18
DISCRETE CHOICES AND EVENT COUNTS

18.1 INTRODUCTION

Chapter 17 presented most of the econometric issues that arise in analyzing discrete dependent variables, including specification, estimation, inference, and a variety of variations on the basic model. All of these were developed in the context of a model of binary choice, the choice between two alternatives. This chapter will use those results in extending the choice model to three specific settings:

Multinomial Choice: The individual chooses among more than two choices, once again, making the choice that provides the greatest utility. Applications include the choice among political candidates, how to commute to work, where to live, or what brand of car, appliance, or food product to buy.

Ordered Choice: The individual reveals the strength of their preferences with respect to a single outcome. Familiar cases involve survey questions about strength of feelings about a particular commodity such as a movie, a book, or a consumer product, or self-assessments of social outcomes such as health in general or self-assessed well-being. Although preferences will probably vary continuously in the space of individual utility, the expression of those preferences for purposes of analyses is given in a discrete outcome on a scale with a limited number of choices, such as the typical five-point scale used in marketing surveys.

Event Counts: The observed outcome is a count of the number of occurrences. In many cases, this is similar to the preceding settings in that the "dependent variable" measures an individual choice, such as the number of visits to the physician or the hospital, the number of derogatory reports in one's credit history, or the number of visits to a particular recreation site. In other cases, the event count might be the outcome of some less focused natural process, such as incidence of a disease in a population or the number of defects per unit of time in a production process, the number of traffic accidents that occur at a particular location per month, or the number of messages that arrive at a switchboard per unit of time over the course of a day. In this setting, we will be doing a more familiar sort of regression modeling.

Most of the methodological underpinnings needed to analyze these cases were presented in Chapter 17. In this chapter, we will be able to develop variations on these basic model types that accommodate different choice situations. As in Chapter 17, we are focused on ~~models with~~ discrete outcomes, so the analysis is framed in terms of models of the probabilities attached to those outcomes.

18.2 MODELS FOR UNORDERED MULTIPLE CHOICES

Some studies of multiple-choice settings include the following:

1. Hensher (1986, 1991), McFadden (1974), and many others have analyzed the travel mode of urban commuters. In Greene (2007b), Hensher and Greene analyze commuting between Sydney and Melbourne by a sample of individuals who choose among air, train, bus, and car as the mode of travel.
2. Schmidt and Strauss (1975a, b) and Boskin (1974) have analyzed occupational choice among multiple alternatives.
3. Rossi and Allenby (1999, 2003) studied consumer brand choices in a repeated choice (panel data) model.
4. Train (2003) studied the choice of electricity supplier by a sample of California electricity customers.
5. Hensher, Rose, and Greene (2006) analyzed choices of automobile models by a sample of consumers offered a hypothetical menu of features.

In each of these cases, there is a single decision among two or more alternatives. In this and the next section, we will encounter two broad types of multinomial choice sets, **unordered choice models** and **ordered choices.** All of the choice sets listed are unordered. In contrast, a bond rating or a preference scale is, by design, a ranking; that is, its purpose. Quite different techniques are used for the two types of models. We will examined models for ordered choices in Section 18.3. This section will examine models for unordered choice sets. General references on the topics discussed here include Hensher, Louviere, and Swait (2000), Train (2009), and Hensher, Rose, and Greene (2006).

18.2.1 RANDOM UTILITY BASIS OF THE MULTINOMIAL LOGIT MODEL

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

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

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 in the binary choice case, two models are usually 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, politics, finance, and transportation engineering. Let Y_i be a random variable that indicates the choice made. McFadden (1974a) has shown that if (and only if) the J disturbances are independent

and identically distributed with Gumbel (type 1 extreme value) distribution,

$$F(\varepsilon_{ij}) = \exp(-\exp(-\varepsilon_{ij})), \tag{18-1}$$

then

$$\text{Prob}(Y_i = j) = \frac{\exp(\mathbf{z}_{ij}'\boldsymbol{\theta})}{\sum_{j=1}^{J}\exp(\mathbf{z}_{ij}'\boldsymbol{\theta})}, \tag{18-2}$$

which leads to what is called the **conditional logit model.** (It is often 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.)

Utility depends on $\mathbf{z}_{ij}$, which includes aspects specific to the individual as well as to the choices. It is useful to distinguish them. Let $\mathbf{z}_{ij} = [\mathbf{x}_{ij}, \mathbf{w}_i]$ and partition $\boldsymbol{\theta}$ conformably into $[\boldsymbol{\beta}', \boldsymbol{\alpha}']'$. 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 (18-2) becomes

$$\text{Prob}(Y_i = j) = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\alpha})}{\sum_{j=1}^{J}\exp(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\alpha})} = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})\exp(\mathbf{w}_i'\boldsymbol{\alpha})}{\left[\sum_{j=1}^{J}\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})\right]\exp(\mathbf{w}_i'\boldsymbol{\alpha})}. \tag{18-3}$$

Terms that do not vary across alternatives—that is, those specific to the individual—fall out of the probability. This is as expected in a model that compares the utilities of the alternatives.

For example, in a model of a shopping center choice by individuals in various cities that depends on the number of stores at the mall, S_{ij}, the distance from the central business district, D_{ij} and the shoppers' incomes, I_i, the utilities for three choices would be

$$\begin{aligned} U_{i1} &= D_{i1}\beta_1 + S_{i1}\beta_2 + \alpha + \gamma I_i + \varepsilon_{i1};\\ U_{i2} &= D_{i2}\beta_1 + S_{i2}\beta_2 + \alpha + \gamma I_i + \varepsilon_{i2};\\ U_{i3} &= D_{i3}\beta_1 + S_{i3}\beta_2 + \alpha + \gamma I_i + \varepsilon_{i3}. \end{aligned}$$

The choice of alternative 1, for example, reveals that

$$\begin{aligned} U_{i1} - U_{i2} &= (D_{i1} - D_{i2})\beta_1 + (S_{i1} - S_{i2})\beta_2 + (\varepsilon_{i1} - \varepsilon_{i2}) > 0 \text{ and}\\ U_{i1} - U_{i3} &= (D_{i1} - D_{i3})\beta_1 + (S_{i1} - S_{i3})\beta_2 + (\varepsilon_{i1} - \varepsilon_{i3}) > 0. \end{aligned}$$

The constant term and *Income* have fallen out of the comparison. The result follows from the fact that random utility model is ultimately based on comparisons of pairs of alternatives, not the alternatives themselves. 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 (alternative specific constants), A_j, for the choices and multiply each of them by the common **w**. We then allow the coefficients on these choice invariant characteristics 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 this example, the matrix of attributes and characteristics would be

$$Z_i = \begin{bmatrix} S_{i1} & D_{i1} & 1 & 0 & I_i & 0 \\ S_{i2} & D_{i2} & 0 & 1 & 0 & I_i \\ S_{i3} & D_{i3} & 0 & 0 & 0 & 0 \end{bmatrix}$$

The probabilities for this model would be

$$\text{Prob}(Y_i = j|\mathbf{Z}_i) = \frac{\exp\begin{pmatrix} Stores_{ij}\,\beta_1 + Distance_{ij}\,\beta_2 \\ A_1\alpha_1 + A_2\alpha_2 + A_3\alpha_3 \\ A_1\,Income_i\gamma_1 + A_2\,Income_i\gamma_2 + A_3\,Income_i\gamma_3 \end{pmatrix}}{\sum_{j=1}^{3}\exp\begin{pmatrix} Stores_{ij}\,\beta_1 + Distance_{ij}\,\beta_2 \\ A_1\alpha_1 + A_2\alpha_2 + A_3\alpha_3 \\ A_1\,Income_i\gamma_1 + A_2\,Income_i\gamma_2 + A_3\,Income_i\gamma_3 \end{pmatrix}}, \boldsymbol{\alpha}_3 = \boldsymbol{\gamma}_3 = 0.$$

18.2.2 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 choice based on a sample of 1,000 observations drawn from the Public Use Sample for three years: l960, 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. (Note the slightly different numbering convention, starting at zero, which is standard.)
2. *Characteristics:* constant, education, experience, race, sex.

The model for occupational choice is

$$\text{Prob}(Y_i = j \mid \mathbf{w}_i) = \frac{\exp(\mathbf{w}_i'\boldsymbol{\alpha}_j)}{\sum_{j=0}^{4}\exp(\mathbf{w}_i'\boldsymbol{\alpha}_j)}, \quad j = 0, 1, \ldots, 4. \tag{18-4}$$

(The binomial logit model in Section 17.3 is conveniently produced as the special case of $J = 1$.)

The model in (18-4) is a multinomial logit model.[1] The estimated equations provide a set of probabilities for the $J + 1$ choices for a decision maker with characteristics $\mathbf{w}_i$. Before proceeding, we must remove an indeterminacy in the model. If we define $\boldsymbol{\alpha}_j^* = \boldsymbol{\alpha}_j + \mathbf{q}$ for any vector $\mathbf{q}$, then recomputing the probabilities defined later using $\boldsymbol{\alpha}_j^*$ instead of $\boldsymbol{\alpha}_j$ produces the identical set of probabilities because all the terms involving $\mathbf{q}$ drop out. A convenient normalization that solves the problem is $\boldsymbol{\alpha}_0 = \mathbf{0}$. (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

$$\text{Prob}(Y_i = j \mid \mathbf{w}_i) = P_{ij} = \frac{\exp(\mathbf{w}_i'\boldsymbol{\alpha}_j)}{1 + \sum_{k=1}^{J}\exp(\mathbf{w}_i'\boldsymbol{\alpha}_k)}, \quad j = 0, 1, \ldots, J, \quad \boldsymbol{\alpha}_0 = \mathbf{0}. \tag{18-5}$$

[1] Nerlove and Press (1973).

The form of the binomial model examined in Section 17.3 results if $J = 1$. The model implies that we can compute J **log-odds**

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

From the point of view of estimation, it is useful that the odds ratio, P_{ij}/P_{ik}, does not depend on the other choices, which follows from the independence of the disturbances in the original model. From a behavioral viewpoint, this fact is not very attractive. We shall return to this problem in Section 18.2.4.

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:

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

The derivatives have the characteristically simple form

$$\frac{\partial \ln L}{\partial \boldsymbol{\alpha}_j} = \sum_{i=1}^{n} (d_{ij} - P_{ij})\mathbf{w}_i \quad \text{for } j = 1, \ldots, J.$$

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

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

where $\mathbf{1}(j = l)$ equals 1 if j equals l and 0 if not. Because 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 inconvenient 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 $\boldsymbol{\alpha}_j$ with the jth outcome, but that would be misleading. By differentiating (18-5), we find that the partial effects of the characteristics on the probabilities are

$$\boldsymbol{\delta}_{ij} = \frac{\partial P_{ij}}{\partial \mathbf{w}_i} = P_{ij}\left[\boldsymbol{\alpha}_j - \sum_{k=0}^{J} P_{ik}\boldsymbol{\alpha}_k\right] = P_{ij}[\boldsymbol{\alpha}_j - \bar{\boldsymbol{\alpha}}]. \tag{18-6}$$

Therefore, every subvector of $\boldsymbol{\alpha}$ enters every partial effect, both through the probabilities and through the weighted average that appears in $\boldsymbol{\delta}_{ij}$. These values can be computed from the parameter estimates. Although the usual focus is on the coefficient estimates, equation (18-6) suggests that there is at least some potential for confusion. Note, for example, that for any particular w_{ik}, $\partial P_{ij}/\partial w_{ik}$ need not have the same sign as α_{jk}. Standard errors can be estimated using the delta method. (See Section 4.4.4.) For purposes of the computation, let $\boldsymbol{\alpha} = [\mathbf{0}, \boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2', \ldots, \boldsymbol{\alpha}_J']'$. We include the fixed $\mathbf{0}$ vector for outcome 0 because although $\boldsymbol{\alpha}_0 = \mathbf{0}$, $\boldsymbol{\delta}_{i0} = -P_{i0}\bar{\boldsymbol{\alpha}}$, which is not $\mathbf{0}$. Note as well that

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

Asy. $\text{Cov}[\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}}_j] = \mathbf{0}$ for $j = 0, \ldots, J$. Then

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

$$\frac{\partial \boldsymbol{\delta}_{ij}}{\partial \boldsymbol{\alpha}_l'} = [\mathbf{1}(j = l) - P_{il}][P_{ij}\mathbf{I} + \boldsymbol{\delta}_{ij}\mathbf{w}_i'] - P_{ij}[\boldsymbol{\delta}_{il}\mathbf{w}_i'].$$

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 characteristic vector includes only a constant term, then the restricted log-likelihood is

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

where p_j is the sample proportion of observations that make choice j. A useful table will give a listing of hits and misses of the prediction rule "predict $Y_i = j$ if $\hat{P}_{ij}$ is the maximum of the predicted probabilities."[3]

Example 18.1 Hollingshead Scale of Occupations

Fair's (1977) study of extramarital affairs is based on a cross section of 601 responses to a survey by *Psychology Today*. One of the covariates is a category of occupations on a seven-point scale, the Hollingshead (1975) scale. [See, also, Bornstein and Bradley (2003).] The Hollingshead scale is intended to be a measure on a prestige scale, a fact which we'll ignore (or disagree with) for the present. The seven levels on the scale are, broadly,

1. Higher executives
2. Managers and proprietors of medium-sized businesses
3. Administrative personnel and owners of small businesses
4. Clerical and sales workers and technicians
5. Skilled manual employees
6. Machine operators and semiskilled employees
7. Unskilled employees

Among the other variables in the data set are *Age*, *Sex*, and *Education*. The data are given in Appendix Table F18.1. Table 18.1 lists estimates of a multinomial logit model. (We emphasize that the data are a self-selected sample of *Psychology Today* readers in 1976, so it is unclear what contemporary population would be represented. The following serves as an uncluttered numerical example that readers could reproduce. Note, as well, that at least by some viewpoint, the outcome for this experiment is ordered.) The log-likelihood for the model is −770.28141 while that for the model with only the constant terms is −982.20533. The likelihood ratio statistic for the hypothesis that all 18 coefficients of the model are zero is 423.85, which is far larger than the critical value of 28.87. In the estimated parameters, it appears that only gender is consistently statistically significant. However, it is unclear how

[3]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. This is not a contradiction of an estimated model with many "significant" coefficients, because the coefficients are not estimated so as to maximize the number of correct predictions.

