates, z_1, z_2, z_3, z_4, z_5, y_2
2. Bivariate probit model estimates, z_1, z_2, z_3, z_4, z_5, y_2
3. Single-equation probit estimates, z_1, z_2, z_3, z_4, z_5
4. Single-equation probit estimates, z_1, z_3, z_5, y_2
5. Single-equation probit estimates, z_1, z_3, z_5
6. Single-equation probit estimates, z_1, z_5
7. Single-equation probit estimates z_1 (constant only).

The specifications are in descending "quality" because we removed the most statistically significant variables from the model at each step. The values are listed in Table 21.9. The matrix below each column is the table of "hits" and "misses" of the prediction rule $\hat{y} = 1$ if $\hat{P} > 0.5, 0$ otherwise. [Note that by construction, model (7) must predict all ones or all zeros.] The column is the actual count and the row is the prediction. Thus, for model (1), 92 of 101 zeros were predicted correctly, whereas five of 31 ones were predicted incorrectly. As one would hope, the fit measures decline as the more significant

TABLE 21.9 Binary Choice Fit Measures

Measure	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
LRI	0.573	0.535	0.495	0.407	0.279	0.206	0.000
R^2_{BL}	0.844	0.844	0.823	0.797	0.754	0.718	0.641
λ	0.565	0.560	0.526	0.444	0.319	0.216	0.000
R^2_{EF}	0.561	0.558	0.530	0.475	0.343	0.216	0.000
R^2_{VZ}	0.708	0.707	0.672	0.589	0.447	0.352	0.000
R^2_{MZ}	0.687	0.679	0.628	0.567	0.545	0.329	0.000
Predictions	$\begin{bmatrix} 92 & 9 \\ 5 & 26 \end{bmatrix}$	$\begin{bmatrix} 93 & 8 \\ 5 & 26 \end{bmatrix}$	$\begin{bmatrix} 92 & 9 \\ 8 & 23 \end{bmatrix}$	$\begin{bmatrix} 94 & 7 \\ 8 & 23 \end{bmatrix}$	$\begin{bmatrix} 98 & 3 \\ 16 & 15 \end{bmatrix}$	$\begin{bmatrix} 101 & 0 \\ 31 & 0 \end{bmatrix}$	$\begin{bmatrix} 101 & 0 \\ 31 & 0 \end{bmatrix}$

variables are removed from the model. The Ben-Akiva measure has an obvious flaw in that with only a constant term, the model still obtains a "fit" of 0.641. From the prediction matrices, it is clear that the explanatory power of the model, such as it is, comes from its ability to predict the ones correctly. The poorer is the model, the greater the number of correct predictions of $y = 0$. But as this number rises, the number of incorrect predictions rises and the number of correct predictions of $y = 1$ declines. All the fit measures appear to react to this feature to some degree. The Efron and Cramer measures, which are nearly identical, and McFadden's LRI appear to be most sensitive to this, with the remaining two only slightly less consistent.

21.7 LOGIT MODELS FOR MULTIPLE CHOICES

Some studies of multiple-choice settings include the following:

1. Hensher (1986), McFadden (1974), and many others have analyzed the travel mode of urban commuters.
2. Schmidt and Strauss (1975a,b) and Boskin (1974) have analyzed occupational choice among multiple alternatives.
3. Terza (1985) has studied the assignment of bond ratings to corporate bonds as a choice among multiple alternatives.

These are all distinct from the multivariate probit model we examined earlier. In that setting, there were several decisions, each between two alternatives. Here there is a single decision among two or more alternatives. We will examine two broad types of choice sets, **ordered** and **unordered.** The choice among means of getting to work—by car, bus, train, or bicycle—is clearly unordered. A bond rating is, by design, a ranking; that is its purpose. As we shall see, quite different techniques are used for the two types of models. Models for unordered choice sets are considered in this section. A model for ordered choices is described in Section 21.8.

Unordered-choice models can be motivated by a random utility model. For the ith consumer faced with J choices, suppose that the utility of choice j is

$$U_{ij} = \mathbf{z}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij}.$$

If the consumer makes choice j in particular, then we assume that U_{ij} is the maximum among the J utilities. Hence, the statistical model is driven by the probability that choice j is made, which is

$$\text{Prob}(U_{ij} > U_{ik}) \quad \text{for all other } k \neq j.$$

The model is made operational by a particular choice of distribution for the disturbances. As before, two models have been considered, logit and probit. Because of the need to evaluate multiple integrals of the normal distribution, the probit model has found rather limited use in this setting. The logit model, in contrast, has been widely used in many fields, including economics, market research, and transportation engineering. Let Y_i be a random variable that indicates the choice made. McFadden (1973) has shown that if (and only if) the J disturbances are independent and identically distributed with

type I extreme value (Gumbel) distribution,

$$F(\varepsilon_{ij}) = \exp(-e^{-\varepsilon_{ij}}),$$

then

$$\text{Prob}(Y_i = j) = \frac{e^{\mathbf{z}_{ij}'\boldsymbol{\beta}}}{\sum_{j=1}^{J} e^{\mathbf{z}_{ij}'\boldsymbol{\beta}}}, \tag{21-44}$$

which leads to what is called the **conditional logit** model.[56]

Utility depends on $\mathbf{x}_{ij}$, which includes aspects specific to the individual as well as to the choices. It is useful to distinguish them. Let $\mathbf{z}_{ij} = [\mathbf{x}_{ij}, \mathbf{w}_i]$. Then $\mathbf{x}_{ij}$ varies across the choices and possibly across the individuals as well. The components of $\mathbf{x}_{ij}$ are typically called the **attributes** of the choices. But $\mathbf{w}_i$ contains the **characteristics** of the individual and is, therefore, the same for all choices. If we incorporate this fact in the model, then (21-44) becomes

$$\text{Prob}(Y_i = j) = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_{ij} + \boldsymbol{\alpha}'\mathbf{w}_i}}{\sum_{j=1}^{J} e^{\boldsymbol{\beta}'\mathbf{x}_{ij} + \boldsymbol{\alpha}'\mathbf{w}_i}} = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_{ij}} e^{\boldsymbol{\alpha}_i'\mathbf{w}_i}}{\sum_{j=1}^{J} e^{\boldsymbol{\beta}'\mathbf{x}_{ij}} e^{\boldsymbol{\alpha}_i'\mathbf{w}_i}}.$$

Terms that do not vary across alternatives—that is, those specific to the individual—fall out of the probability. Evidently, if the model is to allow individual specific effects, then it must be modified. One method is to create a set of dummy variables for the choices and multiply each of them by the common $\mathbf{w}$. We then allow the coefficient to vary across the choices instead of the characteristics. Analogously to the linear model, a complete set of interaction terms creates a singularity, so one of them must be dropped. For example, a model of a shopping center choice by individuals might specify that the choice depends on attributes of the shopping centers such as number of stores and distance from the central business district, both of which are the same for all individuals. and income, which varies across individuals. Suppose that there were three choices. The three regressor vectors would be as follows:

Choice 1:	Stores	Distance	Income	0
Choice 2:	Stores	Distance	0	Income
Choice 3:	Stores	Distance	0	0

The data sets typically analyzed by economists do not contain mixtures of individual- and choice-specific attributes. Such data would be far too costly to gather for most purposes. When they do, the preceding framework can be used. For the present, it is useful to examine the two types of data separately and consider aspects of the model that are specific to the two types of applications.

21.7.1 THE MULTINOMIAL LOGIT MODEL

To set up the model that applies when data are individual specific, it will help to consider an example. Schmidt and Strauss (1975a,b) estimated a model of occupational

[56] It is occasionally labeled the **multinomial logit model,** but this wording conflicts with the usual name for the model discussed in the next section, which differs slightly. Although the distinction turns out to be purely artificial, we will maintain it for the present.

choice based on a sample of 1000 observations drawn from the Public Use Sample for three years, 1960, 1967, and 1970. For each sample, the data for each individual in the sample consist of the following:

1. *Occupation:* 0 = menial, 1 = blue collar, 2 = craft, 3 = white collar, 4 = professional.
2. *Regressors:* constant, education, experience, race, sex.

The model for occupational choice is

$$\text{Prob}(Y_i = j) = \frac{e^{\boldsymbol{\beta}_j' \mathbf{x}_i}}{\sum_{k=0}^{4} e^{\boldsymbol{\beta}_k' \mathbf{x}_i}}, \quad j = 0, 1, \ldots, 4. \tag{21-45}$$

(The binomial logit of Sections 21.3 and 21.4 is conveniently produced as the special case of $J = 1$.)

The model in (21-45) is a **multinomial logit model.**[57] The estimated equations provide a set of probabilities for the $J + 1$ choices for a decision maker with characteristics $\mathbf{x}_i$. Before proceeding, we must remove an indeterminacy in the model. If we define $\boldsymbol{\beta}_j^* = \boldsymbol{\beta}_j + \mathbf{q}$ for any vector $\mathbf{q}$, then recomputing the probabilities defined below using $\boldsymbol{\beta}_j^*$ instead of $\boldsymbol{\beta}_j$ produces the identical set of probabilities because all the terms involving $\mathbf{q}$ drop out. A convenient normalization that solves the problem is $\boldsymbol{\beta}_0 = \mathbf{0}$. (This arises because the probabilities sum to one, so only J parameter vectors are needed to determine the $J + 1$ probabilities.) Therefore, the probabilities are

$$\text{Prob}(Y_i = j \mid \mathbf{x}_i) = \frac{e^{\boldsymbol{\beta}_j' \mathbf{x}_i}}{1 + \sum_{k=1}^{J} e^{\boldsymbol{\beta}_k' \mathbf{x}_i}} \quad \text{for } j = 0, 2, \ldots, J, \boldsymbol{\beta}_0 = \mathbf{0}. \tag{21-46}$$

The form of the binomial model examined in Section 21.4 results if $J = 1$. The model implies that we can compute J log-odds ratios

$$\ln\left[\frac{P_{ij}}{P_{ik}}\right] = \mathbf{x}_i'(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k) = \mathbf{x}_i'\boldsymbol{\beta}_j \quad \text{if } k = 0.$$

From the point of view of estimation, it is useful that the odds ratio, P_j / P_k, does not depend on the other choices, which follows from the independence of disturbances in the original model. From a behavioral viewpoint, this fact is not very attractive. We shall return to this problem in Section 21.7.3.

The log-likelihood can be derived by defining, for each individual, $d_{ij} = 1$ if alternative j is chosen by individual i, and 0 if not, for the $J - 1$ possible outcomes. Then, for each i, one and only one of the d_{ij}'s is 1. The log-likelihood is a generalization of that for the binomial probit or logit model:

$$\ln L = \sum_{i=1}^{n} \sum_{j=0}^{J} d_{ij} \ln \text{Prob}(Y_i = j).$$

The derivatives have the characteristically simple form

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}_j} = \sum_i (d_{ij} - P_{ij}) \mathbf{x}_i \quad \text{for } j = 1, \ldots, J.$$

[57] Nerlove and Press (1973).

The exact second derivatives matrix has $J^2K \times K$ blocks,

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_l'} = -\sum_{i=1}^{n} P_{ij}[\mathbf{1}(j = l) - P_{il}]\mathbf{x}_i\mathbf{x}_i',^{58}$$

where $\mathbf{1}(j = l)$ equals 1 if j equals l and 0 if not. Since the Hessian does not involve d_{ij}, these are the expected values, and Newton's method is equivalent to the method of scoring. It is worth noting that the number of parameters in this model proliferates with the number of choices, which is unfortunate because the typical cross section sometimes involves a fairly large number of regressors.

The coefficients in this model are difficult to interpret. It is tempting to associate $\boldsymbol{\beta}_j$ with the jth outcome, but that would be misleading. By differentiating (21-46), we find that the marginal effects of the characteristics on the probabilities are

$$\boldsymbol{\delta}_j = \frac{\partial P_j}{\partial \mathbf{x}_i} = P_j\left[\boldsymbol{\beta}_j - \sum_{k=0}^{J} P_k\boldsymbol{\beta}_k\right] = P_j[\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}]. \tag{21-47}$$

Therefore, every subvector of $\boldsymbol{\beta}$ enters every marginal effect, both through the probabilities and through the weighted average that appears in $\boldsymbol{\delta}_j$. These values can be computed from the parameter estimates. Although the usual focus is on the coefficient estimates, equation (21-47) suggests that there is at least some potential for confusion. Note, for example, that for any particular x_k, $\partial P_j/\partial x_k$ need not have the same sign as β_{jk}. Standard errors can be estimated using the delta method. (See Section 5.2.4.) For purposes of the computation, let $\boldsymbol{\beta} = [\mathbf{0}, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \ldots, \boldsymbol{\beta}_J']'$. We include the fixed $\mathbf{0}$ vector for outcome 0 because although $\boldsymbol{\beta}_0 = \mathbf{0}$, $\boldsymbol{\gamma}_0 = -P_0\bar{\boldsymbol{\beta}}$, which is not $\mathbf{0}$. Note as well that Asy. $\text{Cov}[\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_j] = \mathbf{0}$ for $j = 0, \ldots, J$. Then

$$\text{Asy. Var}[\hat{\boldsymbol{\delta}}_j] = \sum_{l=0}^{J}\sum_{m=0}^{J}\left(\frac{\partial \boldsymbol{\delta}_j}{\partial \boldsymbol{\beta}_l'}\right)\text{Asy. Cov}[\hat{\boldsymbol{\beta}}_l, \hat{\boldsymbol{\beta}}_m]\left(\frac{\partial \boldsymbol{\delta}_j'}{\partial \boldsymbol{\beta}_m}\right),$$

$$\frac{\partial \boldsymbol{\delta}_j}{\partial \boldsymbol{\beta}_l} = [\mathbf{1}(j = l) - P_l][P_j\mathbf{I} + \boldsymbol{\delta}_j\mathbf{x}'] + P_j[\boldsymbol{\delta}_l\mathbf{x}'].$$

Finding adequate fit measures in this setting presents the same difficulties as in the binomial models. As before, it is useful to report the log-likelihood. If the model contains no covariates and no constant term, then the log-likelihood will be

$$\ln L_c = \sum_{j=0}^{J} n_j \ln\left(\frac{1}{J+1}\right).$$

where n_j is the number of individuals who choose outcome j. If the regressor vector includes only a constant term, then the restricted log-likelihood is

$$\ln L_0 = \sum_{j=0}^{J} n_j \ln\left(\frac{n_j}{n}\right) = \sum_{j=0}^{J} n_j \ln p_j,$$

[58] If the data were in the form of proportions, such as market shares, then the appropriate log-likelihood and derivatives are $\sum_i \sum_j n_i p_{ij}$ and $\sum_i \sum_j n_i(p_{ij} - P_{ij})\mathbf{x}_i$, respectively. The terms in the Hessian are multiplied by n_i.

where p_j is the sample proportion of observations that make choice j. If desired, the likelihood ratio index can also be reported. A useful table will give a listing of hits and misses of the prediction rule "predict $Y_i = j$ if $\hat{P}_j$ is the maximum of the predicted probabilities."[59]

21.7.2 THE CONDITIONAL LOGIT MODEL

When the data consist of choice-specific attributes instead of individual-specific characteristics, the appropriate model is

$$\text{Prob}(Y_i = j \mid \mathbf{z}_{i1}, \mathbf{z}_{i2}, \ldots, \mathbf{z}_{iJ}) = \frac{e^{\boldsymbol{\beta}'\mathbf{z}_{ij}}}{\sum_{j=1}^{J} e^{\boldsymbol{\beta}'\mathbf{z}_{ij}}}. \tag{21-48}$$

Here, in accordance with the convention in the literature, we let $j = 1, 2, \ldots, J$ for a total of J alternatives. The model is otherwise essentially the same as the multinomial logit. Even more care will be required in interpreting the parameters, however. Once again, an example will help to focus ideas.