TABLE 18.1 Estimated Multinomial Logit Model for Occupation (t ratios in parentheses)

	α_0	α_1	α_2	α_3	α_4	α_5	α_6
				Parameters			
Constant	0.0	3.1506	2.0156	−1.9849	−6.6539	−15.0779	−12.8919
	(0.0)	(1.14)	(1.28)	(−1.38)	(−5.49)	(−9.18)	(−4.61)
Age	0.0	−0.0244	−0.0361	−0.0123	0.0038	0.0225	0.0588
	(0.0)	(−0.73)	(−1.64)	(−0.63)	(0.25)	(1.22)	(1.92)
Sex	0.0	6.2361	4.6294	4.9976	4.0586	5.2086	5.8457
	(0.0)	(5.08)	(4.39)	(4.82)	(3.98)	(5.02)	(4.57)
Education	0.0	−0.4391	−0.1661	0.0684	0.4288	0.8149	0.4506
	(0.0)	(−2.62)	(−1.75)	(0.79)	(5.92)	(8.56)	(2.92)
				Partial Effects			
Age	−0.0001	−0.0002	−0.0028	−0.0022	0.0006	0.0036	0.0011
	(−.19)	(−0.92)	(−2.23)	(−1.15)	(0.23)	(1.89)	(1.90)
Sex	−0.2149	0.0164	0.0233	0.1041	−0.1264	0.1667	0.0308
	(−4.24)	(1.98)	(1.00)	(2.87)	(−2.15)	(4.20)	(2.35)
Education	−0.0187	−0.0069	−0.0387	−0.0460	0.0278	0.0810	0.0015
	(−2.22)	(−2.31)	(−6.29)	(−5.1)	(2.12)	(8.61)	(0.56)

to interpret the fact that *Education* is significant in some of the parameter vectors and not others. The partial effects give a similarly unclear picture, though in this case, the effect can be associated with a particular outcome. However, we note that the implication of a test of significance of a partial effect in this model is itself ambiguous. For example, *Education* is not "significant" in the partial effect for outcome 6, though the coefficient on *Education* in $\boldsymbol{\alpha}_6$ is. This is an aspect of modeling with multinomial choice models that calls for careful interpretation by the model builder.

18.2.3 THE CONDITIONAL LOGIT MODEL

When the data consist of choice-specific attributes instead of individual-specific characteristics, the natural model formulation would be

$$\text{Prob}(Y_i = j \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iJ}) = \text{Prob}(Y_i = j \mid \mathbf{X}_i) = P_{ij} = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}{\sum_{j=1}^{J} \exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}. \quad \textbf{(18-7)}$$

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 (18-7) with respect to a particular $\mathbf{x}_m$ to obtain

$$\frac{\partial P_{ij}}{\partial \mathbf{x}_{im}} = [P_{ij}(\mathbf{1}(j = m) - P_{im})]\boldsymbol{\beta}, \quad m = 1, \ldots, J.$$

It is clear that through its presence in P_{ij} and P_{im}, every attribute set $\mathbf{x}_m$ affects all the probabilities. Hensher (1991) suggests that one might prefer to report elasticities of the probabilities. The effect of attribute k of choice m on P_{ij} would be

$$\frac{\partial \ln P_j}{\partial \ln x_{mk}} = x_{mk}[\mathbf{1}(j = m) - P_{im}]\beta_k.$$

Because 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.

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

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

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.

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,

$$\begin{aligned}\frac{\partial \log L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i),\\ \frac{\partial^2 \log L}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} &= -\sum_{i=1}^{n}\sum_{j=1}^{J} P_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)',\end{aligned} \tag{18-8}$$

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. The model cannot contain a constant term, so 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.

18.2.4 THE INDEPENDENCE FROM IRRELEVANT ALTERNATIVES ASSUMPTION

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_{ij}/P_{im} 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 excluded 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.[4]

18.2.5 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:

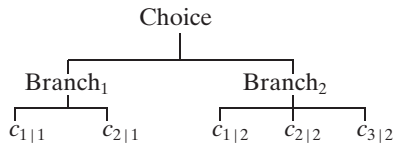
$$U_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij}, \quad j = 1, \ldots, J, [\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iJ}] \sim N[\mathbf{0}, \boldsymbol{\Sigma}]. \tag{18-9}$$

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 $U_{ij} > U_{im}$ for $m \neq j$ involve the $J - 1$ differences, $\varepsilon_j - \varepsilon_m$. 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 necessarily as a model of behavior). Suppose, then, that the J alternatives can be divided into B subgroups (branches) such that the choice set can be written

$$[c_1, \ldots, c_J] = [(c_{1|1}, \ldots, c_{J_1|1}), (c_{1|2}, \ldots, c_{J_2|2}) \ldots, (c_{1|B}, \ldots, c_{J_B|B})].$$

Logically, we may think of the choice process as that of choosing among the B 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 (twigs) might look as follows:



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

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

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

$$\text{Prob}[twig_j, branch_b] = P_{ijb} = \frac{\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta} + \mathbf{z}'_{ib}\boldsymbol{\gamma})}{\sum_{b=1}^{B}\sum_{j=1}^{J_b}\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta} + \mathbf{z}'_{ib}\boldsymbol{\gamma})}.$$

Now write this probability as

$$\begin{aligned} P_{ijb} &= P_{ij|b}P_b \\ &= \left(\frac{\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})}{\sum_{j=1}^{J_b}\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})}\right)\left(\frac{\exp(\mathbf{z}'_{ib}\boldsymbol{\gamma})}{\sum_{l=1}^{L}\exp(\mathbf{z}'_{ib}\boldsymbol{\gamma})}\right)\frac{\left(\sum_{j=1}^{J_b}\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})\right)\left(\sum_{l=1}^{L}\exp(\mathbf{z}'_{ib}\boldsymbol{\gamma})\right)}{\left(\sum_{l=1}^{L}\sum_{j=1}^{J_l}\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta} + \mathbf{z}'_{ib}\boldsymbol{\gamma})\right)}. \end{aligned}$$

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

$$IV_{ib} = \ln\left(\sum_{j=1}^{J_b}\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})\right).$$

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

$$P_{ij|b} = \frac{\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})}{\sum_{j=1}^{J_b}\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})} \quad \text{and} \quad P_b = \frac{\exp[\tau_b(\mathbf{z}'_{ib}\boldsymbol{\gamma} + IV_{ib})]}{\sum_{b=1}^{B}\exp[\tau_b(\mathbf{z}'_{ib}\boldsymbol{\gamma} + IV_{ib})]},$$

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[5]

$$\sigma_b^2 = \frac{\pi^2}{6\tau_b}. \tag{18-10}$$

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

$$\begin{aligned} &\frac{\partial \ln \text{Prob}[choice = m, branch = b]}{\partial x(k) \text{ in } choice\ M \text{ and } branch\ B} \\ &\qquad = \{\mathbf{1}(b = B)[\mathbf{1}(m = M) - P_{M|B}] + \tau_B[\mathbf{1}(b = B) - P_B]P_M \mid B\}\boldsymbol{\beta}_k. \end{aligned}$$

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

[5]See Hensher, Louviere, and Swaite (2000). See Greene and Hensher (2002) for alternative formulations of the nested logit model.

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 $\boldsymbol{\tau}$ parameters by treating the choice among branches as a conditional logit model with attributes $\mathbf{z}_{ib}$ and I_{ib}.

Because 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 14.7 and McFadden (1984).] For **full information maximum likelihood** (FIML) estimation of the model, the log-likelihood is

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

[See Hensher (1986, 1991) and Greene (2007a).] 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. The FIML estimator is now available in several commercial computer packages. The two-step estimator is rarely used in current research.

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.

18.2.6 THE MULTINOMIAL PROBIT MODEL

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

$$U_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij}, \;\; j = 1, \ldots, J, [\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iJ}] \sim N[\mathbf{0}, \boldsymbol{\Sigma}].$$

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

$$\text{Prob}[\text{choice}_{iq}] = \text{Prob}[U_{iq} > U_{ij}, \;\; j = 1, \ldots, J, \; j \neq q].$$

The probability for this occurrence is

$$\text{Prob}[\text{choice}_i q] = \text{Prob}[\varepsilon_{i1} - \varepsilon_{iq} < (\mathbf{x}_{iq} - \mathbf{x}_{i1})'\boldsymbol{\beta}, \ldots, \varepsilon_{iJ} - \varepsilon_{iq} < (\mathbf{x}_{iq} - \mathbf{x}_{iJ})'\boldsymbol{\beta}]$$

for the $J - 1$ other choices, which is a cumulative probability from a $(J - 1)$-variate normal distribution. Because 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. Because 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 17.2.) For more than two choices, this specification is far more general than the MNL model, which assumes that $\mathbf{\Sigma}=\mathbf{I}$. (The scaling is absorbed in the coefficient vector in the MNL model.) It adds the unrestricted correlations to the heteroscedastic model of the previous section.

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 15.6.2.b 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 $\mathbf{\Sigma}$ remain a positive definite matrix. One way often suggested is to construct the Cholesky decomposition of $\mathbf{\Sigma}$, $\mathbf{LL}'$, where $\mathbf{L}$ is a lower triangular matrix, and estimate the elements of $\mathbf{L}$. The normalizations and zero restrictions can be imposed by making the last row of the $J\times J$ matrix $\mathbf{\Sigma}$ equal $(0,0,\ldots,1)$ and using $\mathbf{LL}'$ to create the upper $(J-1)\times(J-1)$ matrix. The additional normalization restriction is obtained by imposing $\mathbf{L}_{11}=1$.

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.

18.2.7 THE MIXED LOGIT MODEL

Another variant of the multinomial logit model is the **random parameters logit model (RPL)** (also called the **mixed logit model**). [See Revelt and Train (1996); Bhat (1996); Berry, Levinsohn, and Pakes (1995); Jain, Vilcassim, and Chintagunta (1994); and Hensher and Greene (2004).] Train's (2003) 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:

$$\beta_{ik}=\boldsymbol{\beta}_k+\mathbf{z}_i'\boldsymbol{\theta}_k+\sigma_k u_{ik}, \qquad \textbf{(18-11)}$$

where $u_{ik}, k=1,\ldots,K$, is multivariate normally distributed with correlation matrix $\mathbf{R}$, σ_k is the standard deviation of the kth 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 model is estimated 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,

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

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

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

which can be estimated by simulation, using

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

where $\mathbf{u}_{ir}$ 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.

The mixed model enjoys two considerable advantages not available in any of the other forms suggested. In a panel data or repeated-choices setting (see Section 18.2.11), one can formulate a random effects model simply by making the variation in the coefficients time invariant. Thus, the model is changed to

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

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.[6] A second degree of flexibility is in (18-11). The random components, u_i are not restricted to normality. Other distributions that can be simulated will be appropriate when the range of parameter variation consistent with consumer behavior must be restricted, for example to narrow ranges or to positive values.

18.2.8 A GENERALIZED MIXED LOGIT MODEL

The development of functional forms for multinomial choice models begins with the conditional (now usually called the multinomial) logit model that we considered in Section 18.2.3. Subsequent proposals including the multinomial probit and nested logit models (and a wide range of variations on these themes) were motivated by a desire to extend the model beyond the IIA assumptions. These were achieved by allowing correlation across the utility functions or heteroscedasticity such as that in the heteroscedastic extreme value model in (18-12). That issue has been settled in the current generation of multinomial choice models, culminating with the mixed logit model that appears to provide all the flexibility needed to depart from the IIA assumptions. [See McFadden and Train (2000) for a strong endorsement of this idea.]

[6] See Hensher (2001) for an application to transportation mode choice in which each individual is observed in several choice situations. A stated choice experiment in which consumers make several choices in sequence about automobile features appears in Hensher, Rose, and Greene (2006).

Recent research in choice modeling has focused on enriching the models to accommodate individual heterogeneity in the choice specification. To a degree, including observable characteristics, such as household income in our application to follow, serves this purpose. In this case, the observed heterogeneity enters the deterministic part of the utility functions. The heteroscedastic HEV model shown in (18-13) moves the observable heterogeneity to the scaling of the utility function instead of the mean. The mixed logit model in (18-11) accommodates both observed and unobserved heterogeneity in the preference parameters. A recent thread of research including Keane (2006), Feibig, Keane, Louviere, and Wasi (2009), and Greene and Hensher (2010) has considered functional forms that accommodate individual heterogeneity in both taste parameters (marginal utilities) and overall scaling of the preference structure. Keane et al.'s generalized mixed logit model is

$$\begin{aligned} U_{i,j} &= \mathbf{x}_{ij}'\boldsymbol{\beta}_i + \varepsilon_{ij}, \\ \boldsymbol{\beta}_i &= \sigma_i\boldsymbol{\beta} + [\gamma + \sigma_i(1-\gamma)]\mathbf{v}_i \\ \sigma_i &= \exp[\bar{\sigma} + \tau w_i] \end{aligned}$$

where $0 \leq \gamma \leq 1$ and w_i is an additional source of unobserved random variation in preferences. In this formulation, the weighting parameter, γ, distributes the individual heterogeneity in the preference weights, $\mathbf{v}_i$ and the overall scaling parameter σ_i. Heterogeneity across individuals in the overall scaling of preference structures is introduced by a nonzero τ while $\bar{\sigma}$ is chosen so that $E_w[\sigma_i] = 1$. Greene and Hensher (2010) proposed including the observable heterogeneity already in the mixed logit model, and adding it to the scaling parameter as well. Also allowing the random parameters to be correlated (via the nonzero elements in $\boldsymbol{\Gamma}$), produces a multilayered form of the generalized mixed logit model,

$$\begin{aligned} \boldsymbol{\beta}_i &= \sigma_i[\boldsymbol{\beta} + \boldsymbol{\Delta}\mathbf{z}_i] + [\gamma + \sigma_i(1-\gamma)]\boldsymbol{\Gamma}\mathbf{v}_i \\ \sigma_i &= \exp[\bar{\sigma} + \boldsymbol{\delta}'\mathbf{h}_i + \tau w_i]. \end{aligned}$$

Ongoing research has continued to produce refinements that will accommodate realistic forms of individual heterogeneity in the basic multinomial logit framework.

18.2.9 APPLICATION: CONDITIONAL LOGIT MODEL FOR TRAVEL MODE CHOICE

Hensher and Greene [Greene (2007a)] 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 F18.2.) 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 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 18.2. 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 18.2, is dominated by drivers.

TABLE 18.2 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	41.657	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.

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 18.3. The model was fit with and without the corrections for choice-based sampling. Because 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 18.4. The predictions are generated by tabulating the integer parts of $m_{jk} = \sum_{i=1}^{210} \hat{p}_{ij} d_{ik}$, $j, k =$ *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. Because 58 respondents chose this mode,

TABLE 18.3 Parameter Estimates

	Unweighted Sample		*Choice-Based Weighting*	
	Estimate	t *Ratio*	*Estimate*	t *Ratio*
β_G	−0.015501	−3.517	−0.01333	−2.711
β_T	−0.09612	−9.207	−0.13405	−5.216
γ_H	0.01329	1.295	−0.00108	−0.097
α_{air}	5.2074	6.684	6.5940	4.075
α_{train}	3.8690	8.731	3.6190	4.317
α_{bus}	3.1632	7.025	3.3218	3.822
Log-likelihood at $\boldsymbol{\beta} = 0$	−291.1218		−291.1218	
Log-likelihood (sample shares)	−283.7588		−218.9929	
Log-likelihood at convergence	−199.1284		−147.5896	

TABLE 18.4 Predicted Choices Based on Model Probabilities (predictions based on choice-based sampling 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 (14)	6 (12)	30
Car	16 (5)	13 (5)	6 (3)	25 (45)	59
Total (Predicted)	58 (39)	63 (40)	30 (23)	59 (108)	210

TABLE 18.5 Results for IIA Test

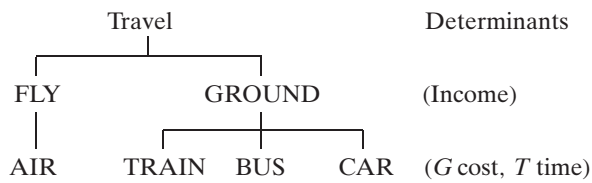
	Full-Choice Set				*Restricted-Choice Set*			
	β_G	β_T	α_{train}	α_{bus}	β_G	β_T	α_{train}	α_{bus}
Estimate	−0.0155	−0.0961	3.869	3.163	−0.0639	−0.0699	4.464	3.105
	Estimated Asymptotic Covariance Matrix				*Estimated Asymptotic Covariance Matrix*			
β_G	0.194e-4				0.000101			
β_T	−0.46e-6	0.000109			−0.000013	0.000221		
α_{train}	−0.00060	−0.0038	0.196		−0.00244	−0.00759	0.410	
α_{bus}	−0.00026	−0.0038	0.161	0.203	−0.00113	−0.00753	0.336	0.371

Note: 0.nnne-p indicates times 10 to the negative p power.
$H = 33.3367$. Critical chi-squared[4] = 9.488.

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 18.5

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.

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



Note that one of the branches has only a single choice, so the conditional probability, $P_{j|fly} = P_{air|fly} = 1$. The estimates marked "unconditional" in Table 18.6 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

TABLE 18.6 Estimates of a Mode Choice Model (standard errors in parentheses)

Parameter	*FIML Estimate*		*Unconditional*	
α_{air}	6.042	(1.199)	5.207	(0.779)
α_{bus}	4.096	(0.615)	3.163	(0.450)
α_{train}	5.065	(0.662)	3.869	(0.443)
β_{GC}	−0.03159	(0.00816)	−0.1550	(0.00441)
β_{TTME}	−0.1126	(0.0141)	−0.09612	(0.0104)
γ_H	0.01533	(0.00938)	0.01329	(0.0103)
τ_{fly}	0.5860	(0.141)	1.0000	(0.000)
τ_{ground}	0.3890	(0.124)	1.0000	(0.000)
σ_{fly}	2.1886	(0.525)	1.2825	(0.000)
σ_{ground}	3.2974	(1.048)	1.2825	(0.000)
$\ln L$	−193.6561		−199.1284	

full log-likelihood for the nested logit model. In this model,

$$\begin{aligned}
\text{Prob}(choice \mid branch) &= P(\alpha_{air}d_{air} + \alpha_{train}d_{train} + \alpha_{bus}d_{bus} + \beta_G GC + \beta_T TTME),\\
\text{Prob}(branch) &= P(\gamma d_{air} HINC + \tau_{fly} IV_{fly} + \tau_{ground} IV_{ground}),\\
\text{Prob}(choice, branch) &= \text{Prob}(choice \mid branch) \times \text{Prob}(branch).
\end{aligned}$$

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 choice model was reestimated under the assumptions of a heteroscedastic extreme value (HEV) specification. In its simplest form, this model allows a separate variance,

$$\sigma_j^2 = \pi^2/(6\theta_j^2) \qquad \textbf{(18-12)}$$

for each ε_{ij} in (18-1). (One of the θ's must be normalized to 1.0 because we can only compare ratios of variances.) The results for this model are shown in Table 18.7. This model is less restrictive than the nested logit model. To make them comparable, we note that we found that $\sigma_{air} = \pi/(\tau_{fly}\sqrt{6}) = 2.1886$ and $\sigma_{train} = \sigma_{bus} = \sigma_{car} = \pi/(\tau_{ground}\sqrt{6}) = 3.2974$. The HEV model thus relaxes an additional restriction because it has three free variances whereas the nested logit model has two. On the other hand, the important degree of freedom is that the HEV model does not impose the IIA assumptions anywhere in the choices, whereas the nexted logit does, within each branch. Table 18.7 contains two additional results for HEV specifications. In the one denoted "Heteroscedastic HEV Model," we have allowed heteroscedasticity across individuals as well as across

TABLE 18.7 Estimates of a Heteroscedastic Extreme Value Model (standard errors in parentheses)

Parameter	*HEV Model*		*Heteroscedastic HEV Model*		*Restricted HEV Model*		*Nested Logit Model*	
α_{air}	7.8326	(10.951)	5.1815	(6.042)	2.973	(0.995)	6.062	(1.199)
α_{bus}	7.1718	(9.135)	5.1302	(5.132)	4.050	(0.494)	4.096	(0.615)
α_{train}	6.8655	(8.829)	4.8654	(5.071)	3.042	(0.429)	5.065	(0.662)
β_{GC}	−0.05156	(0.0694)	−0.03326	(0.0378)	−0.0289	(0.00580)	−0.03159	(0.00816)
β_{TTME}	−0.1968	(0.288)	−0.1372	(0.164)	−0.0828	(0.00576)	−0.1126	(0.0141)
γ	0.04024	(0.0607)	0.03557	(0.0451)	0.0238	(0.0186)	0.01533	(0.00938)
τ_{fly}							0.5860	(0.141)
τ_{ground}							0.3890	(0.124)
θ_{air}	0.2485	(0.369)	0.2890	(0.321)	0.4959	(0.124)		
θ_{train}	0.2595	(0.418)	0.3629	(0.482)	1.0000	(0.000)		
θ_{bus}	0.6065	(1.040)	0.6895	(0.945)	1.0000	(0.000)		
θ_{car}	1.0000	(0.000)	1.0000	(0.000)	1.0000	(0.000)		
ϕ	0.0000	(0.000)	0.00552	(0.00573)	0.0000	(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		−194.5107		−200.3791		−193.6561	

choices by specifying

$$\theta_{ij} = \theta_j \times \exp(\phi HINC_i). \quad \textbf{(18-13)}$$

[See Salisbury and Feinberg (2010) and Louviere and Swait (2010) for an application of this type of HEV model.]

In the "Restricted HEV Model," the variance of $\varepsilon_{i,Air}$ is allowed to differ from the others. Finally, the nested logit model has different variance for *Air* and (*Train, Bus, Car*).

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 18.8 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 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 18.9 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

TABLE 18.8 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

TABLE 18.9 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 $\theta_{air} = 3.798$, $\theta_{train} = 1.182$, $\theta_{bus} = 0.0712$, $\theta_{car} = 0.000$.

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 18.7 and (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 matrix 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 18.9 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, although 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, the question still warrants additional study. Both models are making their way into the applied literature.

18.2.10 ESTIMATING WILLINGNESS TO PAY

One of the standard applications of choice models is to estimate how much consumers value the attributes of the choices. Recall that we are not able to observe the scale of the utilities in the choice model. However, we can use the marginal utility of income, also scaled in the same unobservable way, to effect the valuation. In principle, we could estimate

$$\begin{aligned}\text{WTP} &= (\text{Marginal Utility of Attribute}/\sigma)/(\text{Marginal Utility of Income}/\sigma)\\ &= (\beta_{\text{attribute}}/\sigma)/(\gamma_{\text{Income}}/\sigma),\end{aligned}$$

where σ is the unknown scaling of the utility functions. Note that σ cancels out of the ratio. In our application, for example, we might assess how much consumers would be willing to pay to have shorter waits at the terminal for the public modes of transportation by using

$$\text{WTP}_{\text{time}} = -\beta_{\widehat{\text{TIME}}}/\gamma_{\text{Income}}.$$

(We use the negative because additional time spent waiting at the terminal provides disutility, as evidenced by its coefficient's negative sign.) In settings in which income is not observed, researchers often use the negative of the coefficient on a cost variable as a proxy for the marginal utility of income. Standard errors for estimates of WTP can be computed using the delta method or the method of Krinsky and Robb. (See Sections 4.4.4 and 15.3.)

In the basic multinomial logit model, the estimator of WTP is a simple ratio of parameters. In our estimated model in Table 18.3, for example, using the household income coefficient as the numeraire, the estimate of WTP for a shorter wait at the terminal is $-0.09612/0.01329 = 7.239$. The units of measurement must be resolved in this computation, since terminal time is measured in minutes while the cost is in \$1,000/year. Multiplying this result by \$60 minutes/hour and dividing by the equivalent hourly income of income times 8,760/1,000 gives \$49.54 per hour of waiting time. To compute the estimated asymptotic standard error, for convenience, we first rescaled the terminal time to hours by dividing it by 60 and the income variable to \$/hour by multiplying it by 1,000/8,760. The resulting estimated asymptotic distribution for the estimators is

$$\begin{pmatrix} \hat{\beta}_{\widehat{TIME}} \\ \hat{\gamma}_{HINC} \end{pmatrix} \sim N\left[\begin{pmatrix} -5.76749 \\ 0.11639 \end{pmatrix}, \begin{pmatrix} 0.392365 & 0.00193095 \\ 0.00193095 & 0.00808177 \end{pmatrix} \right].$$

The derivatives of $\text{WTP}_{\widehat{TIME}} = -\beta_{\widehat{TIME}}/\gamma_H$ are $-1/\gamma_H$ for β_{TIME} and –WTP/γ_H for γ_H. This provides an estimator of 38.8304 for the standard error. The confidence interval for this parameter would be -26.56 to $+125.63$. This seems extremely wide. We will return to this issue later.

In the mixed logit model, if either of the coefficients in the computation is random, then the preceding simple computation above will not reveal the heterogeneity in the result. In many studies of WTP using mixed logit models, it is common to allow the utility parameter on the attribute (numerator) to be random and treat the numeraire (income or cost coefficient) as nonrandom. Using our mode choice application, we refit the model with $\beta_{\widehat{TIME},i} = \beta_{\widehat{TIME}} + \sigma_{\widehat{TIME}} v_i$ and all other coefficients nonrandom. We then used the method described in Section 15.10 to estimate $\text{E}[\beta_{\widehat{TIME},i}|\mathbf{X}_i, choice_i]/\gamma_H$ to estimate the expected WTP for each individual in the sample. Income and terminal time were scaled as before. Figure 18.1 displays a kernel estimator of the estimates of WTP_i by this method. Note that the distribution is roughly centered on our earlier estimate of \$49.53. The density estimator reveals the heterogeneity in the population of this parameter.