In this model, the coefficients are not directly tied to the marginal effects. The marginal effects for continuous variables can be obtained by differentiating (21-48) with respect to $\mathbf{x}$ to obtain

$$\frac{\partial P_j}{\partial \mathbf{x}_k} = [P_j(\mathbf{1}(j = k) - P_k)]\boldsymbol{\beta}, \quad k = 1, \ldots, J.$$

(To avoid cluttering the notation, we have dropped the observation subscript.) It is clear that through its presence in P_j and P_k, every attribute set $\mathbf{x}_j$ affects all the probabilities. Hensher suggests that one might prefer to report elasticities of the probabilities. The effect of attribute m of choice k on P_j would be

$$\frac{\partial \log P_j}{\partial \log x_{km}} = x_{km}[\mathbf{1}(j = k) - P_k]\beta_m.$$

Since there is no ambiguity about the scale of the probability itself, whether one should report the derivatives or the elasticities is largely a matter of taste. Some of Hensher's elasticity estimates are given in Table 21.16 later on in this chapter.

Estimation of the conditional logit model is simplest by Newton's method or the method of scoring. The log-likelihood is the same as for the multinomial logit model. Once again, we define $d_{ij} = 1$ if $Y_i = j$ and 0 otherwise. Then

$$\log L = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} \log \text{Prob}(Y_i = j).$$

Market share and frequency data are common in this setting. If the data are in this form, then the only change needed is, once again, to define d_{ij} as the proportion or frequency.

[59] Unfortunately, it is common for this rule to predict all observation with the same value in an unbalanced sample or a model with little explanatory power.

Because of the simple form of L, the gradient and Hessian have particularly convenient forms: Let $\bar{\mathbf{x}}_i = \sum_{j=1}^{J} P_{ij}\mathbf{x}_{ij}$. Then,

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i),$$

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n}\sum_{j=1}^{J} P_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)',$$

The usual problems of fit measures appear here. The log-likelihood ratio and tabulation of actual versus predicted choices will be useful. There are two possible constrained log-likelihoods. Since the model cannot contain a constant term, the constraint $\boldsymbol{\beta} = \mathbf{0}$ renders all probabilities equal to $1/J$. The constrained log-likelihood for this constraint is then $L_c = -n \ln J$. Of course, it is unlikely that this hypothesis would fail to be rejected. Alternatively, we could fit the model with only the $J-1$ choice-specific constants, which makes the constrained log-likelihood the same as in the multinomial logit model, $\ln L_0^* = \sum_j n_j \ln p_j$ where, as before, n_j is the number of individuals who choose alternative j.

21.7.3 THE INDEPENDENCE FROM IRRELEVANT ALTERNATIVES

We noted earlier that the odds ratios in the multinomial logit or conditional logit models are independent of the other alternatives. This property is convenient as regards estimation, but it is not a particularly appealing restriction to place on consumer behavior. The property of the logit model whereby P_j/P_k is independent of the remaining probabilities is called the **independence from irrelevant alternatives (IIA).**

The independence assumption follows from the initial assumption that the disturbances are independent and homoscedastic. Later we will discuss several models that have been developed to relax this assumption. Before doing so, we consider a test that has been developed for testing the validity of the assumption. Hausman and McFadden (1984) suggest that if a subset of the choice set truly is irrelevant, omitting it from the model altogether will not change parameter estimates systematically. Exclusion of these choices will be inefficient but will not lead to inconsistency. But if the remaining odds ratios are not truly independent from these alternatives, then the parameter estimates obtained when these choices are included will be inconsistent. This observation is the usual basis for Hausman's specification test. The statistic is

$$\chi^2 = (\hat{\boldsymbol{\beta}}_s - \hat{\boldsymbol{\beta}}_f)'[\hat{\mathbf{V}}_s - \hat{\mathbf{V}}_f]^{-1}(\hat{\boldsymbol{\beta}}_s - \hat{\boldsymbol{\beta}}_f),$$

where s indicates the estimators based on the restricted subset, f indicates the estimator based on the full set of choices, and $\hat{\mathbf{V}}_s$ and $\hat{\mathbf{V}}_f$ are the respective estimates of the asymptotic covariance matrices. The statistic has a limiting chi-squared distribution with K degrees of freedom.[60]

[60] McFadden (1987) shows how this hypothesis can also be tested using a Lagrange multiplier test.

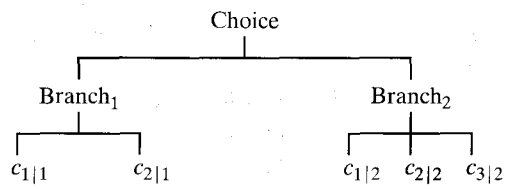
21.7.4 NESTED LOGIT MODELS

If the independence from irrelevant alternatives test fails, then an alternative to the multinomial logit model will be needed. A natural alternative is a multivariate probit model:

$$U_j = \boldsymbol{\beta}'\mathbf{x}_j + \varepsilon_j, \quad j = 1, \ldots, J, [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_J] \sim N[\mathbf{0}, \boldsymbol{\Sigma}].$$

We had considered this model earlier but found that as a general model of consumer choice, its failings were the practical difficulty of computing the multinormal integral and estimation of an unrestricted correlation matrix. Hausman and Wise (1978) point out that for a model of consumer choice, the probit model may not be as impractical as it might seem. First, for J choices, the comparisons implicit in $U_j > U_k$ for $k \neq j$ involve the $J - 1$ differences, $\varepsilon_j - \varepsilon_k$. Thus, starting with a J-dimensional problem, we need only consider derivatives of $(J - 1)$-order probabilities. Therefore, to come to a concrete example, a model with four choices requires only the evaluation of bivariate normal integrals, which, albeit still complicated to estimate, is well within the received technology. For larger models, however, other specifications have proved more useful.

One way to relax the homoscedasticity assumption in the conditional logit model that also provides an intuitively appealing structure is to group the alternatives into subgroups that allow the variance to differ across the groups while maintaining the IIA assumption within the groups. This specification defines a **nested logit model.** To fix ideas, it is useful to think of this specification as a two-(or more) level choice problem (although, once again, the model arises as a modification of the stochastic specification in the original conditional logit model, not as a model of behavior). Suppose, then, that the J alternatives can be divided into L subgroups such that the choice set can be written $[c_1, \ldots, c_J] = (c_{1|1}, \ldots, c_{J1|1}), \ldots, (c_{1|L}, \ldots, c_{JL|L})$. Logically, we may think of the choice process as that of choosing among the L choice sets and then making the specific choice within the chosen set. This method produces a tree structure, which for two branches and, say, five choices might look as follows:



Suppose as well that the data consist of observations on the attributes of the choices $\mathbf{x}_{j|l}$ and attributes of the choice sets $\mathbf{z}_l$.

To derive the mathematical form of the model, we begin with the unconditional probability

$$\text{Prob}[\text{twig}_j, \textit{branch}_l] = P_{jl} = \frac{e^{\mathbf{x}'_{j|l}\boldsymbol{\beta} + \mathbf{z}'_l\boldsymbol{\gamma}}}{\sum_{l=1}^{L}\sum_{j=1}^{Jl} e^{\mathbf{x}'_{j|l}\boldsymbol{\beta} + \mathbf{z}'_l\boldsymbol{\gamma}}}.$$

Now write this probability as

$$P_{jl} = P_{j|l} P_l = \left(\frac{e^{\mathbf{x}'_{j|l}\boldsymbol{\beta}}}{\sum_{j=1}^{J_l} e^{\mathbf{x}'_{j|l}\boldsymbol{\beta}}} \right) \left(\frac{e^{\mathbf{z}'_l\boldsymbol{\gamma}}}{\sum_{l=1}^{L} e^{\mathbf{z}'_l\boldsymbol{\gamma}}} \right) \frac{\left(\sum_{j=1}^{J_l} e^{\mathbf{x}'_{j|l}\boldsymbol{\beta}} \right) \left(\sum_{l=1}^{L} e^{\mathbf{z}'_l\boldsymbol{\gamma}} \right)}{\left(\sum_{l=1}^{L} \sum_{j=1}^{J_l} e^{\mathbf{x}'_{j|l}\boldsymbol{\beta} + \mathbf{z}'_l\boldsymbol{\gamma}} \right)}.$$

Define the **inclusive value** for the lth branch as

$$I_l = \ln \sum_{j=1}^{J_l} e^{\mathbf{x}'_{j|l}\boldsymbol{\beta}}.$$

Then, after canceling terms and using this result, we find

$$P_{j|l} = \frac{e^{\mathbf{x}'_{j|l}\boldsymbol{\beta}}}{\sum_{j=1}^{J_l} e^{\mathbf{x}'_{j|l}\boldsymbol{\beta}}} \quad \text{and} \quad P_l = \frac{e^{\mathbf{z}'_l\boldsymbol{\gamma} + \tau_l I_l}}{\sum_{l=1}^{L} e^{\mathbf{z}'_l\boldsymbol{\gamma} + \tau_l I_l}},$$

where the new parameters τ_l must equal 1 to produce the original model. Therefore. we use the restriction $\tau_l = 1$ to recover the conditional logit model, and the preceding equation just writes this model in another form. The **nested logit** model arises if this restriction is relaxed. The inclusive value coefficients, unrestricted in this fashion, allow the model to incorporate some degree of heteroscedasticity. Within each branch, the IIA restriction continues to hold. The equal variance of the disturbances within the jth branch are now

$$\sigma_j^2 = \frac{\pi^2}{6\tau_j}.^{61}$$

With $\tau_j = 1$, this reverts to the basic result for the multinomial logit model.

As usual, the coefficients in the model are not directly interpretable. The derivatives that describe covariation of the attributes and probabilities are

$$\frac{\partial \ln \text{Prob}[choice_c, branch_b]}{\partial x(k) \text{ in choice } C \text{ and branch } B} = \{\mathbf{1}(b = B)[\mathbf{1}(c = C) - P_{C|B}] + \tau_B[\mathbf{1}(b = B) - P_B]P_{C|B}\}\beta_k.$$

The nested logit model has been extended to three and higher levels. The complexity of the model increases geometrically with the number of levels. But the model has been found to be extremely flexible and is widely used for modeling consumer choice and in the marketing and transportation literatures, to name a few.

There are two ways to estimate the parameters of the nested logit model. A **limited information,** two-step maximum likelihood approach can be done as follows:

1. Estimate $\boldsymbol{\beta}$ by treating the choice within branches as a simple conditional logit model.
2. Compute the inclusive values for all the branches in the model. Estimate $\boldsymbol{\gamma}$ and the τ parameters by treating the choice among branches as a conditional logit model with attributes $\mathbf{z}_l$ and I_l.

[61] See Hensher, Louviere, and Swaite (2000).

Since this approach is a two-step estimator, the estimate of the asymptotic covariance matrix of the estimates at the second step must be corrected. [See Section 4.6, McFadden (1984), and Greene (1995a, Chapter 25).] For **full information maximum likelihood** (FIML) estimation of the model, the log-likelihood is

$$\ln L = \sum_{i=1}^{n} \ln[\text{Prob(twig} \mid \textit{branch})] \times \text{Prob}(\textit{branch})]_i.$$

The information matrix is not block diagonal in $\boldsymbol{\beta}$ and $(\boldsymbol{\gamma}, \boldsymbol{\tau})$, so FIML estimation will be more efficient than two-step estimation.

To specify the nested logit model, it is necessary to partition the choice set into branches. Sometimes there will be a natural partition, such as in the example given by Maddala (1983) when the choice of residence is made first by community, then by dwelling type within the community. In other instances, however, the partitioning of the choice set is ad hoc and leads to the troubling possibility that the results might be dependent on the branches so defined. (Many studies in this literature present several sets of results based on different specifications of the tree structure.) There is no well-defined testing procedure for discriminating among tree structures, which is a problematic aspect of the model.