Willingness to pay measures computed as suggested above are ultimately based on a ratio of two asymptotically normally distributed parameter estimators. In general, ratios of normally distributed random variables do not have a finite variance. This often becomes apparent when using the delta method, as it seems previously. A number of writers, notably, Daly, Hess, and Train (2009), have documented the problem of extreme

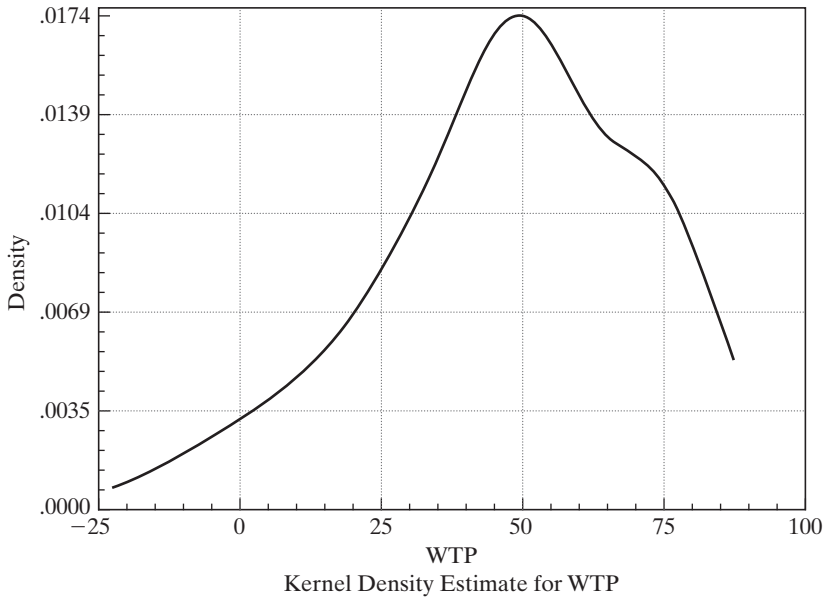


FIGURE 18.1 Estimated Willingness to Pay for Decreased Terminal Time.

results of WTP computations, and why they should be expected. One solution suggested, for example, by Train and Weeks (2005), Sonnier, Ainsle, and Otter (2007), and Scarpa, Thiene, and Train (2008), is to recast the original model in **willingness to pay space**. In the multinomial logit case, this amounts to a trivial reparameterization of the model. Using our application as an example, we would write

$$\begin{aligned} U_{ij} &= \alpha_j + \beta_{GC}[\mathbf{GC}_i + \beta_{\widehat{TIME}}/\beta_{GC} TIME_i] + \gamma_H A_{AIR} HINC_i + \varepsilon_{ij} \\ &= \alpha_j + \beta_{GC}[\mathbf{GC}_i + \lambda_{\widehat{TIME}} TIME_i] + \gamma_H A_{AIR} HINC_i + \varepsilon_{ij}. \end{aligned}$$

This obviously returns the original model, though in the process, it transforms a linear estimation problem into a nonlinear one. But, in principle, with the model reparameterized in "WTP space," we have sidestepped the problem noted earlier – $\lambda_{\widehat{TIME}}$ is the estimator of WTP with no further transformation of the parameters needed. As noted, this will return the numerically identical results for a multinomial logit model. It will not return the identical results for a mixed logit model, in which we write $\lambda_{TIME,i} = \lambda_{\widehat{TIME}} + \theta_{\widehat{TIME}} v_{\widehat{TIME},i}$. Greene and Hensher (2010b) apply this method to the generalized mixed logit model in Section 18.2.8.

18.2.11 PANEL DATA AND STATED CHOICE EXPERIMENTS

Panel data in the unordered discrete choice setting typically come in the form of sequential choices. Train (2009, Chapter 6) reports an analysis of the site choices of 258 anglers who chose among 59 possible fishing sites for a total of 962 visits. Allenby and Rossi (1999) modeled brand choice for a sample of shoppers who made multiple store trips. The mixed logit model is a framework that allows the counterpart to a random

effects model. The random utility model would appear

$$U_{ij,t} = \mathbf{x}'_{ij,t}\boldsymbol{\beta}_i + \varepsilon_{ij,t},$$

where conditioned on $\boldsymbol{\beta}_i$, a multinomial logit model applies. The random coefficients carry the common effects across choice situations. For example, if the random coefficients include choice-specific constant terms, then the random utility model becomes essentially a random effects model. A modification of the model that resembles Mundlak's correction for the random effects model is

$$\boldsymbol{\beta}_i = \boldsymbol{\beta}^0 + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\Gamma}\mathbf{u}_i,$$

where, typically, $\mathbf{z}_i$ would contain demographic and socioeconomic information.

The **stated choice experiment** is similar to the repeated choice situation, with a crucial difference. In a stated choice survey, the respondent is asked about his or her preferences over a series of hypothetical choices, often including one or more that are actually available and others that might not be available (yet). Hensher, Rose, and Greene (2006) describe a survey of Australian commuters who were asked about hypothetical commutation modes in a choice set that included the one they currently took and a variety of alternatives. Revelt and Train (2000) analyzed a stated choice experiment in which California electricity consumers were asked to choose among alternative hypothetical energy suppliers. The advantage of the stated choice experiment is that it allows the analyst to study choice situations over a range of variation of the attributes or a range of choices that might not exist within the observed, actual outcomes. Thus, the original work on the MNL by McFadden et al. concerned survey data on whether commuters would ride a (then-hypothetical) underground train system to work in the San Francisco Bay area. The disadvantage of **stated choice data** is that they are hypothetical. Particularly when they are mixed with **revealed preference data**, the researcher must assume that the same preference patterns govern both types of outcomes. This is likely to be a dubious assumption. One method of accommodating the mixture of underlying preferences is to build different scaling parameters into the model for the stated and revealed preference components of the model. Greene and Hensher (2007) suggest a nested logit model that groups the hypothetical choices in one branch of a tree and the observed choices in another.

18.2.12 AGGREGATE MARKET SHARE DATA—THE BLP RANDOM PARAMETERS MODEL

We note, finally, an important application of the mixed logit model, the structural demand model of Berry, Levinsohn, and Pakes (1995). (Demand models for differentiated products such as automobiles [BLP (1995), Goldberg (1995)], ready-to-eat cereals [Nevo (2001)], and consumer electronics [Das, Olley, and Pakes (1996)], have been constructed using the mixed logit model with market share data.[7] A basic structure is defined for

Markets, denoted $t = 1, \ldots, T$

Consumers in the markets, denoted $i = 1, \ldots, n_t$

Products, denoted $j = 1, \ldots, J$

[7] We draw heavily on Nevo (2000) for this discussion.

The definition of a market varies by application; BLP analyzed the U.S. national automobile market for 20 years; Nevo examined a cross section of cities over 20 quarters so the city-quarter is a market; Das et al. defined a market as the annual sales to consumers in particular income levels.

For market t, we base the analysis on average prices, p_{jt}, aggregate quantities q_{jt}, consumer incomes y_i observed product attributes, $\mathbf{x}_{jt}$ and unobserved (by the analyst) product attributes, Δ_{jt}. The indirect utility function for consumer i, for product j in market t is

$$u_{ijt} = \alpha_i(y_i - p_{jt}) + \mathbf{x}_{jt}'\boldsymbol{\beta}_i + \Delta_{jt} + \varepsilon_{ijt}, \tag{18-14}$$

where α_i is the marginal utility of income and $\boldsymbol{\beta}_i$ are marginal utilities attached to specific observable attributes of the products. The fact that some unobservable product attributes, Δ_{jt} will be reflected in the prices implies that prices will be endogenous in a demand model that is based on only the observable attributes. Heterogeneity in preferences is reflected (as we did earlier) in the formulation of the random parameters,

$$\begin{pmatrix} \alpha_i \\ \boldsymbol{\beta}_i \end{pmatrix} = \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\pi}' \\ \boldsymbol{\Pi} \end{pmatrix} \mathbf{d}_i + \begin{pmatrix} \gamma w_i \\ \boldsymbol{\Gamma}\mathbf{v}_i \end{pmatrix} \tag{18-15}$$

where $\mathbf{d}_i$ is a vector of demographics such as gender and age while $\alpha, \boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\Pi}, \gamma$, and $\boldsymbol{\Gamma}$ are structural parameters to be estimated (assuming they are identified). A utility function is also defined for an "outside good" that is (presumably) chosen if the consumer chooses none of the brands $1, \ldots, J$:

$$u_{i0t} = \alpha_i y_i + \Delta_{0t} + \boldsymbol{\pi}_0'\mathbf{d}_i + \varepsilon_{i0t}.$$

Since there is no variation in income across the choices, $\alpha_i y_i$ will fall out of the logit probabilities, as we saw earlier. A normalization is used instead, $u_{i0t} = \varepsilon_{i0t}$, so that comparisons of utilities are against the outside good. The resulting model can be reconstructed by inserting (18-15) into (18-14),

$$\begin{aligned} u_{ijt} &= \alpha_i y_i + \delta_{jt}(\mathbf{x}_{jt},\ p_{jt},\ \Delta_{jt} : \alpha, \boldsymbol{\beta}) + \tau_{ijt}(\mathbf{x}_{jt}, p_{jt}, \mathbf{v}_i, w_i : \boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\gamma}, \boldsymbol{\Gamma}) + \varepsilon_{ijt} \\ \delta_{jt} &= \mathbf{x}_{jt}'\boldsymbol{\beta} - \alpha p_{jt} + \Delta_{jt} \\ \tau_{jt} &= [-p_{jt}, \mathbf{x}_{jt}'] \left[\begin{pmatrix} \boldsymbol{\pi}' \\ \boldsymbol{\Pi} \end{pmatrix} d_i + \begin{pmatrix} \gamma w_i \\ \boldsymbol{\Gamma}\mathbf{v}_i \end{pmatrix} \right]. \end{aligned}$$

The preceding model defines the random utility model for consumer i in market t. Each consumer is assumed to purchase the one good that maximizes utility. The market share of the jth product in this market is obtained by summing over the choices made by those consumers. With the assumption of homogeneous tastes ($\boldsymbol{\Gamma} = \mathbf{0}$ and $\gamma = 0$) and i.i.d., type I extreme value distributions for ε_{ijt}, it follows that the market share of product j is

$$s_{jt} = \frac{\exp(\mathbf{x}_{jt}'\boldsymbol{\beta} - \alpha p_{jt} + \Delta_{jt})}{1 + \sum_{k=1}^{J} \exp(\mathbf{x}_{kt}'\boldsymbol{\beta} - \alpha p_{kt} + \Delta_{kt})}.$$

The IIA assumptions produce the familiar problems of peculiar and unrealistic substitution patterns among the goods. Alternatives considered include a nested logit, a "generalized extreme value" model and, finally, the mixed logit model, now applied to the aggregate data.

Estimation cannot proceed along the lines of Section 18.2.7 because Δ_{jt} is unobserved and p_{jt} is, therefore, endogenous. BLP propose, instead to use a GMM estimator, based on the moment equations

$$E\{[S_{jt} - s_{jt}(\mathbf{x}_{jt}, p_{jt}|\alpha, \boldsymbol{\beta})]\mathbf{z}_{jt}\} = 0$$

for a suitable set of instruments. Layering in the random parameters specification, we obtain an estimation based on **method of simulated moments**, rather than a maximum simulated log likelihood. The simulated moments would be based on

$$E_{w,v}[s_{jt}(\mathbf{x}_{jt}, p_{jt}|\alpha_i, \boldsymbol{\beta}_i)] = \int_{w,v} \{s_{jt}[\mathbf{x}_{jt}, p_{jt}|\alpha_i(w), \boldsymbol{\beta}_i(v)]\}\, dF(w)\, dF(v).$$

These would be simulated using the method of Section 18.2.7.

18.3 RANDOM UTILITY MODELS FOR ORDERED CHOICES

The analysts at bond rating agencies such as Moody's and Standard and Poor provide an evaluation of the quality of a bond that is, in practice, a discrete listing of the continuously varying underlying features of the security. The rating scales are as follows:

Rating	*S&P Rating*	*Moody's Rating*
Highest quality	AAA	Aaa
High quality	AA	Aa
Upper medium quaity	A	A
Medium grade	BBB	Baa
Somewhat speculative	BB	Ba
Low grade, speculative	B	B
Low grade, default possible	CCC	Caa
Low grade, partial recovery possible	CC	Ca
Default, recovery unlikely	C	C

For another example, *Netflix* (www.netflix.com) is an Internet company that rents movies. Subscribers order the film online for download or home delivery of a DVD. The next time the customer logs onto the web site, they are invited to rate the movie on a five-point scale, where five is the highest, most favorable rating. The ratings of the many thousands of subscribers who rented that movie are averaged to provide a recommendation to prospective viewers. As of April 5, 2009, the average rating of the 2007 movie *National Treasure: Book of Secrets* given by approximately 12,900 visitors to the site was 3.8. Many other Internet sellers of products and services, such as Barnes and Noble, Amazon, Hewlett Packard, and Best Buy, employ rating schemes such as this. Many recently developed national survey data sets, such as the British Household Panel Data Set (BHPS) (http://www.iser.essex.ac.uk/survey/bhps) and the German Socioeconomic Panel (GSOEP) (http://www.diw.de/en/soep), contain questions that elicit self-assessed ratings of health, health satisfaction, or overall well-being. Like the other examples listed, these survey questions are answered on a discrete scale, such as the

zero to 10 scale of the question about health satisfaction in the GSOEP. Ratings such as these provide applications of the models and methods that interest us in this section.[8]

For any individual respondent, we hypothesize that there is a continuously varying strength of preferences that underlies the rating they submit. For convenience and consistency with what follows, we will label that strength of preference "utility," U^*. Continuing the Netflix example, we describe utility as ranging over the entire real line:

$$-\infty < U_{im}^* < +\infty$$

where i indicates the individual and m indicates the movie. Individuals are invited to "rate" the movie on an integer scale from 1 to 5. Logically, then, the translation from underlying utility to a rating could be viewed as a *censoring* of the underlying utility,

$$\begin{aligned}
R_{im} &= 1 \text{ if } -\infty < U_{im}^* \leq \mu_1,\\
R_{im} &= 2 \text{ if } \mu_1 < U_{im}^* \leq \mu_2,\\
R_{im} &= 3 \text{ if } \mu_2 < U_{im}^* \leq \mu_3,\\
R_{im} &= 4 \text{ if } \mu_3 < U_{im}^* \leq \mu_4,\\
R_{im} &= 5 \text{ if } \mu_4 < U_{im}^* < \infty.
\end{aligned}$$

The same mapping would characterize the bond ratings, since the qualities of bonds that produce the ratings will vary continuously and the self-assessed health and well-being questions in the panel survey data sets based on an underlying utility or preference structure. The crucial feature of the description thus far is that underlying the discrete response is a continuous range of preferences. Therefore, the observed rating represents a censored version of the true underlying preferences. Providing a rating of five could be an outcome ranging from general enjoyment to wild enthusiasm. Note that the *thresholds*, μ_j, number $(J-1)$ where J is the number of possible ratings (here, five) – $J-1$ values are needed to divide the range of utility into J cells. The thresholds are an important element of the model; they divide the range of utility into cells that are then identified with the observed outcomes. Importantly, the difference between two levels of a rating scale (e.g., one compared to two, two compared to three) is not the same as on a utility scale; hence we have a strictly nonlinear transformation captured by the thresholds, which are estimable parameters in an **ordered choice model**.

The model as suggested thus far provides a crude description of the mechanism underlying an observed rating. Any individual brings their own set of *characteristics* to the utility function, such as age, income, education, gender, where they live, family situation, and so on, which we denote $x_{i1}, x_{i2}, \ldots, x_{iK.}$ They also bring their own aggregate of unmeasured and unmeasurable (by the statistician) idiosyncrasies, denoted ε_{im} How these features enter the utility function is uncertain, but it is conventional to use a linear function, which produces a familiar *random utility function*,

$$U_{im}^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K \mathbf{x}_{iK} + \varepsilon_{im}.$$

[8] Greene and Hensher (2010) provide a survey of ordered choice modeling. Other textbook and monograph treatments include DeMaris (2004), Long (1997), Johnson and Abbot (1999), and Long and Freese (2006). Introductions to the model also appear in journal articles such as Winship and Mare (1984), Becker and Kennedy (1992), Daykin and Moffatt (2002), and Boes and Winkelmann (2006).

Example 18.2 Movie Ratings

The web site www.imdb.com invites visitors to rate movies that they have seen, in the same fashion as the Netflix site. This site uses a 10 point scale. On December 1, 2008, they reported the results in Figure 18.2 for the movie *National Treasure: Book of Secrets* for 41,771 users of the site: The earlier panel at the left shows the overall ratings. The panel at the right shows how the average rating varies across age, gender, and whether the rater is a U.S. viewer or not.

The rating mechanism we have constructed is

$$R_{im} = 1 \text{ if } -\infty < \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{im} \leq \mu_1,$$
$$R_{im} = 2 \text{ if } \mu_1 < \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{im} \leq \mu_2,$$
$$R_{im} = 3 \text{ if } \mu_2 < \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{im} \leq \mu_3,$$
$$R_{im} = 4 \text{ if } \mu_3 < \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{im} \leq \mu_4,$$
$$R_{im} = 5 \text{ if } \mu_4 < \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{im} < \infty.$$

Relying on a central limit to aggregate the innumerable small influences that add up to the individual idiosyncrasies and movie attraction, we assume that the random component, ε_{im}, is normally distributed with zero mean and (for now) constant variance. The assumption of normality will allow us to attach probabilities to the ratings. In particular, arguably the most interesting one is

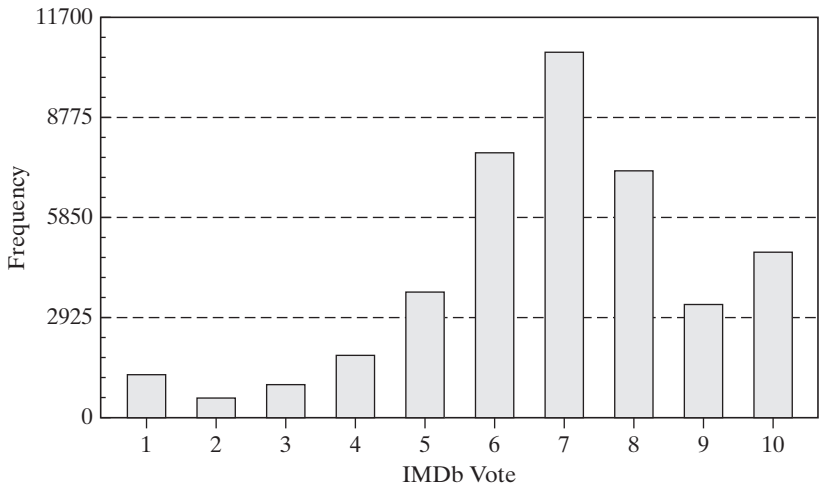
$$\text{Prob}(R_{im} = 5 \mid \mathbf{x}_i) = \text{Prob}[\varepsilon_{im} > \mu_4 - \mathbf{x}_i'\boldsymbol{\beta}].$$

The structure provides the framework for an econometric model of how individuals rate movies (that they rent from Netflix). The resemblance of this model to familiar models of binary choice is more than superficial. For example, one might translate this econometric model directly into a probit model by focusing on the variable

$$E_{im} = 1 \text{ if } R_{im} = 5$$
$$E_{im} = 0 \text{ if } R_{im} < 5.$$

Thus, the model is an extension of a binary choice model to a setting of more than two choices. But, the crucial feature of the model is the ordered nature of the observed outcomes and the correspondingly ordered nature of the underlying preference scale.

FIGURE 18.2 IMDb.com Ratings (www.imdb.com/title/tt0465234/ratings).



The model described here is an *ordered choice model*. (The choice of the normal distribution for the random term makes it an *ordered probit model*.) Ordered choice models are appropriate for a wide variety of settings in the social and biological sciences. The essential ingredient is the *mapping from an underlying, naturally ordered preference scale to a discrete ordered observed outcome*, such as the rating scheme described. The model of ordered choice pioneered by Aitcheson and Silvey (1957), Snell (1964), and Walker and Duncan (1967) and articulated in its modern form by Zavoina and McElvey (1975) has become a widely used tool in many fields. The number of applications in the current literature is large and increasing rapidly, including

- Bond ratings [Terza (1985a)]
- Congressional voting on a Medicare bill [McElvey and Zavoina (1975)]
- Credit ratings [Cheung (1996) , Metz, and Cantor (2006)]
- Driver injury severity in car accidents [Eluru, Bhat, and Hensher (2008)]
- Drug reactions [Fu, Gordon, Liu, Dale, and Christensen.(2004)]
- Education [Machin and Vignoles (2005), Carneiro, Hansen, and Heckman (2003), Cunha, Heckman, and Navarro (2007)]
- Financial failure of firms [Hensher and Jones (2007)]
- Happiness [Winkelmann (2005), Zigante (2007)]
- Health status [Jones, Koolman, and Rice (2003)]
- Life satisfaction [Clark, Georgellis, and Sanfey (2001), Groot and ven den Brink (2003)]
- Monetary policy [Eichengreen, Watson, and Grossman (1985)]
- Nursing labor supply [Brewer, Kovner, Greene, and Cheng (2008)]
- Obesity [Greene, Harris, Hollingsworth, and Maitra (2008)]
- Political efficacy [King, Murray, Salomon, and Tandon [2004)]
- Pollution [Wang and Kockelman (2009)]
- Promotion and rank in nursing [Pudney and Shields [2000)]
- Stock price movements [Tsay (2005)]
- Tobacco use [Harris and Zhao (2007), Kasteridis, Munkin, and Yen (2008)]
- Work disability [Kapteyn et al. (2007)]

18.3.1 THE ORDERED PROBIT MODEL

The ordered probit 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

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

which is a form of censoring. The μ's are unknown parameters to be estimated with $\boldsymbol{\beta}$.

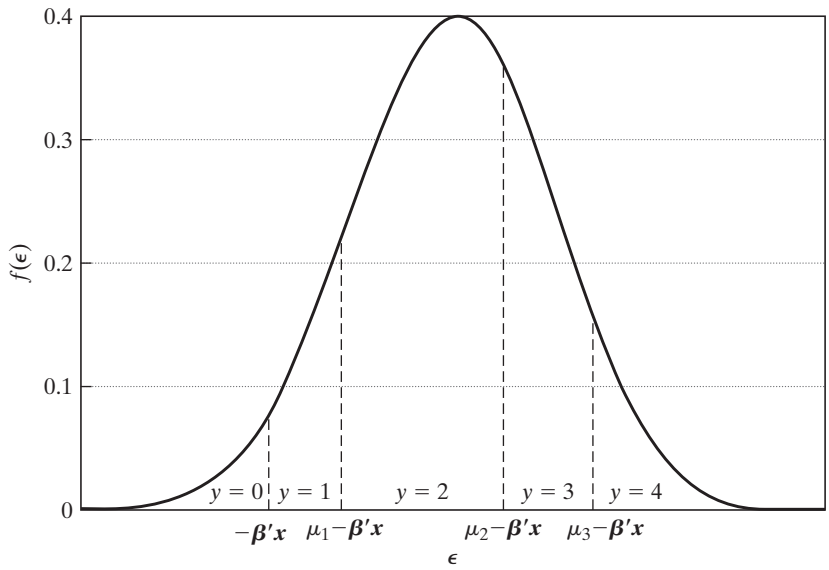


FIGURE 18.3 Probabilities in the Ordered Probit Model.

We assume that ε is normally distributed across observations.[9] 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:

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

For all the probabilities to be positive, we must have

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

Figure 18.3 shows the implications of the structure. This is an extension of the univariate probit model we examined in chapter 17. 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

[9] 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 18.4 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

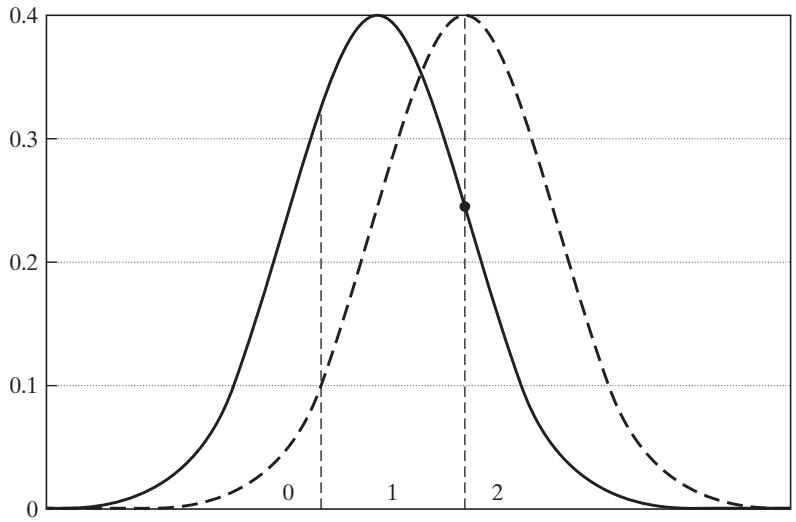


FIGURE 18.4 Effects of Change in x on Predicted Probabilities.

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.

Example 18.3 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." Because 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 Armed Forces 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 data used in this study are not available for distribution.) The sample size was 5,641. The results are reported in Table 18.10. The extremely large t ratio on the AFQT score is to be expected, as 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

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

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

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

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 18.11.

TABLE 18.10 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	—

TABLE 18.11 Marginal Effect of a Binary Variable

	$-\hat{\beta}'\mathbf{x}$	$\hat{\mu}-\hat{\beta}'\mathbf{x}$	*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

18.3.2 A SPECIFICATION TEST FOR THE ORDERED CHOICE MODEL

The basic formulation of the ordered choice model implies that for constructed binary variables,

$$w_{ij} = 1 \text{ if } y_i \leq j, 0 \text{ otherwise, } j = 1, 2, \ldots, J-1, \qquad \textbf{(18-16)}$$
$$\text{Prob}(w_{ij} = 1 \mid \mathbf{x}_i) = F(\mathbf{x}_i'\boldsymbol{\beta} - \mu_j).$$

The first of these, when $j = 1$, is the binary choice model of Section 17.2. One implication is that we could estimate the slopes, but not the threshold parameters, in the ordered choice model just by using w_{i1} and $\mathbf{x}_i$ in a binary probit or logit model. (Note that this result also implies the validity of combining adjacent cells in the ordered choice model.) But, (18-16) also defines a set of $J-1$ binary choice models with different constants but common slope vector, $\boldsymbol{\beta}$. This equality of the parameter vectors in (18-16) has been labeled the **parallel regression assumption**. Although it is merely an implication of the model specification, this has been viewed as an implicit restriction on the model. [See, e.g., Long (1997, p. 141).] Brant (1990) suggests a test of the parallel regressions assumption based on (18-16). One can, in principle, fit $J-1$ such binary choice models separately. Each will produce its own constant term and a consistent estimator of the common $\boldsymbol{\beta}$. Brant's Wald test examines the linear restrictions $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_{J-1}$, or H_0: $\boldsymbol{\beta}_q - \boldsymbol{\beta}_1 = \mathbf{0}$, $q = 2, \ldots, J-1$. The Wald statistic will be

$$\chi^2[(J-2)K] = (\mathbf{R}\hat{\boldsymbol{\beta}}^*)'[\mathbf{R} \times Asy.Var[\hat{\boldsymbol{\beta}}^*] \times \mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}^*),$$

where $\hat{\boldsymbol{\beta}}^*$ is obtained by stacking the individual binary logit or probit estimates of $\boldsymbol{\beta}$ (without the constant terms). [See Brant (1990), Long (1997), or Greene and Hensher (2010, page 187) for details on computing the statistic.]