21.7.5 A HETEROSCEDASTIC LOGIT MODEL

Bhat (1995) and Allenby and Ginter (1995) have developed an extension of the conditional logit model that works around the difficulty of specifying the tree for a nested model. Their model is based on the same random utility structure as before,

$$U_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij} + \varepsilon_{ij}.$$

The logit model arises from the assumption that ε_{ij} has a homoscedastic extreme value (HEV) distribution with common variance $\pi^2/6$. The authors' proposed model simply relaxes the assumption of equal variances. Since the comparisons are all pairwise, one of the variances is set to 1.0; the same comparisons of utilities will result if all equations are multiplied by the same constant, so the indeterminacy is removed by setting one of the variances to one. The model that remains, then, is exactly as before, with the additional assumption that $\text{Var}[\varepsilon_{ij}] = \sigma_j$, with $\sigma_J = 1.0$.

21.7.6 MULTINOMIAL MODELS BASED ON THE NORMAL DISTRIBUTION

A natural alternative model that relaxes the independence restrictions built into the multinomial logit (MNL) model is the **multinomial probit** (MNP) model. The structural equations of the MNP model are

$$U_j = \mathbf{x}_j'\boldsymbol{\beta}_j + \varepsilon_j, \quad j = 1, \ldots, J, [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_J] \sim N[\mathbf{0}, \boldsymbol{\Sigma}].$$

The term in the log-likelihood that corresponds to the choice of alternative q is

$$\text{Prob}[\textit{choice } q] = \text{Prob}[U_q > U_j, j = 1, \ldots, J, j \neq q].$$

The probability for this occurrence is

$$\text{Prob}[\textit{choice } q] = \text{Prob}[\varepsilon_1 - \varepsilon_q > (\mathbf{x}_q - \mathbf{x}_1)'\boldsymbol{\beta}, \ldots, \varepsilon_J - \varepsilon_q > (\mathbf{x}_q - \mathbf{x}_J)'\boldsymbol{\beta}]$$

for the $J-1$ other choices, which is a cumulative probability from a $(J-1)$-variate normal distribution. As in the HEV model, since we are only making comparisons, one of the variances in this $J-1$ variate structure—that is, one of the diagonal elements in the reduced $\boldsymbol{\Sigma}$—must be normalized to 1.0. Since only comparisons are ever observable in this model, for identification, $J-1$ of the covariances must also be normalized, to zero. The MNP model allows an unrestricted $(J-1) \times (J-1)$ correlation structure and $J-2$ free standard deviations for the disturbances in the model. (Thus, a two choice model returns to the univariate probit model of Section 21.2.) For more than two choices, this specification is far more general than the MNL model, which assumes that $\boldsymbol{\Sigma} = \mathbf{I}$. (The scaling is absorbed in the coefficient vector in the MNL model.)

The main obstacle to implementation of the MNP model has been the difficulty in computing the multivariate normal probabilities for any dimensionality higher than 2. Recent results on accurate simulation of multinormal integrals, however, have made estimation of the MNP model feasible. (See Section E.5.6 and a symposium in the November 1994 issue of the *Review of Economics and Statistics*.) Yet some practical problems remain. Computation is exceedingly time consuming. It is also necessary to ensure that $\boldsymbol{\Sigma}$ remain a positive definite matrix. One way often suggested is to construct the Cholesky decomposition of $\boldsymbol{\Sigma}$, $\mathbf{LL}'$, where $\mathbf{L}$ is a lower triangular matrix, and estimate the elements of $\mathbf{L}$. Maintaining the normalizations and zero restrictions will still be cumbersome, however. An alternative is estimate the correlations, $\mathbf{R}$, and a diagonal matrix of standard deviations, $\mathbf{S} = \text{diag}(\sigma_1, \ldots, \sigma_{J-2}, 1, 1)$ separately. The normalizations, $\mathbf{R}_{jj} = 1$, and exclusions, $\mathbf{R}_{jl} = 0$, are simple to impose, and $\boldsymbol{\Sigma}$ is just $\mathbf{SRS}$. $\mathbf{R}$ is otherwise restricted only in that $-1 < \mathbf{R}_{jl} < +1$. The resulting matrix must be positive definite. Identification appears to be a serious problem with the MNP model. Although the unrestricted MNP model is fully identified in principle, convergence to satisfactory results in applications with more than three choices appears to require many additional restrictions on the standard deviations and correlations, such as zero restrictions or equality restrictions in the case of the standard deviations.

21.7.7 A RANDOM PARAMETERS MODEL

Another variant of the multinomial logit model is the random parameters logit (RPL) model (also called the "mixed logit model"). [See Revelt and Train (1996); Bhat (1996); Berry, Levinsohn, and Pakes (1995); and Jain, Vilcassim, and Chintagunta (1994).] Train's formulation of the RPL model (which encompasses the others) is a modification of the MNL model. The model is a **random coefficients** formulation. The change to the basic MNL model is the parameter specification in the distribution of the parameters across individuals, i;

$$\beta_{ik} = \beta_k + \mathbf{z}_i'\boldsymbol{\theta}_k + \sigma_k u_{ik},$$

where u_{ik} is normally distributed with correlation matrix $\mathbf{R}$, σ_k is the standard deviation of the distribution, $\beta_k + \mathbf{z}_i'\boldsymbol{\theta}_k$ is the mean of the distribution, and $\mathbf{z}_i$ is a vector of person specific characteristics (such as age and income) that do not vary across choices. This formulation contains all the earlier models. For example, if $\boldsymbol{\theta}_k = \mathbf{0}$ for all the coefficients and $\sigma_k = 0$ for all the coefficients except for choice specific constants, then the original MNL model with a normal-logistic mixture for the random part of the MNL model arises (hence the name).

The authors propose estimation of the model by simulating the log-likelihood function rather than direct integration to compute the probabilities, which would be infeasible because the mixture distribution composed of the original ε_{ij} and the random part of the coefficient is unknown. For any individual,

$$\text{Prob}[\textit{choice } q \mid \mathbf{u}_i] = \text{MNL probability} \mid \boldsymbol{\beta}_i(\mathbf{u}_i),$$

with all restrictions imposed on the coefficients. The appropriate probability is

$$E_u[\text{Prob}(\textit{choice } q \mid \mathbf{u})] = \int_{u_1,\ldots,u_k} \text{Prob}[\textit{choice } q \mid \mathbf{u}] f(\mathbf{u})\, d\mathbf{u},$$

which can be estimated by simulation, using

$$\text{Est.}E_u[\text{Prob}(\textit{choice } q \mid \mathbf{u})] = \frac{1}{R}\sum_{r=1}^{R} \text{Prob}[\textit{choice } q \mid \hat{\boldsymbol{\beta}}_i(\boldsymbol{e}_{ir})]$$

where $\boldsymbol{e}_{ir}$ is the rth of R draws for observation i. (There are nkR draws in total. The draws for observation i must be the same from one computation to the next, which can be accomplished by assigning to each individual their own seed for the random number generator and restarting it each time the probability is to be computed.) By this method, the log-likelihood and its derivatives with respect to $(\beta_k, \boldsymbol{\theta}_k, \sigma_k)$, $k = 1, \ldots, K$ and $\mathbf{R}$ are simulated to find the values that maximize the simulated log-likelihood. This is precisely the approach we used in Example 17.10.

The RPL model enjoys a considerable advantage not available in any of the other forms suggested. In a panel data setting, one can formulate a random effects model simply by making the variation in the coefficients time invariant. Thus, the model is changed to

$$U_{ijt} = \mathbf{x}_{ijt}'\boldsymbol{\beta}_{ijt} + \varepsilon_{ijt}, \quad i = 1, \ldots, n,\ j = 1, \ldots, J, t = 1, \ldots T$$
$$\beta_{ijt,k} = \beta_k + \mathbf{z}_{it}'\boldsymbol{\theta}_{ik} + \sigma_k u_{ik},$$

The time variation in the coefficients is provided by the choice invariant variables which may change through time. Habit persistence is carried by the time invariant random effect, u_{ik}. If only the constant terms vary and they are assumed to be uncorrelated, then this is logically equivalent to the familiar random effects model. But, much greater generality can be achieved by allowing the other coefficients to vary randomly across individuals and by allowing correlation of these effects.[62]

21.7.8 APPLICATION: CONDITIONAL LOGIT MODEL FOR TRAVEL MODE CHOICE

Hensher and Greene [Greene (1995a)] report estimates of a model of travel mode choice for travel between Sydney and Melbourne, Australia. The data set contains 210 observations on choice among four travel modes, *air, train, bus,* and *car.* (See Appendix Table F21.2.) The attributes used for their example were: choice-specific constants; two choice-specific continuous measures; GC, a measure of the generalized cost of the travel that is equal to the sum of in-vehicle cost, INVC and a wagelike measure

[62] See Hensher (2001) for an application to transportation mode choice in which each individual is observed in several choice situations.

TABLE 21.10 Summary Statistics for Travel Mode Choice Data

	GC	*TTME*	*INVC*	*INVT*	*HINC*	*Number Choosing*	*p*	*True prop.*
Air	102.648	61.010	85.522	133.710	34.548	58	0.28	0.14
	113.522	46.534	97.569	124.828	41.274			
Train	130.200	35.690	51.338	608.286	34.548	63	0.30	0.13
	106.619	28.524	37.460	532.667	23.063			
Bus	115.257	41.650	33.457	629.462	34.548	30	0.14	0.09
	108.133	25.200	33.733	618.833	29.700			
Car	94.414	0	20.995	573.205	34.548	59	0.28	0.64
	89.095	0	15.694	527.373	42.220			

Note: The upper figure is the average for all 210 observations. The lower figure is the mean for the observations that made that choice.

times INVT, the amount of time spent traveling; and TTME, the terminal time (zero for car); and for the choice between air and the other modes, HINC, the household income. A summary of the sample data is given in Table 21.10. The sample is **choice based** so as to balance it among the four choices—the true population allocation, as shown in the last column of Table 21.10, is dominated by drivers.

The model specified is

$$U_{ij} = \alpha_{air} d_{i,air} + \alpha_{train} d_{i,train} + \alpha_{bus} d_{i,bus} + \beta_G GC_{ij} + \beta_T TTME_{ij} + \gamma_H d_{i,air} HINC_i + \varepsilon_{ij}.$$

where for each j, ε_{ij} has the same independent, type 1 extreme value distribution,

$$F_\varepsilon(\varepsilon_{ij}) = \exp(-\exp(-\varepsilon_{ij}))$$

which has standard deviation $\pi^2/6$. The mean is absorbed in the constants. Estimates of the conditional logit model are shown in Table 21.11. The model was fit with and without the corrections for choice based sampling. Since the sample shares do not differ radically from the population proportions, the effect on the estimated parameters is fairly modest. Nonetheless, it is apparent that the choice based sampling is not completely innocent. A cross tabulation of the predicted versus actual outcomes is given in Table 21.12. The predictions are generated by tabulating the integer parts of $m_{jk} = \sum_{i=1}^{210} \hat{p}_{ij} d_{ik}$.

TABLE 21.11 Parameter Estimates (t Values in Parentheses)

	Unweighted Sample		*Choice Based Weighting*	
	Estimate	*t Ratio*	*Estimate*	*t Ratio*
β_G	−0.15501	−3.517	−0.01333	−2.724
β_T	−0.19612	−9.207	−0.13405	−7.164
γ_H	0.01329	1.295	−0.00108	−0.087
α_{air}	5.2074	6.684	6.5940	5.906
α_{train}	3.8690	8.731	3.6190	7.447
α_{bus}	3.1632	7.025	3.3218	5.698
Log likelihood at $\beta = 0$	−291.1218		−291.1218	
Log likelihood (sample shares)	−283.7588		−223.0578	
Log likelihood at convergence	−199.1284		−147.5896	

TABLE 21.12 Predicted Choices Based on Model Probabilities (Predictions Based on Choice Based Sampling are in Parentheses.)

	Air	*Train*	*Bus*	*Car*	*Total (Actual)*
Air	32 (30)	8 (3)	5 (3)	13 (23)	58
Train	7 (3)	37 (30)	5 (3)	14 (27)	63
Bus	3 (1)	5 (2)	15 (4)	6 (12)	30
Car	16 (5)	13 (5)	6 (3)	25 (45)	59
Total (Predicted)	58 (39)	63 (40)	30 (23)	59 (108)	210

$j, k = air, train, bus, car,$ where $\hat{p}_{ij}$ is the predicted probability of outcome j for observation i and d_{ik} is the binary variable which indicates if individual i made choice k.

Are the odds ratios *train/bus* and *car/bus* really independent from the presence of the *air* alternative? To use the Hausman test, we would eliminate choice *air,* from the choice set and estimate a three-choice model. Since 58 respondents chose this mode, we would lose 58 observations. In addition, for every data vector left in the sample, the air specific constant and the interaction, $d_{i,air} \times HINC_i$ would be zero for every remaining individual. Thus, these parameters could not be estimated in the restricted model. We would drop these variables. The test would be based on the two estimators of the remaining four coefficients in the model, $[\beta_G, \beta_T, \alpha_{train}, \alpha_{bus}]$. The results for the test are as shown in Table 21.13.

The hypothesis that the odds ratios for the other three choices are independent from *air* would be rejected based on these results, as the chi-squared statistic exceeds the critical value.

Since IIA was rejected, they estimated a nested logit model of the following type:

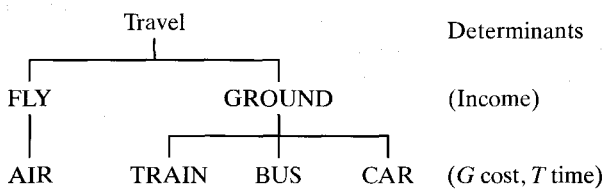


TABLE 21.13 Results for IIA Test

	Full Choice Set				*Restricted Choice Set*			
	β_G	β_T	α_{train}	α_{bus}	β_G	β_T	α_{train}	α_{bus}
Estimate	−0.0155	−0.0961	3.869	3.163	−0.0639	−0.0699	4.464	3.105
	Estimated Asymptotic Covariance Matrix				***Estimated Asymptotic Covariance Matrix***			
β_G	0.194e-5				0.000101			
β_T	−0.46e-7	0.000110			−0.0000013	0.000221		
α_{train}	−0.00060	−0.0038	0.196		−0.000244	−0.00759	0.410	
α_{bus}	−0.00026	−0.0037	0.161	0.203	−0.000113	−0.00753	0.336	0.371

Note: 0.nnne-p indicates times 10 to the negative p power.
$H = 33.3363$. Critical chi-squared[4] = 9.488.

TABLE 21.14 Estimates of a Mode Choice Model (Standard Errors in Parentheses)

Parameter	*FIML Estimate*		*LIML Estimate*		*Unconditional*	
α_{air}	6.042	(1.199)	−0.0647	(2.1485)	5.207	(0.779)
α_{bus}	4.096	(0.615)	3.105	(0.609)	3.163	(0.450)
α_{train}	5.065	(0.662)	4.464	(0.641)	3.869	(0.443)
β_{GC}	−0.03159	(0.00816)	−0.06368	(0.0100)	−0.1550	(0.00441)
β_{TTME}	−0.1126	(0.0141)	−0.0699	(0.0149)	−0.09612	(0.0104)
γ_H	0.01533	(0.00938)	0.02079	(0.01128)	0.01329	(0.0103)
τ_{flyu}	0.5860	(0.141)	0.2266	(0.296)	1.0000	(0.000)
τ_{ground}	0.3890	(0.124)	0.1587	(0.262)	1.0000	(0.000)
σ_{fly}	2.1886	(0.525)	5.675	(2.350)	1.2825	(0.000)
σ_{ground}	3.2974	(1.048)	8.081	(4.219)	1.2825	(0.000)
log L	−193.6561		−115.3354 + (−87.9382)		−199.1284	

Note that one of the branches has only a single choice, so the conditional probability, $P_{j|fly} = P_{air|fly} = 1$. The model is fit by both FIML and LIML methods. Three sets of estimates are shown in Table 21.14. The set marked "unconditional" are the simple conditional (multinomial) logit (MNL) model for choice among the four alternatives that was reported earlier. Both inclusive value parameters are constrained (by construction) to equal 1.0000. The FIML estimates are obtained by maximizing the full log likelihood for the nested logit model. In this model,

$$\text{Prob}(choice \mid branch) = P(\alpha_{air}d_{air} + \alpha_{train}d_{train} + \alpha_{bus}d_{bus} + \beta_G GC + \beta_T TTME),$$

$$\text{Prob}(branch) = P(\gamma d_{air} HINC + \tau_{fly} IV_{fly} + \tau_{ground} IV_{ground}),$$

$$\text{Prob}(choice, branch) = \text{Prob}(choice \mid branch) \times \text{Prob}(branch).$$

Finally, the limited information estimator is estimated in two steps. At the first step, a choice model is estimated for the three choices in the ground branch:

$$\text{Prob}(choice \mid ground) = P(\alpha_{train}d_{train} + \alpha_{bus}d_{bus} + \beta_G GC + \beta_T TTME)$$

This model uses only the observations that chose one of the three ground modes; for these data, this subset was 152 of the 210 observations. Using the estimates from this model, we compute, for all 210 observations, $IV_{fly} = \log[\exp(\mathbf{z}'_{air}\boldsymbol{\beta})]$ for *air* and 0 for ground, and $IV_{ground} = \log[\sum_{j=ground} \exp(\mathbf{z}'_j\boldsymbol{\beta})]$ for ground modes and 0 for *air*. Then, the choice model

$$\text{Prob}(branch) = P(\alpha_{air}d_{air} + \gamma_H d_{air} HINC + \tau_{fly} IV_{fly} + \tau_{ground} IV_{ground})$$

is fit separately. Since the Hessian is not block diagonal, the FIML estimator is more efficient. To obtain appropriate standard errors, we must make the Murphy and Topel correction for two-step estimation; see Section 17.7 and Theorem 17.8. It is simplified a bit here because different samples are used for the two steps. As such, the matrix $\mathbf{R}$ in the theorem is not computed. To compute $\mathbf{C}$, we require the matrix of derivatives of log Prob(*branch*) with respect to the direct parameters, α_{air}, γ_H, τ_{fly}, τ_{ground}, and with respect to the choice parameters, $\boldsymbol{\beta}$. Since this model is a simple binomial (two choice) logit model, these are easy to compute, using (21-19). Then the corrected asymptotic covariance matrix is computed using Theorem 17.8 with $\mathbf{R} = \mathbf{0}$.

TABLE 21.15 Estimates of a Heteroscedastic Extreme Value Model (Standard Errors in Parentheses)

Parameter	*HEV Estimate*		*Nested Logit Estimate*		*Restricted HEV*	
α_{air}	7.8326	(10.951)	6.062	(1.199)	2.973	(0.995)
α_{bus}	7.1718	(9.135)	4.096	(0.615)	4.050	(0.494)
α_{train}	6.8655	(8.829)	5.065	(0.662)	3.042	(0.429)
β_{GC}	−0.05156	(0.0694)	−0.03159	(0.00816)	−0.0289	(0.00580)
β_{TTME}	−0.1968	(0.288)	−0.1126	(0.0141)	−0.0828	(0.00576)
γ	0.04024	(0.0607)	0.01533	(0.00938)	0.0238	(0.0186)
τ_{fly}		—	0.5860	(0.141)		—
τ_{ground}		—	0.3890	(0.124)		—
θ_{air}	0.2485	(0.369)			0.4959	(0.124)
θ_{train}	0.2595	(0.418)			1.0000	(0.000)
θ_{bus}	0.6065	(1.040)			1.0000	(0.000)
θ_{car}	1.0000	(0.000)			1.0000	(0.000)
Implied Standard Deviations						
σ_{air}	5.161	(7.667)				
σ_{train}	4.942	(7.978)				
σ_{bus}	2.115	(3.623)				
σ_{car}	1.283	(0.000)				
$\ln L$	−195.6605		−193.6561		−200.3791	

The likelihood ratio statistic for the nesting (heteroscedasticity) against the null hypothesis of homoscedasticity is $-2[-199.1284-(-193.6561)] = 10.945$. The 95 percent critical value from the chi-squared distribution with two degrees of freedom is 5.99, so the hypothesis is rejected. We can also carry out a Wald test. The asymptotic covariance matrix for the two inclusive value parameters is [0.01977/0.009621, 0.01529]. The Wald statistic for the joint test of the hypothesis that $\tau_{fly} = \tau_{ground} = 1$, is

$$W = (0.586 - 1.0 \quad 0.389 - 1.0) \begin{bmatrix} 0.1977 & 0.009621 \\ 0.009621 & 0.01529 \end{bmatrix}^{-1} \begin{pmatrix} 0.586 - 1.0 \\ 0.389 - 1.0 \end{pmatrix} = 24.475$$

The hypothesis is rejected, once again.

The nested logit model was reestimated under assumptions of the heteroscedastic extreme value model. The results are shown in Table 21.15. This model is less restrictive than the nested logit model. To make them comparable, we note that we found that $\sigma_{air} = \pi/(\tau_{air}\sqrt{6}) = 2.1886$ and $\sigma_{train} = \sigma_{bus} = \sigma_{car} = \pi/(\tau_{ground}\sqrt{6}) = 3.2974$. The heteroscedastic extreme value (HEV) model thus relaxes one variance restriction, because it has three free variance parameters instead of two. On the other hand, the important degree of freedom here is that the HEV model does not impose the IIA assumption anywhere in the choice set, whereas the nested logit does, within each branch.

A primary virtue of the HEV model, the nested logit model, and other alternative models is that they relax the IIA assumption. This assumption has implications for the cross elasticities between attributes in the different probabilities. Table 21.16 lists the estimated elasticities of the estimated probabilities with respect to changes in the generalized cost variable. Elasticities are computed by averaging the individual sample values rather than computing them once at the sample means. The implication of the IIA

TABLE 21.16 Estimated Elasticities with Respect to Generalized Cost

	Cost Is That of Alternative			
Effect on	***Air***	***Train***	***Bus***	***Car***
Multinomial Logit				
Air	−1.136	0.498	0.238	0.418
Train	0.456	−1.520	0.238	0.418
Bus	0.456	0.498	−1.549	0.418
Car	0.456	0.498	0.238	−1.061
Nested Logit				
Air	−0.858	0.332	0.179	0.308
Train	0.314	−4.075	0.887	1.657
Bus	0.314	1.595	−4.132	1.657
Car	0.314	1.595	0.887	−2.498
Heteroscedastic Extreme Value				
Air	−1.040	0.367	0.221	0.441
Train	0.272	−1.495	0.250	0.553
Bus	0.688	0.858	−6.562	3.384
Car	0.690	0.930	1.254	−2.717

assumption can be seen in the table entries. Thus, in the estimates for the multinomial logit (MNL) model, the cross elasticities for each attribute are all equal. In the nested logit model, the IIA property only holds within the branch. Thus, in the first column, the effect of GC of air affects all ground modes equally, whereas the effect of GC for train is the same for bus and car but different from these two for air. All these elasticities vary freely in the HEV model.

Table 21.17 lists the estimates of the parameters of the multinomial probit and random parameters logit models. For the multinomial probit model, we fit three specifications: (1) free correlations among the choices, which implies an unrestricted 3×3 correlation matrix and two free standard deviations; (2) uncorrelated disturbances, but free standard deviations, a model that parallels the heteroscedastic extreme value model; and (3) uncorrelated disturbances and equal standard deviations, a model that is the same as the original conditional logit model save for the normal distribution of the disturbances instead of the extreme value assumed in the logit model. In this case, the scaling of the utility functions is different by a factor of $(\pi^2/6)^{1/2} = 1.283$, as the probit model assumes ε_j has a standard deviation of 1.0.

We also fit three variants of the random parameters logit. In these cases, the choice specific variance for each utility function is $\sigma_j^2 + \theta_j^2$ where σ_j^2 is the contribution of the logit model, which is $\pi^2/6 = 1.645$, and θ_j^2 is the estimated constant specific variance estimated in the random parameters model. The combined estimated standard deviations are given in the table. The estimates of the specific parameters, θ_j are given in the footnotes. The estimated models are: (1) unrestricted variation and correlation among the three intercept parameters—this parallels the general specification of the multinomial probit model; (2) only the constant terms randomly distributed but uncorrelated, a model that is parallel to the multinomial probit model with no cross equation correlation and to the heteroscedastic extreme value model shown in Table 21.15:

TABLE 21.17 Parameter Estimates for Normal Based Multinomial Choice Models

	Multinomial Probit			*Random Parameters Logit*		
Parameter	*Unrestricted*	*Homoscedastic*	*Uncorrelated*	*Unrestricted*	*Constants*	*Uncorrelated*
α_{air}	1.358	3.005	3.171	5.519	4.807	12.603
σ_{air}	4.940	1.000[a]	3.629	4.009[d]	3.225[b]	2.803[c]
α_{train}	4.298	2.409	4.277	5.776	5.035	13.504
σ_{train}	1.899	1.000[a]	1.581	1.904	1.290[b]	1.373
α_{bus}	3.609	1.834	3.533	4.813	4.062	11.962
σ_{bus}	1.000[a]	1.000[a]	1.000[a]	1.424	3.147[b]	1.287
α_{car}	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000
σ_{car}	1.000[a]	1.000	1.000[a]	1.283[a]	1.283[a]	1.283[a]
β_G	−0.0351	−0.0113	−0.0325	−0.0326	−0.0317	−0.0544
$\sigma_{\beta G}$	—	—	—	0.000[a]	0.000[a]	0.00561
β_T	−0.0769	−0.0563	−0.0918	−0.126	−0.112	−0.2822
$\sigma_{\beta T}$	—	—	—	0.000[a]	0.000[a]	0.182
γ_H	0.0593	0.0126	0.0370	0.0334	0.0319	0.0846
σ_γ	—	—	—	0.000[a]	0.000[a]	0.0768
ρ_{AT}	0.581	0.000[a]	0.000[a]	0.543	0.000[a]	0.000[a]
ρ_{AB}	0.576	0.000[a]	0.000[a]	0.532	0.000[a]	0.000[a]
ρ_{BT}	0.718	0.000[a]	0.000[a]	0.993	0.000[a]	0.000[a]
log L	−196.9244	−208.9181	−199.7623	−193.7160	−199.0073	−175.5333

[a]Restricted to this fixed value.
[b]Computed as the square root of $(\pi^2/6 + \theta_j^2)$, $\theta_{air} = 2.959$, $\theta_{train} = 0.136$, $\theta_{bus} = 0.183$, $\theta_{car} = 0.000$.
[c]$\theta_{air} = 2.492$, $\theta_{train} = 0.489$, $\theta_{bus} = 0.108$, $\theta_{car} = 0.000$.
[d]Derived standard deviations for the random constants are $\theta_{air} = 3.798$, $\theta_{train} = 1.182$, $\theta_{bus} = 0.0712$, $\theta_{car} = 0.000$.

(3) random but uncorrelated parameters. This model is more general than the others, but is somewhat restricted as the parameters are assumed to be uncorrelated. Identification of the correlation model is weak in this model—after all, we are attempting to estimate a 6 × 6 correlation matrix for all unobserved variables. Only the estimated parameters are shown in Table 21.17. Estimated standard errors are similar to (although generally somewhat larger than) those for the basic multinomial logit model.