Rejection of the null hypothesis calls the model specification into question. An alternative model in which there is a different $\boldsymbol{\beta}$ for each value of y has two problems: it does not force the probabilities to be positive and it is internally inconsistent. On the latter point, consider the suggested latent regression, $y^* = \mathbf{x}'\boldsymbol{\beta}_j + \varepsilon$. If the "$\boldsymbol{\beta}$" is different for each j, then it is not possible to construct a data generating mechanism for y^* (or, for example, simulate it); the realized value of y^* cannot be defined without knowing y (i.e., the realized j), since the applicable $\boldsymbol{\beta}$ depends on j, but y is supposed to be determined from y^* through, for example, (18-16). There is no parametric restriction other than the one we seek to avoid that will preserve the ordering of the probabilities for all values of the data and maintain the coherency of the model. This still leaves the question of what specification failure would logically explain the finding. Some suggestions in Brant (1990) include (1) misspecification of the latent regression, $\mathbf{x}'\boldsymbol{\beta}$; (2) heteroscedasticity

of ε; and (3) misspecification of the distributional form for the latent variable, that is, "nonlogistic link function."

Example 18.4 Brant Test for an Ordered Probit Model of Health Satisfaction

In Example 17.4, we studied the health care usage of a sample of households in the German Socioeconomic Panel (GSOEP). The data include a self-reported measure of "health satisfaction," (HSAT) that is coded 0–10. This variable provides a natural application of the ordered choice models in this chapter. The data are an unbalanced panel. For purposes of this exercise, we have used the fifth (1984) wave of the data set, which is a cross section of 4,483 observations. We then collapsed the 10 cells into 5 [(0–2),(3–5), (6–8),(9),(10)] for this example. The utility function is

$$HSAT_i^* = \beta_1 + \beta_2 AGE_i + \beta_3 INCOME_i + \beta_4 KIDS_i + \beta_5 EDUC_i + \beta_6 MARRIED_i \quad \beta_7 WORKING_i + \varepsilon_i .$$

Variables KIDS, MARRIED, and WORKING, are binary indicators of whether there are children in the household, marital status, and whether the individual was working at the time of the survey. (These data are examined further in Example 18.6.) The model contains six variables, and there are four binary choice models fit, so there are $(J-2)(K) = (3)(6) = 18$ restrictions. The chi-squared for the probit model is 87.836. The critical value for 95 percent is 28.87, so the homogeneity restriction is rejected. The corresponding value for the logit model is 77.84, which leads to the same conclusion.

18.3.3 BIVARIATE ORDERED PROBIT MODELS

There are several extensions of the ordered probit model that follow the logic of the bivariate probit model we examined in Section 17.5. A direct analog to the base case two-equation model is used in the study in Example 18.3.

Example 18.5 Calculus and Intermediate Economics Courses

Butler et al. (1994) analyzed the relationship between the level of calculus attained and grades in intermediate economics courses for a sample of Vanderbilt students. The two-step estimation approach involved the following strategy. (We are stylizing the precise formulation a bit to compress the description.) Step 1 involved a direct application of the ordered probit model of Section 18.3.1 to the level of calculus achievement, which is coded 0, 1, . . . , 6:

$$\begin{aligned} m_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \ \varepsilon_i \mid \mathbf{x}_i \sim \text{N}[0, 1], \\ m_i &= 0 \text{ if } -\infty < m_i^* \leq 0 \\ &= 1 \text{ if } 0 < m_i^* \leq \mu_1 \\ &\quad \cdots \\ &= 6 \text{ if } \mu_5 < m_i^* < +\infty. \end{aligned}$$

The authors argued that although the various calculus courses can be ordered discretely by the material covered, the differences between the levels cannot be measured directly. Thus, this is an application of the ordered probit model. The independent variables in this first-step model included SAT scores, foreign language proficiency, indicators of intended major, and several other variables related to areas of study.

The second step of the estimator involves regression analysis of the grade in the intermediate microeconomics or macroeconomics course. Grades in these courses were translated to a granular continuous scale (A = 4.0, A− = 3.7, etc.). A linear regression is specified,

$$Grade_i = \mathbf{z}_i'\boldsymbol{\delta} + u_i, \quad \text{where} \quad u_i \mid \mathbf{z}_i \sim \text{N}\left[0, \sigma_u^2\right].$$

Independent variables in this regression include, among others, (1) dummy variables for which outcome in the ordered probit model applies to the student (with the zero reference case omitted), (2) grade in the last calculus course, (3) several other variables related to prior courses, (4) class size, (5) freshman GPA, and so on. The unobservables in the *Grade* equation and the math attainment are clearly correlated, a feature captured by the additional assumption that $(\varepsilon_i, u_i \mid \mathbf{x}_i, \mathbf{z}_i) \sim \mathrm{N}_2[(0,0),(1,\sigma_u^2),\rho\sigma_u]$. A nonzero ρ captures this "selection" effect. With this in place, the dummy variables in (1) have now become endogenous. The solution is a "selection" correction that we will examine in detail in Chapter 19. The modified equation becomes

$$\begin{aligned} \mathit{Grade}_i \mid m_i &= \mathbf{z}_i'\boldsymbol{\delta} + E[u_i \mid m_i] + v_i \\ &= \mathbf{z}_i'\boldsymbol{\delta} + (\rho\sigma_u)[\lambda(\mathbf{x}_i'\boldsymbol{\beta}, \mu_1, \ldots, \mu_5)] + v_i. \end{aligned}$$

They thus adopt a "control function" approach to accommodate the endogeneity of the math attainment dummy variables. [See Section 17.3.5 and (17-32) for another application of this method.] The term $\lambda(\mathbf{x}_i'\boldsymbol{\beta}, \mu_1, \ldots, \mu_5)$ is a generalized residual that is constructed using the estimates from the first-stage ordered probit model. [A precise statement of the form of this variable is given in Li and Tobias (2006).] Linear regression of the course grade on $\mathbf{z}_i$ and this constructed regressor is computed at the second step. The standard errors at the second step must be corrected for the use of the estimated regressor using what amounts to a Murphy and Topel (2002) correction. (See Section 14.7.)

Li and Tobias (2006) in a replication of and comment on Butler et al. (1994), after roughly replicating the classical estimation results with a Bayesian estimator, observe that the preceding *Grade* equation above could also be treated as an ordered probit model. The resulting **bivariate ordered probit** model would be

$$\begin{aligned} m_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, & \text{and} \qquad g_i^* &= \mathbf{z}_i'\boldsymbol{\delta} + u_i, \\ m_i &= 0 \text{ if } -\infty < m_i^* \le 0 & g_i &= 0 \text{ if } -\infty < g_i^* \le 0 \\ &= 1 \text{ if } 0 < m_i^* \le \mu_1 & &= 1 \text{ if } 0 < g_i^* \le \alpha_1 \\ &\cdots & &\cdots \\ &= 6 \text{ if } \mu_5 < m_i^* < +\infty. & &= 11 \text{ if } \mu_9 < g_i^* < +\infty \end{aligned}$$

where

$$(\varepsilon_i, u_i \mid \mathbf{x}_i, \mathbf{z}_i) \sim \mathbf{N}_2\left[(0,0),\left(1,\sigma_u^2\right),\rho\sigma_u\right].$$

Li and Tobias extended their analysis to this case simply by "transforming" the dependent variable in Butler et al.'s second equation. Computing the log-likelihood using sets of bivariate normal probabilities is fairly straightforward for the bivariate ordered probit model. [See Greene (2007).] However, the classical study of these data using the bivariate ordered approach remains to be done, so a side-by-side comparison to Li and Tobias's Bayesian alternative estimator is not possible. The endogeneity of the calculus dummy variables in (1) remains a feature of the model, so both the MLE and the Bayesian posterior are less straightforward than they might appears. Whether the results in Section 17.5.5 on the recursive bivariate probit model extend to this case also remains to be determined.

The bivariate ordered probit model has been applied in a number of settings in the recent empirical literature, including husband and wife's education levels [Magee et al. (2000)], family size [(Calhoun (1991)], and many others. In two early contributions to the field of pet econometrics, Butler and Chatterjee analyze ownership of cats and dogs (1995) and dogs and televisions (1997).

18.3.4 PANEL DATA APPLICATIONS

The ordered probit model is used to model discrete scales that represent indicators of a continuous underlying variable such as strength of preference, performance, or level of attainment. Many of the recently assembled national panel data sets contain survey questions that ask about subjective assessments of health, satisfaction, or well-being, all of which are applications of this interpretation. Examples include the following:

- The European Community Household Panel (ECHP) includes questions about job satisfaction [see D'Addio (2004)].
- The British Household Panel Survey (BHPS) includes questions about health status [see Contoyannis et al. (2004)].
- The German Socioeconomic Household Panel (GSOEP) includes questions about subjective well-being [see Winkelmann (2004)] and subjective assessment of health satisfaction [see Riphahn et al. (2003) and Example 18.4.]

Ostensibly, the applications would fit well into the ordered probit frameworks already described. However, given the panel nature of the data, it will be desirable to augment the model with some accommodation of the individual heterogeneity that is likely to be present. The two standard models, fixed and random effects, have both been applied to the analyses of these survey data.

18.3.4.a Ordered Probit Models with Fixed Effects

D'Addio et al. (2003), using methodology developed by Frijters et al. (2004) and Ferrer-i-Carbonel et al. (2004), analyzed survey data on job satisfaction using the Danish component of the European Community Household Panel. Their estimator for an ordered logit model is built around the logic of Chamberlain's estimator for the binary logit model. [See Section 17.4.4.] Because the approach is robust to individual specific threshold parameters and allows time-invariant variables, it differs sharply from the fixed effects models we have considered thus far as well as from the ordered probit model of Section 23.10.1.[10] Unlike Chamberlain's estimator for the binary logit model, however, their conditional estimator is not a function of minimal sufficient statistics. As such, the incidental parameters problem remains an issue.

Das and van Soest (2000) proposed a somewhat simpler approach. [See, as well, Long's (1997) discussion of the "parallel regressions assumption," which employs this device in a cross-section framework]. Consider the base case ordered logit model with fixed effects,

$$y_{it}^* = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \varepsilon_{it} \mid \mathbf{X}_i \sim \mathrm{N}[0, 1],$$
$$y_{it} = j \quad \text{if} \quad \mu_{j-1} < y_{it}^* < \mu_j, \; j = 0, 1, \ldots, J \quad \text{and} \quad \mu_{-1} = -\infty, \mu_0 = 0, \mu_J = +\infty.$$

The model assumptions imply that

$$\mathrm{Prob}(y_{it} = \mathrm{j} \mid \mathbf{X}_i) = \Lambda(\mu_j - \alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}) - \Lambda(\mu_{j-1} - \alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}),$$

[10] Cross-section versions of the ordered probit model with individual specific thresholds appear in Terza (1985a), Pudney and Shields (2000), and Greene (2007).

where $\Lambda(t)$ is the cdf of the logistic distribution. Now, define a binary variable

$$w_{it,j} = 1 \text{ if } y_{it} > j, \quad j = 0, \ldots, J-1.$$

It follows that

$$\begin{aligned}\text{Prob}[w_{it,j} = 1 \mid \mathbf{X}_i] &= \Lambda(\alpha_i - \mu_j + \mathbf{x}_{it}'\boldsymbol{\beta}) \\ &= \Lambda(\theta_i + x_{it}'\boldsymbol{\beta}).\end{aligned}$$

The "j " specific constant, which is the same for all individuals, is absorbed in θ_i. Thus, a fixed effects binary logit model applies to each of the $J-1$ binary random variables, $w_{it,j}$. The method in Section 17.4.4 can now be applied to each of the $J-1$ random samples. This provides $J-1$ estimators of the parameter vector $\boldsymbol{\beta}$ (but no estimator of the threshold parameters). The authors propose to reconcile these different estimators by using a minimum distance estimator of the common true $\boldsymbol{\beta}$. (See Section 13.3.) The minimum distance estimator at the second step is chosen to minimize

$$q = \sum_{j=0}^{J-1}\sum_{m=0}^{J-1}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})'\left[\mathbf{V}_{jm}^{-1}\right](\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}),$$

where $[\mathbf{V}_{jm}^{-1}]$ is the j, m block of the inverse of the $(J-1)K \times (J-1)K$ partitioned matrix $\mathbf{V}$ that contains Asy. Cov$[\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_m]$. The appropriate form of this matrix for a set of cross-section estimators is given in Brant (1990). Das and van Soest (2000) used the counterpart for Chamberlain's fixed effects estimator but do not provide the specifics for computing the off-diagonal blocks in $\mathbf{V}$.

The full ordered probit model with fixed effects, including the individual specific constants, can be estimated by unconditional maximum likelihood using the results in Section 14.9.6.d. The likelihood function is concave [see Pratt (1981)], so despite its superficial complexity, the estimation is straightforward. (In the following application, with more than 27,000 observations and 7,293 individual effects, estimation of the full model required roughly five seconds of computation.) No theoretical counterpart to the Hsiao (1986, 2003) and Abrevaya (1997) results on the small T bias (incidental parameters problem) of the MLE in the presence of fixed effects has been derived for the ordered probit model. The Monte Carlo results in Greene (2004) (see, as well, Chapter 15), suggest that biases comparable to those in the binary choice models persist in the ordered probit model as well. As in the binary choice case, the complication of the fixed effects model is the small sample bias, not the computation. The Das and van Soest approach finesses this problem—their estimator is consistent—but at the cost of losing the information needed to compute partial effects or predicted probabilities.