The standard deviations and correlations shown for the multinomial probit model are parameters of the distribution of ε_{ij}, the overall randomness in the model. The counterparts in the random parameters model apply to the distributions of the parameters. Thus, the full disturbance in the model in which only the constants are random is $\varepsilon_{iair} + u_{air}$ for air, and likewise for train and bus. Likewise, the correlations shown for the first two models are directly comparable, though it should be noted that in the random parameters model, the disturbances have a distribution that is that of a sum of an extreme value and a normal variable, while in the probit model, the disturbances are normally distributed. With these considerations, the "unrestricted" models in each case are comparable and are, in fact, fairly similar.

None of this discussion suggests a preference for one model or the other. The likelihood values are not comparable, so a direct test is precluded. Both relax the IIA assumption, which is a crucial consideration. The random parameters model enjoys a significant practical advantage, as discussed earlier, and also allows a much richer specification of the utility function itself. But, he question still warrants additional study. Both models are making their way into the applied literature.

21.8 ORDERED DATA

Some multinomial-choice variables are inherently ordered. Examples that have appeared in the literature include the following:

1. Bond ratings
2. Results of taste tests
3. Opinion surveys
4. The assignment of military personnel to job classifications by skill level
5. Voting outcomes on certain programs
6. The level of insurance coverage taken by a consumer: none, part, or full
7. Employment: unemployed, part time, or full time

In each of these cases, although the outcome is discrete, the multinomial logit or probit model would fail to account for the ordinal nature of the dependent variable.[63] Ordinary regression analysis would err in the opposite direction, however. Take the outcome of an opinion survey. If the responses are coded 0, 1, 2, 3, or 4, then linear regression would treat the difference between a 4 and a 3 the same as that between a 3 and a 2, whereas in fact they are only a ranking.

The ordered probit and logit models have come into fairly wide use as a framework for analyzing such responses (Zavoina and McElvey, 1975). The model is built around a latent regression in the same manner as the binomial probit model. We begin with

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

As usual, y^* is unobserved. What we do observe is

$$\begin{aligned} y &= 0 \quad \text{if } y^* \leq 0, \\ &= 1 \quad \text{if } 0 < y^* \leq \mu_1, \\ &= 2 \quad \text{if } \mu_1 < y^* \leq \mu_2, \\ &\vdots \\ &= J \quad \text{if } \mu_{J-1} \leq y^*, \end{aligned}$$

which is a form of censoring. The μs are unknown parameters to be estimated with $\boldsymbol{\beta}$. Consider, for example, an opinion survey. The respondents have their own intensity of feelings, which depends on certain measurable factors $\mathbf{x}$ and certain unobservable factors ε. In principle, they could respond to the questionnaire with their own y^* if asked to do so. Given only, say, five possible answers, they choose the cell that most closely represents their own feelings on the question.

[63] In two papers, Beggs, Cardell, and Hausman (1981) and Hausman and Ruud (1986), the authors analyze a richer specification of the logit model when respondents provide their rankings of the full set of alternatives in addition to the identity of the most preferred choice. This application falls somewhere between the conditional logit model and the ones we shall discuss here in that, rather than provide a single choice among J either unordered or ordered alternatives, the consumer chooses one of the $J!$ possible orderings of the set of unordered alternatives.

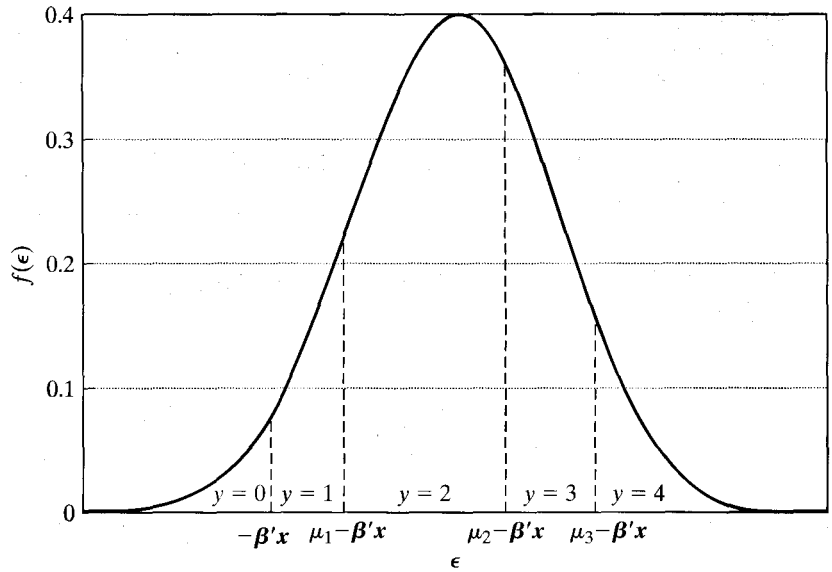


FIGURE 21.4 Probabilities in the Ordered Probit Model.

As before, we assume that ε is normally distributed across observations.[64] For the same reasons as in the binomial probit model (which is the special case of $J = 1$), we normalize the mean and variance of ε to zero and one. We then have the following probabilities:

$$\begin{aligned}
\text{Prob}(y = 0 \mid \mathbf{x}) &= \Phi(-\mathbf{x}'\boldsymbol{\beta}), \\
\text{Prob}(y = 1 \mid \mathbf{x}) &= \Phi(\mu_1 - \mathbf{x}'\boldsymbol{\beta}) - \Phi(-\mathbf{x}'\boldsymbol{\beta}), \\
\text{Prob}(y = 2 \mid \mathbf{x}) &= \Phi(\mu_2 - \mathbf{x}'\boldsymbol{\beta}) - \Phi(\mu_1 - \mathbf{x}'\boldsymbol{\beta}), \\
&\vdots \\
\text{Prob}(y = J \mid \mathbf{x}) &= 1 - \Phi(\mu_{J-1} - \mathbf{x}'\boldsymbol{\beta}).
\end{aligned}$$

For all the probabilities to be positive, we must have

$$0 < \mu_1 < \mu_2 < \cdots < \mu_{J-1}.$$

Figure 21.4 shows the implications of the structure. This is an extension of the univariate probit model we examined earlier. The log-likelihood function and its derivatives can be obtained readily, and optimization can be done by the usual means.

As usual, the marginal effects of the regressors $\mathbf{x}$ on the probabilities are not equal to the coefficients. It is helpful to consider a simple example. Suppose there are three categories. The model thus has only one unknown threshold parameter. The three

[64] Other distributions, particularly the logistic, could be used just as easily. We assume the normal purely for convenience. The logistic and normal distributions generally give similar results in practice.

probabilities are

$$\text{Prob}(y = 0 \mid \mathbf{x}) = 1 - \Phi(\mathbf{x}'\boldsymbol{\beta}),$$
$$\text{Prob}(y = 1 \mid \mathbf{x}) = \Phi(\mu - \mathbf{x}'\boldsymbol{\beta}) - \Phi(-\mathbf{x}'\boldsymbol{\beta}),$$
$$\text{Prob}(y = 2 \mid \mathbf{x}) = 1 - \Phi(\mu - \mathbf{x}'\boldsymbol{\beta}).$$

For the three probabilities, the marginal effects of changes in the regressors are

$$\frac{\partial \text{Prob}(y = 0 \mid \mathbf{x})}{\partial \mathbf{x}} = -\phi(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta},$$
$$\frac{\partial \text{Prob}(y = 1 \mid \mathbf{x})}{\partial \mathbf{x}} = [\phi(-\mathbf{x}'\boldsymbol{\beta}) - \phi(\mu - \mathbf{x}'\boldsymbol{\beta})]\boldsymbol{\beta},$$
$$\frac{\partial \text{Prob}(y = 2 \mid \mathbf{x})}{\partial \mathbf{x}} = \phi(\mu - \mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}.$$

Figure 21.5 illustrates the effect. The probability distributions of y and y^* are shown in the solid curve. Increasing one of the x's while holding $\boldsymbol{\beta}$ and μ constant is equivalent to shifting the distribution slightly to the right, which is shown as the dashed curve. The effect of the shift is unambiguously to shift some mass out of the leftmost cell. Assuming that $\boldsymbol{\beta}$ is positive (for this x), $\text{Prob}(y = 0 \mid \mathbf{x})$ must decline. Alternatively, from the previous expression, it is obvious that the derivative of $\text{Prob}(y = 0 \mid \mathbf{x})$ has the opposite sign from $\boldsymbol{\beta}$. By a similar logic, the change in $\text{Prob}(y = 2 \mid \mathbf{x})$ [or $\text{Prob}(y = J \mid \mathbf{x})$ in the general case] must have the same sign as $\boldsymbol{\beta}$. Assuming that the particular $\boldsymbol{\beta}$ is positive, we are shifting some probability into the rightmost cell. But what happens to the middle cell is ambiguous. It depends on the two densities. In the general case, relative to the signs of the coefficients, only the signs of the changes in $\text{Prob}(y = 0 \mid \mathbf{x})$ and $\text{Prob}(y = J \mid \mathbf{x})$ are unambiguous! The upshot is that we must be very careful

FIGURE 21.5 Effects of Change in x on Predicted Probabilities.

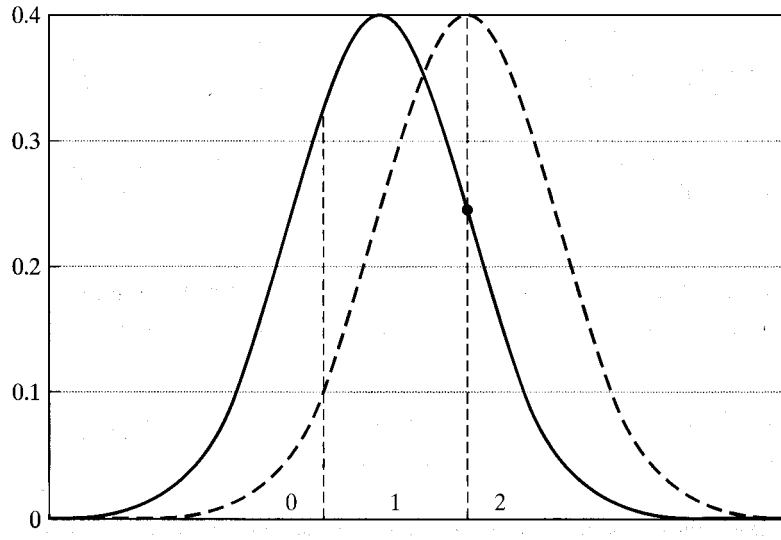


TABLE 21.18 Estimated Rating Assignment Equation

Variable	*Estimate*	*t Ratio*	*Mean of Variable*
Constant	−4.34	—	—
ENSPA	0.057	1.7	0.66
EDMA	0.007	0.8	12.1
AFQT	0.039	39.9	71.2
EDYRS	0.190	8.7	12.1
MARR	−0.48	−9.0	0.08
AGEAT	0.0015	0.1	18.8
μ	1.79	80.8	—

in interpreting the coefficients in this model. Indeed, without a fair amount of extra calculation, it is quite unclear how the coefficients in the ordered probit model should be interpreted.[65]

Example 21.11 Rating Assignments

Marcus and Greene (1985) estimated an ordered probit model for the job assignments of new Navy recruits. The Navy attempts to direct recruits into job classifications in which they will be most productive. The broad classifications the authors analyzed were technical jobs with three clearly ranked skill ratings: "medium skilled," "highly skilled," and "nuclear qualified/highly skilled." Since the assignment is partly based on the Navy's own assessment and needs and partly on factors specific to the individual, an ordered probit model was used with the following determinants: (1) ENSPE = a dummy variable indicating that the individual entered the Navy with an "A school" (technical training) guarantee, (2) EDMA = educational level of the entrant's mother, (3) AFQT = score on the Air Force Qualifying Test, (4) EDYRS = years of education completed by the trainee, (5) MARR = a dummy variable indicating that the individual was married at the time of enlistment, and (6) AGEAT = trainee's age at the time of enlistment. The sample size was 5,641. The results are reported in Table 21.18. The extremely large t ratio on the AFQT score is to be expected, since it is a primary sorting device used to assign job classifications.

To obtain the marginal effects of the continuous variables, we require the standard normal density evaluated at $-\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}} = -0.8479$ and $\hat{\mu} - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}} = 0.9421$. The predicted probabilities are $\Phi(-0.8479) = 0.198$, $\Phi(0.9421) - \Phi(-0.8479) = 0.628$, and $1 - \Phi(0.9421) = 0.174$. (The actual frequencies were 0.25, 0.52, and 0.23.) The two densities are $\phi(-0.8479) = 0.278$ and $\phi(0.9421) = 0.255$. Therefore, the derivatives of the three probabilities with respect to AFQT, for example, are

$$\frac{\partial P_0}{\partial \text{AFQT}} = (-0.278)0.039 = -0.01084,$$

$$\frac{\partial P_1}{\partial \text{AFQT}} = (0.278 - 0.255)0.039 = 0.0009,$$

$$\frac{\partial P_2}{\partial \text{AFQT}} = 0.255(0.039) = 0.00995.$$

[65] This point seems uniformly to be overlooked in the received literature. Authors often report coefficients and t ratios, occasionally with some commentary about significant effects, but rarely suggest upon what or in what direction those effects are exerted.

TABLE 21.19 Marginal Effect of a Binary Variable

	$-\hat{\beta}'x$	$\hat{\mu} - \hat{\beta}'x$	$Prob[y = 0]$	$Prob[y = 1]$	$Prob[y = 2]$
MARR = 0	−0.8863	0.9037	0.187	0.629	0.184
MARR = 1	−0.4063	1.3837	0.342	0.574	0.084
Change			0.155	−0.055	−0.100

Note that the marginal effects sum to zero, which follows from the requirement that the probabilities add to one. This approach is not appropriate for evaluating the effect of a dummy variable. We can analyze a dummy variable by comparing the probabilities that result when the variable takes its two different values with those that occur with the other variables held at their sample means. For example, for the MARR variable, we have the results given in Table 21.19.

21.9 MODELS FOR COUNT DATA

Data on patents suggested in Section 21.2 are typical of **count data.** In principle, we could analyze these data using multiple linear regression. But the preponderance of zeros and the small values and clearly discrete nature of the dependent variable suggest that we can improve on least squares and the linear model with a specification that accounts for these characteristics. The **Poisson regression model** has been widely used to study such data.[66]

The Poisson regression model specifies that each y_i is drawn from a Poisson distribution with parameter λ_i, which is related to the regressors $\mathbf{x}_i$. The primary equation of the model is

$$\text{Prob}(Y_i = y_i \mid \mathbf{x}_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots.$$

The most common formulation for λ_i is the **loglinear model,**

$$\ln \lambda_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

It is easily shown that the expected number of events *per period* is given by

$$E[y_i \mid \mathbf{x}_i] = \text{Var}[y_i \mid \mathbf{x}_i] = \lambda_i = e^{\mathbf{x}_i'\boldsymbol{\beta}},$$

so

$$\frac{\partial E[y_i \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \lambda_i \boldsymbol{\beta}.$$

With the parameter estimates in hand, this vector can be computed using any data vector desired.

In principle, the Poisson model is simply a nonlinear regression.[67] But it is far easier to estimate the parameters with maximum likelihood techniques. The log-likelihood

[66]There are several recent surveys of specification and estimation of models for counts. Among them are Cameron and Trivedi (1998), Greene (1996a), Winkelmann (2000), and Wooldridge (1997).

[67]We have estimated a Poisson regression model using two-step nonlinear least squares in Example 17.9.

function is

$$\ln L = \sum_{i=1}^{n} [-\lambda_i + y_i \mathbf{x}_i'\boldsymbol{\beta} - \ln y_i!].$$

The likelihood equations are

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

The Hessian is

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

The Hessian is negative definite for all $\mathbf{x}$ and $\boldsymbol{\beta}$. Newton's method is a simple algorithm for this model and will usually converge rapidly. At convergence, $[\sum_{i=1}^{n} \hat{\lambda}_i \mathbf{x}_i \mathbf{x}_i']^{-1}$ provides an estimator of the asymptotic covariance matrix for the parameter estimates. Given the estimates, the prediction for observation i is $\hat{\lambda}_i = \exp(\mathbf{x}'\hat{\boldsymbol{\beta}})$. A standard error for the prediction interval can be formed by using a linear Taylor series approximation. The estimated variance of the prediction will be $\hat{\lambda}_i^2 \mathbf{x}_i' \mathbf{V} \mathbf{x}_i$, where $\mathbf{V}$ is the estimated asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$.

For testing hypotheses, the three standard tests are very convenient in this model. The Wald statistic is computed as usual. As in any discrete choice model, the likelihood ratio test has the intuitive form

$$\text{LR} = 2 \sum_{i=1}^{n} \ln \left(\frac{\hat{P}_i}{\hat{P}_{\text{restricted},i}} \right),$$

where the probabilities in the denominator are computed with using the restricted model. Using the BHHH estimator for the asymptotic covariance matrix, the LM statistic is simply

$$\text{LM} = \left[\sum_{i=1}^{n} \mathbf{x}_i'(y_i - \hat{\lambda}_i) \right]' \left[\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i'(y_i - \hat{\lambda}_i)^2 \right]^{-1} \left[\sum_{i=1}^{n} \mathbf{x}_i (y_i - \hat{\lambda}_i) \right] = \mathbf{i}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{i},$$

where each row of $\mathbf{G}$ is simply the corresponding row of $\mathbf{X}$ multiplied by $e_i = (y_i - \hat{\lambda}_i)$, $\hat{\lambda}_i$ is computed using the restricted coefficient vector, and $\mathbf{i}$ is a column of ones.

21.9.1 MEASURING GOODNESS OF FIT

The Poisson model produces no natural counterpart to the R^2 in a linear regression model, as usual, because the conditional mean function is nonlinear and, moreover, because the regression is heteroscedastic. But many alternatives have been suggested.[68]

[68] See the surveys by Cameron and Windmeijer (1993), Gurmu and Trivedi (1994), and Greene (1995b).

A measure based on the standardized residuals is

$$R_p^2 = 1 - \frac{\sum_{i=1}^{n}\left[\frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}\right]^2}{\sum_{i=1}^{n}\left[\frac{y_i - \bar{y}}{\sqrt{\bar{y}}}\right]^2}.$$

This measure has the virtue that it compares the fit of the model with that provided by a model with only a constant term. But it can be negative, and it can fall when a variable is dropped from the model. For an individual observation, the **deviance** is

$$d_i = 2[y_i \ln(y_i/\hat{\lambda}_i) - (y_i - \hat{\lambda}_i)] = 2[y_i \ln(y_i/\hat{\lambda}_i) - e_i],$$

where, by convention, $0 \ln(0) = 0$. If the model contains a constant term, then $\sum_{i=1}^{n} e_i = 0$. The sum of the deviances,

$$G^2 = \sum_{i=1}^{n} d_i = 2\sum_{i=1}^{n} y_i \ln(y_i/\hat{\lambda}_i),$$

is reported as an alternative fit measure by some computer programs. This statistic will equal 0.0 for a model that produces a perfect fit. (Note that since y_i is an integer while the prediction is continuous, it could not happen.) Cameron and Windmeijer (1993) suggest that the fit measure based on the deviances,

$$R_d^2 = 1 - \frac{\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right]}{\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\bar{y}}\right)\right]},$$

has a number of desirable properties. First, denote the log-likelihood function for the model in which ψ_i is used as the prediction (e.g., the mean) of y_i as $\ell(\psi_i, y_i)$. The Poisson model fit by MLE is, then, $\ell(\hat{\lambda}_i, y_i)$, the model with only a constant term is $\ell(\bar{y}, y_i)$, and a model that achieves a perfect fit (by predicting y_i with itself) is $l(y_i, y_i)$. Then

$$R_d^2 = \frac{\ell(\hat{\lambda}, y_i) - \ell(\bar{y}, y_i)}{\ell(y_i, y_i) - \ell(\bar{y}, y_i)}.$$

Both numerator and denominator measure the improvement of the model over one with only a constant term. The denominator measures the maximum improvement, since one cannot improve on a perfect fit. Hence, the measure is bounded by zero and one and increases as regressors are added to the model.[69] We note, finally, the passing resemblance of R_d^2 to the "pseudo-R^2," or "likelihood ratio index" reported by some statistical packages (e.g., Stata),

$$R_{\text{LRI}}^2 = 1 - \frac{\ell(\hat{\lambda}_i, y_i)}{\ell(\bar{y}, y_i)}.$$

[69]Note that multiplying both numerator and denominator by 2 produces the ratio of two likelihood ratio statistics, each of which is distributed as chi-squared.

Many modifications of the Poisson model have been analyzed by economists.[70] In this and the next few sections, we briefly examine a few of them.

21.9.2 TESTING FOR OVERDISPERSION

The Poisson model has been criticized because of its implicit assumption that the variance of y_i equals its mean. Many extensions of the Poisson model that relax this assumption have been proposed by Hausman, Hall, and Griliches (1984), McCullagh and Nelder (1983), and Cameron and Trivedi (1986), to name but a few.

The first step in this extended analysis is usually a test for overdispersion in the context of the simple model. A number of authors have devised tests for "overdispersion" within the context of the Poisson model. [See Cameron and Trivedi (1990), Gurmu (1991), and Lee (1986).] We will consider three of the common tests, one based on a regression approach, one a conditional moment test, and a third, a Lagrange multiplier test, based on an alternative model. Conditional moment tests are developed in Section 17.6.4.

Cameron and Trivedi (1990) offer several different tests for overdispersion. A simple regression based procedure used for testing the hypothesis

$$H_0: \text{Var}[y_i] = E[y_i],$$
$$H_1: \text{Var}[y_i] = E[y_i] + \alpha g(E[y_i])$$

is carried out by regressing

$$z_i = \frac{(y_i - \hat{\lambda}_i)^2 - y_i}{\hat{\lambda}_i \sqrt{2}},$$

where $\hat{\lambda}_i$ is the predicted value from the regression, on either a constant term or $\hat{\lambda}_i$ without a constant term. A simple t test of whether the coefficient is significantly different from zero tests H_0 versus H_1.

Cameron and Trivedi's regression based test for overdispersion is formulated around the alternative $\text{Var}[y_i] = E[y_i] + g(E[y_i])$. This is a very specific type of overdispersion. Consider the more general hypothesis that $\text{Var}[y_i]$ is completely given by $E[y_i]$. The alternative is that the variance is systematically related to the regressors in a way that is not completely accounted for by $E[y_i]$. Formally, we have $E[y_i] = \exp(\boldsymbol{\beta}'\mathbf{x}_i) = \lambda_i$. The null hypothesis is that $\text{Var}[y_i] = \lambda_i$ as well. We can test the hypothesis using the conditional moment test described in Section 17.6.4. The expected first derivatives and the moment restriction are

$$E[\mathbf{x}_i(y_i - \lambda_i)] = \mathbf{0} \quad \text{and} \quad E\{\mathbf{z}_i[(y_i - \lambda_i)^2 - \lambda_i]\} = \mathbf{0}.$$

To carry out the test, we do the following. Let $e_i = y_i - \hat{\lambda}_i$ and $\mathbf{z}_i = \mathbf{x}_i$ without the constant term.

1. Compute the Poisson regression by maximum likelihood.
2. Compute $\mathbf{r} = \sum_{i=1}^{n} \mathbf{z}_i[e_i^2 - \hat{\lambda}_i] = \sum_{i=1}^{n} \mathbf{z}v_i$ based on the maximum likelihood estimates.

[70] There have been numerous surveys of models for count data, including Cameron and Trivedi (1986) and Gurmu and Trivedi (1994).

3. Compute $\mathbf{M}'\mathbf{M} = \sum_{i=1}^{n} \mathbf{z}_i\mathbf{z}_i' v_i^2$, $\mathbf{D}'\mathbf{D} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i' e_i^2$, and $\mathbf{M}'\mathbf{D} = \sum_{i=1}^{n} \mathbf{z}_i\mathbf{x}_i' v_i e_i$.
4. Compute $\mathbf{S} = \mathbf{M}'\mathbf{M} - \mathbf{M}'\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{M}$.
5. $C = \mathbf{r}'\mathbf{S}^{-1}\mathbf{r}$ is the chi-squared statistic. It has K degrees of freedom.

The next section presents the **negative binomial model.** This model relaxes the Poisson assumption that the mean equals the variance. The Poisson model is obtained as a parametric restriction on the negative binomial model, so a Lagrange multiplier test can be computed. In general, if an alternative distribution for which the Poisson model is obtained as a parametric restriction, such as the negative binomial model, can be specified, then a Lagrange multiplier statistic can be computed. [See Cameron and Trivedi (1986, p. 41).] The LM statistic is

$$\text{LM} = \left[\frac{\sum_{i=1}^{n} \hat{w}_i[(y_i - \hat{\lambda}_i)^2 - y_i]}{\sqrt{2\sum_{i=1}^{n} \hat{w}_i \hat{\lambda}_i^2}}\right]^2.$$

The weight, $\hat{w}_i$, depends on the assumed alternative distribution. For the negative binomial model discussed later, $\hat{w}_i$ equals 1.0. Thus, under this alternative, the statistic is particularly simple to compute:

$$\text{LM} = \frac{(\mathbf{e}'\mathbf{e} - n\bar{y})^2}{2\hat{\boldsymbol{\lambda}}'\hat{\boldsymbol{\lambda}}}.$$

The main advantage of this test statistic is that one need only estimate the Poisson model to compute it. Under the hypothesis of the Poisson model, the limiting distribution of the LM statistic is chi-squared with one degree of freedom.

21.9.3 HETEROGENEITY AND THE NEGATIVE BINOMIAL REGRESSION MODEL

The assumed equality of the conditional mean and variance functions is typically taken to be the major shortcoming of the Poisson regression model. Many alternatives have been suggested [see Hausman, Hall, and Griliches (1984), Cameron and Trivedi (1986, 1998), Gurmu and Trivedi (1994), Johnson and Kotz (1993), and Winkelmann (1997) for discussion.] The most common is the negative binomial model, which arises from a natural formulation of cross-section heterogeneity. We generalize the Poisson model by introducing an individual, unobserved effect into the conditional mean,

$$\ln \mu_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i = \ln \lambda_i + \ln u_i,$$

where the disturbance ε_i reflects either specification error as in the classical regression model or the kind of cross-sectional heterogeneity that normally characterizes microeconomic data. Then, the distribution of y_i conditioned on $\mathbf{x}_i$ *and* u_i (i.e., ε_i) remains Poisson with conditional mean and variance μ_i:

$$f(y_i \mid \mathbf{x}_i, u_i) = \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!}.$$

The unconditional distribution $f(y_i \mid \mathbf{x}_i)$ is the expected value (over u_i) of $f(y_i \mid \mathbf{x}_i, u_i)$.

$$f(y_i \mid \mathbf{x}_i) = \int_0^{\infty} \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!} g(u_i)\, du_i.$$

The choice of a density for u_i defines the unconditional distribution. For mathematical convenience, a gamma distribution is usually assumed for $u_i = \exp(\varepsilon_i)$.[71] As in other models of heterogeneity, the mean of the distribution is unidentified if the model contains a constant term (because the disturbance enters multiplicatively) so $E[\exp(\varepsilon_i)]$ is assumed to be 1.0. With this normalization,

$$g(u_i) = \frac{\theta^\theta}{\Gamma(\theta)} e^{-\theta u_i} u_i^{\theta-1}.$$

The density for y_i is then

$$\begin{aligned} f(y_i \mid \mathbf{x}_i) &= \int_0^\infty \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!} \frac{\theta^\theta u_i^{\theta-1} e^{-\theta u_i}}{\Gamma(\theta)} du_i \\ &= \frac{\theta^\theta \lambda_i^{y_i}}{\Gamma(y_i+1)\Gamma(\theta)} \int_0^\infty e^{-(\lambda_i+\theta)u_i} u_i^{\theta+y_i-1} du_i \\ &= \frac{\theta^\theta \lambda_i^{y_i} \Gamma(\theta+y_i)}{\Gamma(y_i+1)\Gamma(\theta)(\lambda_i+\theta)^{\theta+y_i}} \\ &= \frac{\Gamma(\theta+y_i)}{\Gamma(y_i+1)\Gamma(\theta)} r_i^{y_i}(1-r_i)^\theta, \quad \text{where } r_i = \frac{\lambda_i}{\lambda_i+\theta}, \end{aligned}$$

which is one form of the negative binomial distribution. The distribution has conditional mean λ_i and conditional variance $\lambda_i(1 + (1/\theta)\lambda_i)$. [This model is Negbin II in Cameron and Trivedi's (1986) presentation.] The negative binomial model can be estimated by maximum likelihood without much difficulty. A test of the Poisson distribution is often carried out by testing the hypothesis $\theta = 0$ using the Wald or likelihood ratio test.