18.3.4.b Ordered Probit Models with Random Effects

The random effects ordered probit model model has been much more widely used than the fixed effects model. Applications include Groot and van den Brink (2003), who studied training levels of employees, with firm effects; Winkelmann (2003b), who examined subjective measures of well-being with individual and family effects; Contoyannis et al. (2004), who analyzed self-reported measures of health status; and numerous others. In the simplest case, the method of the Butler and Moffitt (1982) quadrature method (Section 14.9.6.b) can be extended to this model.

Example 18.6 Health Satisfaction

The GSOEP German Health Care data that we have used in Example 17.4, and others includes a self-reported measure of health satisfaction, *HSAT*, that takes values 0, 1, . . . , 10.[11] This is a typical application of a scale variable that reflects an underlying continuous variable, "health." The frequencies and sample proportions for the reported values are as follows:

HSAT	*Frequency*	*Proportion*
0	447	1.6%
1	255	0.9%
2	642	2.3%
3	1173	4.2%
4	1390	5.0%
5	4233	15.4%
6	2530	9.2%
7	4231	15.4%
8	6172	22.5%
9	3061	11.2%
10	3192	11.6%

We have fit pooled and panel data versions of the ordered probit model to these data. The model used is

$$y_{it}^* = \beta_1 + \beta_2\, Age_{it} + \beta_3\, Income_{it} + \beta_4\, Kids_{it}\beta_6\, Education_{it} + \beta_6\, Married_{it} + \beta_7\, Working_{it} + \varepsilon_{it} + \mathrm{c_i},$$

where c_i will be the common fixed or random effect. (We are interested in comparing the fixed and random effects estimators, so we have not included any time-invariant variables such as gender in the equation.) Table 18.12 lists five estimated models. (Standard errors for the estimated threshold parameters are omitted.) The first is the pooled ordered probit model. The second and third are fixed effects. Column 2 shows the unconditional fixed effects estimates using the results of Section 14.9.6.d. Column 3 shows the Das and van Soest estimator. For the minimum distance estimator, we used an inefficient weighting matrix, the block-diagonal matrix in which the jth block is the inverse of the jth asymptotic covariance matrix for the individual logit estimators. With this weighting matrix, the estimator is

$$\hat{\boldsymbol{\beta}}_{MDE} = \left[\sum_{j=0}^{9} \mathbf{V}_j^{-1}\right]^{-1} \sum_{j=0}^{9} \mathbf{V}_j^{-1}\hat{\boldsymbol{\beta}}_j,$$

and the estimator of the asymptotic covariance matrix is approximately equal to the bracketed inverse matrix. The fourth set of results is the random effects estimator computed using the maximum simulated likelihood method. This model can be estimated using Butler and Moffitt's quadrature method; however, we found that even with a large number of nodes, the quadrature estimator converged to a point where the log-likelihood was far lower than the MSL estimator, and at parameter values that were implausibly different from the other estimates. Using different starting values and different numbers of quadrature points did not change this outcome. The MSL estimator for a random constant term (see Section 15.6) is considerably slower but produces more reasonable results. The fifth set of results is the Mundlak form of the random effects model, which includes the group means in the models as controls to accommodate possible correlation between the latent heterogeneity and the included variables. As noted in Example 18.2, the components of the ordered choice model must be interpreted with some care. By construction, the partial effects of the variables on

[11] In the original data set, 40 (of 27,326) observations on this variable were coded with noninteger values between 6 and 7. For purposes of our example, we have recoded all 40 observations to 7.

TABLE 18.12 Estimated Ordered Probit Models for Health Satisfaction

	(1)	(2)	(3)	(4)	(5) Random Effects Mundlak Controls	
Variable	*Pooled*	*Fixed Effects Unconditional*	*Fixed Effects Conditional*	*Random Effects*	*Variables*	*Means*
Constant	2.4739			3.8577	3.2603	
	(0.04669)			(0.05072)	(0.05323)	
Age	−0.01913	−0.07162	−0.1011	−0.03319	−0.06282	0.03940
	(0.00064)	(0.002743)	(0.002878)	(0.00065)	(0.00234)	(0.002442)
Income	0.1811	0.2992	0.4353	0.09436	0.2618	0.1461
	(0.03774)	(0.07058)	(0.07462)	(0.03632)	(0.06156)	(0.07695)
Kids	0.06081	−0.06385	−0.1170	0.01410	−0.05458	0.1854
	(0.01459)	(0.02837)	(0.03041)	(0.01421)	(0.02566)	(0.03129)
Education	0.03421	0.02590	0.06013	0.04728	0.02296	0.02257
	(0.002828)	(0.02677)	(0.02819)	(0.002863)	(0.02793)	(0.02807)
Married	0.02574	0.05157	0.08505	0.07327	0.04605	−0.04829
	(0.01623)	(0.04030)	(0.04181)	(0.01575)	(0.03506)	(0.03963)
Working	0.1292	−0.02659	−0.007969	0.07108	−0.02383	0.2702
	(0.01403)	(0.02758)	(0.02830)	(0.01338)	(0.02311)	(0.02856)
μ_1	0.1949	0.3249		0.2726	0.2752	
μ_2	0.5029	0.8449		0.7060	0.7119	
μ_3	0.8411	1.3940		1.1778	1.1867	
μ_4	1.111	1.8230		1.5512	1.5623	
μ_5	1.6700	2.6992		2.3244	2.3379	
μ_6	1.9350	3.1272		2.6957	2.7097	
μ_7	2.3468	3.7923		3.2757	3.2911	
μ_8	3.0023	4.8436		4.1967	4.2168	
μ_9	3.4615	5.5727		4.8308	4.8569	
σ_u	0.0000	0.0000		1.0078	0.9936	
ln L	−56813.52	−41875.63		−53215.54	−53070.43	

TABLE 18.13 Estimated Marginal Effects: Pooled Model

HSAT	*Age*	*Income*	*Kids*	*Education*	*Married*	*Working*
0	0.0006	−0.0061	−0.0020	−0.0012	−0.0009	−0.0046
1	0.0003	−0.0031	−0.0010	−0.0006	−0.0004	−0.0023
2	0.0008	−0.0072	−0.0024	−0.0014	−0.0010	−0.0053
3	0.0012	−0.0113	−0.0038	−0.0021	−0.0016	−0.0083
4	0.0012	−0.0111	−0.0037	−0.0021	−0.0016	−0.0080
5	0.0024	−0.0231	−0.0078	−0.0044	−0.0033	−0.0163
6	0.0008	−0.0073	−0.0025	−0.0014	−0.0010	−0.0050
7	0.0003	−0.0024	−0.0009	−0.0005	−0.0003	−0.0012
8	−0.0019	0.0184	0.0061	0.0035	0.0026	0.0136
9	−0.0021	0.0198	0.0066	0.0037	0.0028	0.0141
10	−0.0035	0.0336	0.0114	0.0063	0.0047	0.0233

the probabilities of the outcomes must change sign, so the simple coefficients do not show the complete picture implied by the estimated model. Table 18.13 shows the partial effects for the pooled model to illustrate the computations.

Winkelmann (2003b) used the random effects approach to analyze the **subjective well-being** (SWB) question (also coded 0 to 10) in the German Socioeconomic

Panel (GSOEP) data set. The ordered probit model in this study is based on the latent regression

$$y^*_{imt} = \mathbf{x}'_{imt}\boldsymbol{\beta} + \varepsilon_{imt} + u_{im} + v_i.$$

The independent variables include age, gender, employment status, income, family size, and an indicator for good health. An unusual feature of the model is the nested random effects (see Section 14.9.6.b), which include a family effect, v_i, as well as the individual family member (i in family m) effect, u_{im}. The GLS/MLE approach we applied to the linear regression model in Section 14.9.6.b is unavailable in this nonlinear setting. Winkelmann instead employed a Hermite quadrature procedure to maximize the log-likelihood function.

Contoyannis, Jones, and Rice (2004) analyzed a self-assessed health scale that ranged from 1 (very poor) to 5 (excellent) in the British Household Panel Survey. Their model accommodated a variety of complications in survey data. The latent regression underlying their ordered probit model is

$$h^*_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{H}'_{i,t-1}\boldsymbol{\gamma} + \alpha_i + \varepsilon_{it},$$

where $\mathbf{x}_{it}$ includes marital status, race, education, household size, age, income, and number of children in the household. The lagged value, $\mathbf{H}_{i,t-1}$, is a set of binary variables for the observed health status in the previous period. (This is the same device that was used by Butler et al. in Example 18.3.) In this case, the lagged values capture state dependence—the assumption that the health outcome is redrawn randomly in each period is inconsistent with evident runs in the data. The initial formulation of the regression is a fixed effects model. To control for the possible correlation between the effects, α_i, and the regressors, and the initial conditions problem that helps to explain the state dependence, they use a hybrid of Mundlak's (1978) correction and a suggestion by Wooldridge (2002a) for modeling the initial conditions,

$$\alpha_i = \alpha_0 + \bar{\mathbf{x}}'\boldsymbol{\alpha}_1 + \mathbf{H}'_{i,1}\boldsymbol{\delta} + u_i,$$

where u_i is exogenous. Inserting the second equation into the first produces a random effects model that can be fit using the quadrature method we considered earlier.

18.3.5 EXTENSIONS OF THE ORDERED PROBIT MODEL

The basic specification of the ordered probit model can be extended in the same directions as we considered in constructing models for binary choice in Chapter 17. These include heteroscedasticity in the random utility function [see Section 17.3.7.b, Keele and Park (2005), and Wang and Kockelman (2005), for an application] and heterogeneity in the preferences (i.e., random parameters and latent classes). [An extensive study of heterogeneity in health satisfaction based on 22 waves of the GSOEP is Jones and Schurer (2010).] Two specification issues that are specific to the ordered choice model are accommodating heterogeneity in the threshold parameters and reconciling differences in the meaning of the preference scale across different groups. We will sketch the model extensions in this section. Further details are given in Chapters 6 and 7 of Hensher and Greene (2010).

18.3.5.a Threshold Models—Generalized Ordered Choice Models

The model analyzed thus far assumes that the thresholds μ_j are the same for every individual in the sample. Terza (1985a), Pudney and Shields (2000), King, Murray, Salomon and Tandon (KMST, 2004), Boes and Winkelmann (2006a), Greene, Harris, Hollingsworth and Maitra (2008), and Greene and Hensher (2009) all present applications that include individual variation in the thresholds of the ordered choice model.

In his analysis of bond ratings, Terza (1985) suggested the generalization,

$$\mu_{ij} = \mu_j + \mathbf{x}_i'\boldsymbol{\delta}.$$

With three outcomes, the probabilities are

$$y_i^* = \alpha + \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,$$

and

$$\begin{aligned} y_i = \ & 0 \text{ if } y_i^* \le 0, \\ & 1 \text{ if } 0 < y_i^* \le \mu + \mathbf{x}_i'\boldsymbol{\delta}, \\ & 2 \text{ if } y_i^* > \mu + \mathbf{x}_i'\boldsymbol{\delta}. \end{aligned}$$

For three outcomes, the model has two thresholds, $\mu_0 = 0$ and $\mu_1 = \mu + \mathbf{x}_i'\delta$. The three probabilities can be written

$$\begin{aligned} P_0 &= \text{Prob}(y_i = 0 \,|\, \mathbf{x}_i) = \Phi[-(\alpha + \mathbf{x}_i'\boldsymbol{\beta})] \\ P_1 &= \text{Prob}(y_i = 1 \,|\, \mathbf{x}_i) = \Phi[(\mu + \mathbf{x}_i'\boldsymbol{\delta}) - (\alpha + \mathbf{x}_i'\boldsymbol{\beta})] - \Phi[-(\alpha + \mathbf{x}_i'\boldsymbol{\beta})] \\ P_2 &= \text{Prob}(y_i = 2 \,|\, \mathbf{x}_i) = 1 - \Phi[(\mu + \mathbf{x}_i'\boldsymbol{\delta}) - (\alpha + \mathbf{x}_i'\boldsymbol{\beta})]. \end{aligned}$$

For applications of this approach, see, for example, Kerkhofs and Lindeboom (1995), Groot and van den Brink (2003) and Lindeboom and van Doorslayer (2003). Note that if $\boldsymbol{\delta}$ is unrestricted, then $\text{Prob}(y_i = 1 \,|\, \mathbf{x}_i)$ can be negative. This is a shortcoming of the model when specified in this form. Subsequent development of the generalized model involves specifications that avoid this internal inconsistency. Note, as well, that if the model is recast in terms of μ and $\boldsymbol{\gamma} = [\alpha, (\boldsymbol{\beta} - \boldsymbol{\delta})]$, then the model is not distinguished from the original ordered probit model with a constant threshold parameter. This identification issue emerges prominently in Pudney and Shield's (2000) continued development of this model.

Pudney and Shields's (2000) "generalized ordered probit model," was also formulated to accommodate *observable* individual heterogeneity in the threshold parameters. Their application was in the context of job promotion for UK nurses in which the steps on the promotion ladder are individual specific. In their setting, in contrast to Terza's, some of the variables in the threshold equations are explicitly different from those in the regression. The authors constructed a generalized model and a test of "threshold constancy" by defining $\mathbf{q}_i$ to include a constant term and those variables that are unique to the threshold model Variables that are common to both the thresholds and the regression are placed in $\mathbf{x}_i$ and the model is reparameterized as

$$\Pr(y_i = g \,|\, \mathbf{x}_i, \mathbf{q}_i) = \Phi[\mathbf{q}_i'\boldsymbol{\delta}_g - \mathbf{x}_i'(\boldsymbol{\beta} - \boldsymbol{\delta}_g)] - \boldsymbol{\Phi}[\mathbf{q}_i'\boldsymbol{\delta}_{g-1} - \mathbf{x}_i'(\boldsymbol{\beta} - \boldsymbol{\delta}_{g-1})].$$

An important point noted by the authors is that the same model results if these common variables are placed in the thresholds instead. This is a minor algebraic result, but it exposes an ambiguity in the interpretation of the model—whether a particular variable affects the regression or the thresholds is one of the issues that was developed in the original model specification.