21.9.4 APPLICATION: THE POISSON REGRESSION MODEL

The number of accidents per service month for a sample of ship types is listed in Appendix Table F21.3. The ships are of five types constructed in one of four periods. The observation is over two periods. Since ships constructed from 1975 to 1979 could not have operated from 1960 to 1974, there is one missing observation in each group. The second observation for group E is also missing, for reasons unexplained by the authors.[72] The substantive variables in the model are number of accidents in the observation period and aggregate number of service months for the ship type by construction year for the period of operation.

Estimates of the parameters of a Poisson regression model are shown in Table 21.20. The model is

$$\ln E[\text{accident per month}] = \mathbf{x}'\boldsymbol{\beta}.$$

[71] An alternative approach based on the normal distribution is suggested in Terza (1998), Greene (1995a, 1997a), and Winkelmann (1997). The normal-Poisson mixture is also easily extended to the random effects model discussed in the next section. There is no closed form for the normal-Poisson mixture model, but it can be easily approximated by using Hermite quadrature.

[72] Data are from McCullagh and Nelder (1983). See Exercise 8 in Chapter 7 for details.

TABLE 21.20 Estimated Poisson Regressions (Standard Errors in Parentheses)

	Mean Dependent Variable	*10.47*				
Variable	*Full Model*		*No Ship Type Effect*		*No Period Effect*	
Constant	−6.4029	(0.2175)	−6.9470	(0.1269)	−5.7999	(0.1784)
Type = A						
Type = B	−0.5447	(0.1776)			−0.7437	(0.1692)
Type = C	−0.6888	(0.3290)			−0.7549	(0.3276)
Type = D	−0.0743	(0.2906)			−0.1843	(0.2876)
Type = E	0.3205	(0.2358)			0.3842	(0.2348)
60–64						
65–69	0.6959	(0.1497)	0.7536	(0.1488)		
70–74	0.8175	(0.1698)	1.0503	(0.1576)		
75–79	0.4450	(0.2332)	0.6999	(0.2203)		
Period = 60–74						
Period = 75–79	0.3839	(0.1183)	0.3875	(0.1181)	0.5001	(0.1116)
Log service	1.0000		1.0000		1.0000	
Log L	−68.41455		−80.20123		−84.11514	
G^2	38.96262		62.53596		70.34967	
R_p^2	0.94560		0.89384		0.90001	
R_d^2	0.93661		0.89822		0.88556	

The model therefore contains the ship type, construction period, and operation period effects, and the aggregate number of months with a coefficient of 1.0.[73] The model is shown in Table 21.20, with sets of estimates for the full model and with the model omitting the type and construction period effects. Predictions from the estimated full model are shown in the last column of Appendix Table F21.3.

The hypothesis that the year of construction is not a significant factor in explaining the number of accidents is strongly rejected by the likelihood ratio test:

$$\chi^2 = 2[84.11514 - 68.41455] = 31.40118.$$

The critical chi-squared value for three degrees of freedom is 7.82. The ship type effect is likewise significant,

$$\chi^2 = 2[80.20123 - 68.41455] = 23.57336,$$

against a critical value for four degrees of freedom of 9.49. The LM tests for the two restrictions give the same conclusions, but much less strongly. The value is 28.526 for the ship type effect and 31.418 for the period effects.

In their analysis of these data, McCullagh and Nelder assert, without evidence, that there is overdispersion in the data. Some of their analysis follows on an assumption that the standard deviation of y_i is 1.3 times the mean. The t statistics for the two regressions in Cameron and Trivedi's regression based tests are 0.934 and −0.613, respectively, so based on these tests, we do not reject H_0: no overdispersion. The LM statistic for the same

[73] When the length of the period of observation varies by observation by T_i and the model is of the rate of occurrence of events *per unit of time*, then the mean of the observed distribution is $T_i\lambda_i$. This assumption produces the coefficient of 1.0 on the number of periods of service in the model.

hypothesis is 0.75044 with one degree of freedom. The critical value from the table is 3.84, so again, the hypothesis of the Poisson model is not rejected. However, the conditional moment test is contradictory; $C = \mathbf{r}'\mathbf{S}^{-1}\mathbf{r} = 26.555$. There are eight degrees of freedom. The 5 percent critical value from the chi-squared table is 15.507, so the hypothesis is now rejected. This test is much more general, since the form of overdispersion is not specified, which may explain the difference. Note that this result affirms McCullagh and Nelder's conjecture.

21.9.5 POISSON MODELS FOR PANEL DATA

The familiar approaches to accommodating heterogeneity in panel data have fairly straightforward extensions in the count data setting. [Hausman, Hall, and Griliches (1984) give full details for these models.] We will examine them for the Poisson model. The authors [and Allison (2000)] also give results for the negative binomial model.

Consider first a fixed effects approach. The Poisson distribution is assumed to have conditional mean

$$\log \lambda_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i.$$

where now, $\mathbf{x}_{it}$ has been redefined to exclude the constant term. The approach used in the linear model of transforming y_{it} to group mean deviations does not remove the heterogeneity, nor does it leave a Poisson distribution for the transformed variable. However, the Poisson model with fixed effects can be fit using the methods described for the probit model in Section 21.5.1b. The extension to the Poisson model requires only the minor modifications, $g_{it} = (y_{it} - \lambda_{it})$ and $h_{it} = -\lambda_{it}$. Everything else in that derivation applies with only a simple change in the notation. The first order conditions for maximizing the log-likelihood function for the Poisson model will include

$$\frac{\partial \ln L}{\partial \alpha_i} = \sum_{t=1}^{T}(y_{it} - e^{\alpha_i}\mu_{it}) = 0 \quad \text{where } \mu_{it} = e^{\mathbf{x}_{it}'\boldsymbol{\beta}}.$$

This implies an explicit solution for α_i in terms of $\boldsymbol{\beta}$ in this model,

$$\hat{\alpha}_i = \ln\left(\frac{(1/n)\sum_{t=1}^{T} y_{it}}{(1/n)\sum_{t=1}^{T} \hat{\mu}_{it}}\right) = \ln\left(\frac{\bar{y}_i}{\bar{\hat{\mu}}_i}\right)$$

Unlike the regression or the probit model, this does not require that there be within group variation in y_{it}—all the values can be the same. It does require that at least one observation for individual i be nonzero, however. The rest of the solution for the fixed effects estimator follows the same lines as that for the probit model. An alternative approach, albeit with little practical gain, would be to concentrate the log likelihood function by inserting this solution for α_i back into the original log likelihood, then maximizing the resulting function of $\boldsymbol{\beta}$. While logically this makes sense, the approach suggested earlier for the probit model is simpler to implement.

An estimator that is not a function of the fixed effects is found by obtaining the joint distribution of $(y_{i1}, \ldots, y_{iT_i})$ conditional on their sum. For the Poisson model, a

close cousin to the logit model discussed earlier is produced:

$$p\left(y_{i1}, y_{i2}, \ldots, y_{iT_i} \,\middle|\, \sum_{i=1}^{T_i} y_{it}\right) = \frac{\left(\sum_{t=1}^{T_i} y_{it}\right)!}{\left(\prod_{t=1}^{T_i} y_{it}!\right)} \prod_{t=1}^{T_i} p_{it}^{y_{it}},$$

where

$$p_{it} = \frac{e^{\mathbf{x}_{it}'\boldsymbol{\beta}+\alpha_i}}{\sum_{t=1}^{T_i} e^{\mathbf{x}_{it}'\boldsymbol{\beta}+\alpha_i}} = \frac{e^{\mathbf{x}_{it}'\boldsymbol{\beta}}}{\sum_{t=1}^{T_i} e^{\mathbf{x}_{it}'\boldsymbol{\beta}}}.$$

The contribution of group i to the conditional log-likelihood is

$$\ln L_i = \sum_{t=1}^{T_i} y_{it} \ln p_{it}.$$

Note, once again, that the contribution to $\ln L$ of a group in which $y_{it} = 0$ in every period is zero. Cameron and Trivedi (1998) have shown that these two approaches give identical results.

The fixed effects approach has the same flaws and virtues in this setting as in the probit case. It is not necessary to assume that the heterogeneity is uncorrelated with the included, exogenous variables. If the uncorrelatedness of the regressors and the heterogeneity can be maintained, then the random effects model is an attractive alternative model. Once again, the approach used in the linear regression model, partial deviations from the group means followed by generalized least squares (see Chapter 13), is not usable here. The approach used is to formulate the joint probability conditioned upon the heterogeneity, then integrate it out of the joint distribution. Thus, we form

$$p(y_{i1}, \ldots, y_{iT_i} \mid u_i) = \prod_{t=1}^{T_i} p(y_{it} \mid u_i).$$

Then the random effect is swept out by obtaining

$$\begin{aligned} p(y_{i1}, \ldots, y_{iT_i}) &= \int_{u_i} p(y_{i1}, \ldots, y_{iT_i}, u_i)\, du_i \\ &= \int_{u_i} p(y_{i1}, \ldots, y_{iT_i} \mid u_i) g(u_i)\, du_i \\ &= E_{u_i}[p(y_{i1}, \ldots, y_{iT_i} \mid u_i)]. \end{aligned}$$

This is exactly the approach used earlier to condition the heterogeneity out of the Poisson model to produce the negative binomial model. If, as before, we take $p(y_{it} \mid u_i)$ to be Poisson with mean $\lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)$ in which $\exp(u_i)$ is distributed as gamma with mean 1.0 and variance $1/\alpha$, then the preceding steps produce the negative binomial distribution,

$$p(y_{i1}, \ldots, y_{iT_i}) = \frac{\left[\prod_{t=1}^{T_i} \lambda_{it}^{y_{it}}\right] \Gamma\left(\theta + \sum_{t=1}^{T_i} y_{it}\right)}{\left[\Gamma(\theta) \prod_{t=1}^{T_i} y_{it}!\right] \left[\left(\sum_{t=1}^{T_i} \lambda_{it}\right)^{\sum_{t=1}^{T_i} y_{it}}\right]} Q_i^{\theta} (1 - Q_i)^{\sum_{t=1}^{T_i} y_{it}},$$

where

$$Q_i = \frac{\theta}{\theta + \sum_{t=1}^{T_i} \lambda_{it}}.$$

For estimation purposes, we have a negative binomial distribution for $Y_i = \sum_t y_{it}$ with mean $\Lambda_i = \sum_t \lambda_{it}$.

There is a mild preference in the received literature for the fixed effects estimators over the random effects estimators. The virtue of dispensing with the assumption of uncorrelatedness of the regressors and the group specific effects is substantial. On the other hand, the assumption does come at a cost. In order to compute the probabilities or the marginal effects it is necessarily to estimate the constants, α_i. The unscaled coefficients in these models are of limited usefulness because of the nonlinearity of the conditional mean functions.

Other approaches to the random effects model have been proposed. Greene (1994, 1995a) and Terza (1995) specify a normally distributed heterogeneity, on the assumption that this is a more natural distribution for the aggregate of small independent effects. Brannas and Johanssen (1994) have suggested a semiparametric approach based on the GMM estimator by superimposing a very general form of heterogeneity on the Poisson model. They assume that conditioned on a random effect ε_{it}, y_{it} is distributed as Poisson with mean $\varepsilon_{it}\lambda_{it}$. The covariance structure of ε_{it} is allowed to be fully general. For $t, s = 1, \ldots, T$, $\text{Var}[\varepsilon_{it}] = \sigma_i^2$, $\text{Cov}[\varepsilon_{it}, \varepsilon_{js}] = \gamma_{ij}(|t - s|)$. For long time series, this model is likely to have far too many parameters to be identified without some restrictions, such as first-order homogeneity ($\boldsymbol{\beta}_i = \boldsymbol{\beta}\,\forall i$), uncorrelatedness across groups, [$\gamma_{ij}(.) = 0$ for $i \neq j$], groupwise homoscedasticity ($\sigma_i^2 = \sigma^2\,\forall i$), and nonautocorrelatedness [$\gamma(r) = 0 \forall r \neq 0$]. With these assumptions, the estimation procedure they propose is similar to the procedures suggested earlier. If the model imposes enough restrictions, then the parameters can be estimated by the method of moments. The authors discuss estimation of the model in its full generality. Finally, the latent class model discussed in Section 16.2.3 and the random parameters model in Section 17.8 extend naturally to the Poisson model. Indeed, most of the received applications of the latent class structure have been in the Poisson regression framework. [See Greene (2001) for a survey.]

21.9.6 HURDLE AND ZERO-ALTERED POISSON MODELS

In some settings, the zero outcome of the data generating process is qualitatively different from the positive ones. Mullahy (1986) argues that this fact constitutes a shortcoming of the Poisson (or negative binomial) model and suggests a "hurdle" model as an alternative.[74] In his formulation, a binary probability model determines whether a zero or a nonzero outcome occurs, then, in the latter case, a (truncated) Poisson distribution describes the positive outcomes. The model is

$$\text{Prob}(y_i = 0 \mid \mathbf{x}_i) = e^{-\theta}$$

$$\text{Prob}(y_i = j \mid \mathbf{x}_i) = \frac{(1 - e^{-\theta})\, e^{-\lambda_i} \lambda_i^j}{j!(1 - e^{-\lambda_i})}, \quad j = 1, 2, \ldots.$$

[74] For a similar treatment in a continuous data application, see Cragg (1971).

This formulation changes the probability of the zero outcome and scales the remaining probabilities so that the sum to one. It adds a new restriction that $\text{Prob}(y_i = 0 \mid \mathbf{x}_i)$ no longer depends on the covariates, however. Therefore, a natural next step is to parameterize this probability. Mullahy suggests some formulations and applies the model to a sample of observations on daily beverage consumption.

Mullahy (1986), Heilbron (1989), Lambert (1992), Johnson and Kotz (1993), and Greene (1994) have analyzed an extension of the hurdle model in which the zero outcome can arise from one of two regimes.[75] In one regime, the outcome is always zero. In the other, the usual Poisson process is at work, which can produce the zero outcome or some other. In Lambert's application, she analyzes the number of defective items produced by a manufacturing process in a given time interval. If the process is under control, then the outcome is always zero (by definition). If it is not under control, then the number of defective items is distributed as Poisson and may be zero or positive in any period. The model at work is therefore

$$\text{Prob}(y_i = 0 \mid \mathbf{x}_i) = \text{Prob}(\text{regime } 1) + \text{Prob}(y_i = 0 \mid \mathbf{x}_i, \text{ regime } 2)\text{Prob}(\text{regime } 2),$$

$$\text{Prob}(y_i = j \mid \mathbf{x}_i) = \text{Prob}(y_i = j \mid \mathbf{x}_i, \text{ regime } 2)\text{Prob}(\text{regime } 2), \quad j = 1, 2, \ldots.$$

Let z denote a binary indicator of regime 1 ($z = 0$) or regime 2 ($z = 1$), and let y^* denote the outcome of the Poisson process in regime 2. Then the observed y is $z \times y^*$. A natural extension of the splitting model is to allow z to be determined by a set of covariates. These covariates need not be the same as those that determine the conditional probabilities in the Poisson process. Thus, the model is

$$\text{Prob}(z_i = 1 \mid \mathbf{w}_i) = F(\mathbf{w}_i, \gamma),$$

$$\text{Prob}(y_i = j \mid \mathbf{x}_i, z_i = 1) = \frac{e^{-\lambda_i}\lambda_i^j}{j!}.$$

The mean in this distribution is

$$E[y_i \mid \mathbf{x}_i] = F \times 0 + (1 - F) \times E[y_i^* \mid \mathbf{x}_i, y_i^* > 0] = (1 - F) \times \frac{\lambda_i}{1 - e^{-\lambda_i}}.$$

Lambert (1992) and Greene (1994) consider a number of alternative formulations. including logit and probit models discussed in Sections 21.3 and 21.4, for the probability of the two regimes.

Both of these modifications substantially alter the Poisson formulation. First, note that the equality of the mean and variance of the distribution no longer follows; both modifications induce overdispersion. On the other hand, the overdispersion does not arise from heterogeneity; it arises from the nature of the process generating the zeros. As such, an interesting identification problem arises in this model. If the data do appear to be characterized by overdispersion, then it seems less than obvious whether it should be attributed to heterogeneity or to the regime splitting mechanism. Mullahy (1986) argues the point more strongly. He demonstrates that overdispersion will always induce excess zeros. As such, in a splitting model, we are likely to misinterpret the excess zeros as due to the splitting process instead of the heterogeneity.

[75] The model is variously labeled the "With Zeros," or WZ, model [Mullahy (1986)], the "Zero Inflated Poisson," or ZIP, model [Lambert (1992)], and "Zero-Altered Poisson," or ZAP, model [Greene (1994)].

It might be of interest to test simply whether there is a regime splitting mechanism at work or not. Unfortunately, the basic model and the zero-inflated model are not nested. Setting the parameters of the splitting model to zero, for example, does not produce Prob[$z = 0$] = 0. In the probit case, this probability becomes 0.5, which maintains the regime split. The preceding tests for over- or underdispersion would be rather indirect. What is desired is a test of non-Poissonness. An alternative distribution may (but need not) produce a systematically different proportion of zeros than the Poisson. Testing for a different distribution, as opposed to a different set of parameters, is a difficult procedure. Since the hypotheses are necessarily nonnested, the power of any test is a function of the alternative hypothesis and may, under some, be small. Vuong (1989) has proposed a test statistic for **nonnested models** that is well suited for this setting when the alternative distribution can be specified. Let $f_j(y_i \mid \mathbf{x}_i)$ denote the predicted probability that the random variable Y equals y_i under the assumption that the distribution is $f_j(y_i \mid \mathbf{x}_i)$, for $j = 1, 2$, and let

$$m_i = \log\left(\frac{f_1(y_i \mid \mathbf{x}_i)}{f_2(y_i \mid \mathbf{x}_i)}\right).$$

Then Vuong's statistic for testing the nonnested hypothesis of Model 1 versus Model 2 is

$$v = \frac{\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} m_i\right]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m})^2}}.$$

This is the standard statistic for testing the hypothesis that $E[m_i]$ equals zero. Vuong shows that v has a limiting standard normal distribution. As he notes, the statistic is bidirectional. If $|v|$ is less than two, then the test does not favor one model or the other. Otherwise, large values favor Model 1 whereas small (negative) values favor Model 2. Carrying out the test requires estimation of both models and computation of both sets of predicted probabilities.

In Greene (1994), it is shown that the Vuong test has some power to discern this phenomenon. The logic of the testing procedure is to allow for overdispersion by specifying a negative binomial count data process, then examine whether, *even allowing for the overdispersion,* there still appear to be excess zeros. In his application, that appears to be the case.

Example 21.12 A Split Population Model for Major Derogatory Reports

Greene (1995c) estimated a model of consumer behavior in which the dependent variable of interest was the number of major derogatory reports recorded in the credit history for a sample of applicants for a type of credit card. The basic model predicts y_i, the number of major derogatory credit reports, as a function of $\mathbf{x}_i$ = [1, age, income, average expenditure]. The data for the model appear in Appendix Table F21.4. There are 1,319 observations in the sample (10% of the original data set.) Inspection of the data reveals a preponderance of zeros. Indeed, of 1,319 observations, 1060 have $y_i = 0$, whereas of the remaining 259, 137 have 1, 50 have 2, 24 have 3, 17 have 4, and 11 have 5—the remaining 20 range from 6 to 14. Thus, for a Poisson distribution, these data are actually a bit extreme. We propose to use Lambert's zero inflated Poisson model instead, with the Poisson distribution built around

$$\ln \lambda_i = \beta_1 + \beta_2 \text{ age} + \beta_3 \text{ income} + \beta_4 \text{ expenditure}.$$

For the splitting model, we use a logit model, with covariates $\mathbf{z}$ = [1, age, income, own/rent]. The estimates are shown in Table 21.21. Vuong's diagnostic statistic appears to confirm

TABLE 21.21 Estimates of a Split Population Model

	Poisson and Logit Models		*Split Population Model*	
Variable	*Poisson for y*	*Logit for y > 0*	*Poisson for y*	*Logit for y > 0*
Constant	−0.8196	−2.2442	1.0010	2.1540
	(0.1453)	(0.2515)	(0.1267)	(0.2900)
Age	0.007181	0.02245	−0.005073	−0.02469
	(0.003978)	(0.007313)	(0.003218)	(0.008451)
Income	0.07790	0.06931	0.01332	−0.1167
	(0.02394)	(0.04198)	(0.02249)	(0.04941)
Expend	−0.004102		−0.002359	
	(0.0003740)		(0.0001948)	
Own/Rent		−0.3766		0.3865
		(0.1578)		(0.1709)
Log L	−1396.719	−645.5649	−1093.0280	
$n\hat{P}(0 \mid \hat{\mathbf{x}})$	938.6		1061.5	

intuition that the Poisson model does not adequately describe the data; the value is 6.9788. Using the model parameters to compute a prediction of the number of zeros, it is clear that the splitting model does perform better than the basic Poisson regression.

21.10 SUMMARY AND CONCLUSIONS

This chapter has surveyed techniques for modeling discrete choice. We examined four classes of models: binary choice, ordered choice, multinomial choice, and models for counts. The first three of these are quite far removed from the regression models (linear and nonlinear) that have been the focus of the preceding 20 chapters. The most important difference concerns the modeling approach. Up to this point, we have been primarily interested in modeling the conditional mean function for outcomes that vary continuously. In this chapter, we have shifted our approach to one of modeling the conditional probabilities of events.

Modeling binary choice—the decision between two alternatives—is a growth area in the applied econometrics literature. Maximum likelihood estimation of fully parameterized models remains the mainstay of the literature. But, we also considered semiparametric and nonparametric forms of the model and examined models for time series and panel data. The ordered choice model is a natural extension of the binary choice setting and also a convenient bridge between models of choice between two alternatives and more complex models of choice among multiple alternatives. Multinomial choice modeling is likewise a large field, both within economics and, especially, in many other fields, such as marketing, transportation, political science, and so on. The multinomial logit model and many variations of it provide an especially rich framework within which modelers have carefully matched behavioral modeling to empirical specification and estimation. Finally, models of count data are closer to regression models than the other three fields. The Poisson regression model is essentially a nonlinear regression, but, as in the other cases, it is more fruitful to do the modeling in terms of the probabilities of discrete choice rather than as a form of regression analysis.

Key Terms and Concepts

- Attributes
- Binary choice model
- Bivariate probit
- Bootstrapping
- Butler and Moffitt method
- Choice based sampling
- Chow test
- Conditional likelihood function
- Conditional logit
- Count data
- Fixed effects model
- Full information ML
- Generalized residual
- Goodness of fit measure
- Grouped data
- Heterogeneity
- Heteroscedasticity
- Incidental parameters problem
- Inclusive value
- Independence from irrelevant alternatives
- Index function model
- Individual data
- Initial conditions
- Kernel density estimator
- Kernel function
- Lagrange multiplier test
- Latent regression
- Likelihood equations
- Likelihood ratio test
- Limited information ML
- Linear probability model
- Logit
- Marginal effects
- Maximum likelihood
- Maximum score estimator
- Maximum simulated likelihood
- Mean-squared deviation
- Minimal sufficient statistic
- Minimum chi-squared estimator
- Multinomial logit
- Multinomial probit
- Multivariate probit
- Negative binomial model
- Nested logit
- Nonnested models
- Normit
- Ordered choice model
- Overdispersion
- Persistence
- Poisson model
- Probit
- Proportions data
- Quadrature
- Qualitative choice
- Qualitative response
- Quasi-MLE
- Random coefficients
- Random effects model
- Random parameters model
- Random utility model
- Ranking
- Recursive model
- Robust covariance estimation
- Sample selection
- Scoring method
- Semiparametric estimation
- State dependence
- Unbalanced sample
- Unordered
- Weibull model

Exercises

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

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

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

		y	
		0	1
d	0	24	28
	1	32	16

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

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

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

estimate a probit model and test the hypothesis that x is not influential in determining the probability that y equals one.
4. Construct the Lagrange multiplier statistic for testing the hypothesis that all the slopes (but not the constant term) equal zero in the binomial logit model. Prove that the Lagrange multiplier statistic is nR^2 in the regression of $(y_i - p)$ on the xs, where P is the sample proportion of 1s.
5. We are interested in the ordered probit model. Our data consist of 250 observations, of which the response are

y	0	1	2	3	4
n	50	40	45	80	35

Using the preceding data, obtain maximum likelihood estimates of the unknown parameters of the model. [Hint: Consider the probabilities as the unknown parameters.]
6. The following hypothetical data give the participation rates in a particular type of recycling program and the number of trucks purchased for collection by 10 towns in a small mid-Atlantic state:

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

The town of Eleven is contemplating initiating a recycling program but wishes to achieve a 95 percent rate of participation. Using a probit model for your analysis,
   a. How many trucks would the town expect to have to purchase in order to achieve their goal? [Hint: See Section 21.4.6.] Note that you will use $n_i = 1$.
   b. If trucks cost \$20,000 each, then is a goal of 90 percent reachable within a budget of \$6.5 million? (That is, should they *expect* to reach the goal?)
   c. According to your model, what is the marginal value of the 301st truck in terms of the increase in the percentage participation?
7. A data set consists of $n = n_1 + n_2 + n_3$ observations on y and x. For the first n_1 observations, $y = 1$ and $x = 1$. For the next n_2 observations, $y = 0$ and $x = 1$. For the last n_3 observations, $y = 0$ and $x = 0$. Prove that neither (21-19) nor (21-21) has a solution.

8. Data on $t =$ strike duration and $x =$ unanticipated industrial production for a number of strikes in each of 9 years are given in Appendix Table F22.1. Use the Poisson regression model discussed in Section 21.9 to determine whether x is a significant determinant of the *number of strikes* in a given year.
9. Asymptotics. Explore whether averaging individual marginal effects gives the same answer as computing the marginal effect at the mean.
10. Prove (21-28).
11. In the panel data models estimated in Example 21.5.1, neither the logit nor the probit model provides a framework for applying a Hausman test to determine whether fixed or random effects is preferred. Explain. (Hint: Unlike our application in the linear model, the incidental parameters problem persists here.)