As will be evident in the application in the next section, the specification of the threshold parameters is a crucial feature of the ordered choice model. KMST (2004), Greene (2007a), Eluru, Bhat, and Hensher (2008), and Greene and Hensher (2009) employ a "hierarchical ordered probit," or HOPIT model,

$$\begin{aligned} y_i^* &= \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \\ y_i &= j \text{ if } \mu_{i,j-1} \leq y_i^* < \mu_{ij}, \\ \mu_0 &= 0, \\ \mu_{i,j} &= \exp(\lambda_j + \boldsymbol{\gamma}'\mathbf{z}_i) \quad \text{(case 1)}, \\ \text{or } \mu_{i,j} &= \exp(\lambda_j + \boldsymbol{\gamma}_j'\mathbf{z}_i) \quad \text{(case 2)}. \end{aligned}$$

Case 2 is the Terza (1985) and Pudney and Shields (2000) model with an exponential rather than linear function for the thresholds. This formulation addresses two problems: (1) The thresholds are mathematically distinct from the regression; (2) by this construction, the threshold parameters must be positive. With a slight modification, the ordering of the thresholds can also be imposed. In case 1,

$$\mu_{i,j} = [\exp(\lambda_1) + \exp(\lambda_2) + \cdots + \exp(\lambda_j)] \times \exp(\boldsymbol{\gamma}'\mathbf{z}_i),$$

and in case 2,

$$\mu_{i,j} = \mu_{i,j-1} + \exp(\lambda_j + \boldsymbol{\gamma}_j'\mathbf{z}_i).$$

In practical terms, the model can now be fit with the constraint that all predicted probabilities are greater than zero. This is a numerical solution to the problem of ordering the thresholds for all data vectors.

This extension of the ordered choice model shows a case of **identification through functional form**. As we saw in the previous two models, the parameters $(\lambda_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta})$ would not be separately identified if all the functions were linear. The contemporary literature views models that are unidentified without a change in functional form with some skepticism. However, the underlying theory of this model does not insist on linearity of the thresholds (or the utility function, for that matter), but it *does* insist on the ordering of the thresholds, and one might equally criticize the original model for being unidentified *because the model builder insists on a linear form*. That is, there is no obvious reason that the threshold parameters must be linear functions of the variables, or that linearity enjoys some claim to first precedence in the utility function. This is a methodological issue that cannot be resolved here. The nonlinearity of the preceding specification, or others that resemble it, does provide the benefit of a simple way to achieve other fundamental results, for example, coherency of the model (all positive probabilities).

18.3.5.b Thresholds and Heterogeneity—Anchoring Vignettes

The introduction of observed heterogeneity into the threshold parameters attempts to deal with a fundamentally restrictive assumption of the ordered choice model. Survey respondents rarely view the survey questions exactly the same way. This is certainly true

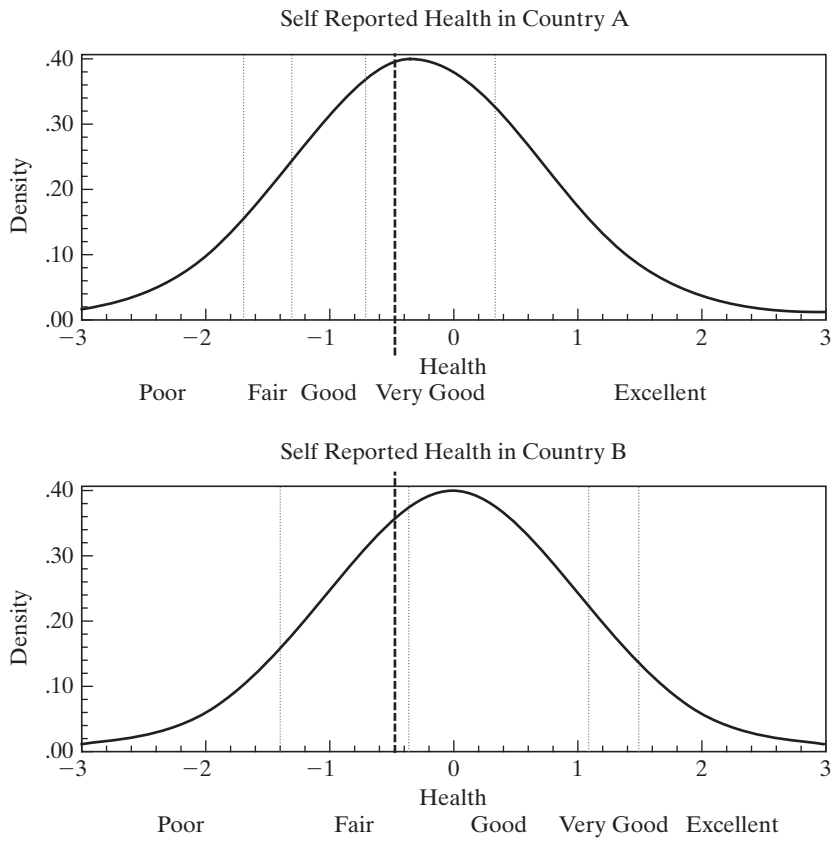


FIGURE 18.5 Differential item Functioning in Ordered Choices.

in surveys of health satisfaction or subjective well-being. [See Boes and Winkelmann (2006b) and Ferrer-i-Carbonell and Frijters (2004).] KMST (2004) identify two very basic features of survey data that will make this problematic. First, they often measure concepts that are definable only with reference to examples, such as freedom, health, satisfaction, and so on. Second, individuals do, in fact, often understand survey questions very differently, particularly with respect to answers at the extremes. A widely used term for this interpersonal incomparability is **differential item functioning (DIF)**. Kapteyn, Smith, and Van Soest (KSV, 2007) and Van Soest, Delaney, Harmon, Kapteyn and Smith (2007) suggest the results in Figure 18.5 to describe the implications of DIF. The figure shows the distribution of Health (or drinking behavior in the latter study) in two hypothetical countries. The density for country A (the upper figure) is to the left of that for country B, implying that, on average, people in country A are less healthy than those in country B. But, the people in the two countries culturally offer very different response scales if asked to report their health on a five-point scale, as shown. In the figure, those in country A have a much more positive view of a given, objective health status than those in country B. A person in country A with health status indicated by the dotted line would report that they are in "Very Good" health while a person in

country B with the same health status would report only "Fair." A simple frequency of the distribution of self-assessments of health status in the two countries would suggest that people in country A are much healthier than those in country B when, in fact, the opposite is true. Correcting for the influences of DIF in such a situation would be essential to obtaining a meaningful comparison of the two countries. The impact of DIF is an accepted feature of the model within a population but could be strongly distortionary when comparing very disparate groups, such as across countries, as in KMST (political groups), Murray, Tandon, Mathers, and Sudana (2002) (health outcomes), Tandon et al. (2004), and KSV (work disability), Sirven, Santos-Egglmann, and Spagnoli (2008), and Gupta, Kristensens, and Possoli (2008) (health), Angelini et al. (2008) (life satisfaction), Kristensen and Johansson (2008), and Bago d'Uva et al. (2008), all of whom used the ordered probit model to make cross group comparisons.

KMST proposed the use of *anchoring vignettes* to resolve this difference in perceptions across groups. The essential approach is to use a series of examples that, it is believed, all respondents will agree on to estimate each respondent's DIF and correct for it. The idea of using vignettes to anchor perceptions in survey questions is not itself new; KMST cite a number of earlier uses. The innovation is their method for incorporating the approach in a formal model for the ordered choices. The bivariate and multivariate probit models that they develp combine the elements described in Sections 18.3.1–18.3.3 and the HOPIT model in Section 18.3.4.a.

18.4 MODELS FOR COUNTS OF EVENTS

We have encountered behavioral variables that involve counts of events at several points in this text. In Examples 14.10 and 17.20, we examined the number of times an individual visited the physician using the GSOEP data. The credit default data that we used in Examples 7.10 and 17.22 also include another behavioral variable, the number of derogatory reports in an individual's credit history. Finally, in Example 17.23, we analyzed data on firm innovation. Innovation is often analyzed [for example, by Hausman, Hall, and Griliches (1984) and many others] in terms of the number of patents that the firm obtains (or applies for). In each of these cases, the variable of interest is a count of events. This obviously differs from the discrete dependent variables we analyzed in the previous two sections. A count is a quantitative measure that is, at least in principle, amenable to analysis using multiple linear regression. However, the typical preponderance of zeros and small values and the discrete nature of the outcome variable suggest that the regression approach can be improved by a method that explicitly accounts for these aspects.

Like the basic multinomial logit model for unordered data in Section 18.2 and the simple probit and logit models for binary and ordered data in Sections 17.2 and 18.3, the Poisson regression model is the fundamental starting point for the analysis of count data. We will develop the elements of modeling for count data in this framework in Sections 18.4.1–18.4.3, and then turn to more elaborate, flexible specifications in subsequent sections. Sections 18.4.4 and 18.4.5 will present the negative binomial and other alternatives to the Poisson functional form. Section 18.4.6 will describe the implications for the model specification of some complicating features of observed data, truncation, and censoring. Truncation arises when certain values, such as zero, are absent from the

observed data because of the sampling mechanism, not as a function of the data generating process. Data on recreation site visitation that are gathered at the site, for example, will, by construction, not contain any zeros. Censoring arises when certain ranges of outcomes are all coded with the same value. In the example analyzed the response variable is censored at 12, though values larger than 12 are possible "in the field." As we have done in the several earlier treatments, in Section 18.4.7, we will examine extensions of the count data models that are made possible when the analysis is based on panel data. Finally, Section 18.4.8 discusses some behavioral models that involve more than one equation. For an example, based on the large number of zeros in the observed data, it appears that our count of doctor visits might be generated by a two-part process, a first step in which the individual decides whether or not to visit the physician at all, and a second decision, given the first, how many times to do so. The "hurdle model" that applies here and some related variants are discussed in Section 18.4.8.

18.4.1 THE POISSON REGRESSION MODEL

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

$$\text{Prob}(Y = y_i \mid \mathbf{x}_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots. \tag{18-17}$$

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. But it is far easier to estimate the parameters with maximum likelihood techniques. The log-likelihood 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 $\hat{\lambda}_i = \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})$. 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

$$\text{LM} = \left[\sum_{i=1}^{n} \mathbf{x}_i(y_i - \hat{\lambda}_i)\right]' \left[\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'(y_i - \hat{\lambda}_i)^2\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{x}_i(y_i - \hat{\lambda}_i)\right] = \mathbf{i}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{i}, \tag{18-18}$$

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.

18.4.2 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.[12] 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 rise 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 because y_i is an integer

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

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.[13] 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),

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

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

18.4.3 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.

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

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

[13] 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.

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.

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

$$\text{LM} = \left[\frac{\sum_{i=1}^n \hat{w}_i[(y_i - \hat{\lambda}_i)^2 - y_i]}{\sqrt{2\sum_{i=1}^n \hat{w}_i \hat{\lambda}_i^2}}\right]^2. \tag{18-19}$$

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:

$$\text{LM} = \frac{(\mathbf{e}'\mathbf{e} - n\bar{y})^2}{2\hat{\boldsymbol{\lambda}}'\hat{\boldsymbol{\lambda}}}. \tag{18-20}$$

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.

18.4.4 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 (2003) for discussion]. The most common is the negative binomial model, which arises from a natural formulation of cross-section heterogeneity. [See Hilbe (2007).] 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 $u_i = \exp(\varepsilon_i)$.[14] 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 2 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 $\alpha = 1/\theta = 0$ using the Wald or likelihood ratio test.

18.4.5 FUNCTIONAL FORMS FOR COUNT DATA MODELS

The equidispersion assumption of the Poisson regression model, $E[y_i \mid \mathbf{x}_i] = \text{Var}[y_i \mid \mathbf{x}_i]$, is a major shortcoming. Observed data rarely, if ever, display this feature. The very large amount of research activity on functional forms for count models is often focused on testing for equidispersion and building functional forms that relax this assumption. In practice, the Poisson model is typically only the departure point for an extended specification search.

One easily remedied minor issue concerns the units of measurement of the data. In the Poisson and negative binomial models, the parameter λ_i is the expected number of events *per unit of time*. Thus, there is a presumption in the model formulation, for

[14] An alternative approach based on the normal distribution is suggested in Terza (1998), Greene (1995a, 1997a, 2007d), Winkelmann (1997) and Riphahn, Wambach and Million (2003). 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 or simulation. See Sections 14.9.6.b and 17.4.8.

example, the Poisson, that the same amount of time is observed for each i. In a spatial context, such as measurements of the incidence of a disease per group of N_i persons, or the number of bomb craters per square mile (London, 1940), the assumption would be that the same physical area or the same size of population applies to each observation. Where this differs by individual, it will introduce a type of heteroscedasticity in the model. The simple remedy is to modify the model to account for the **exposure**, T_i, of the observation as follows:

$$\text{Prob}(y_i = j \mid \mathbf{x}_i, T_i) = \frac{\exp(-T_i\phi_i)(T_i\phi_i)^j}{j!}, \quad \phi_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}),\ j = 0, 1, \ldots.$$

The original model is returned if we write $\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \ln T_i)$. Thus, when the exposure differs by observation, the appropriate accommodation is to include the log of exposure in the regression part of the model with a coefficient of 1.0. (For less than obvious reasons, the term "offset variable" is commonly associated with the exposure variable T_i.) Note that if T_i is the same for all i, $\ln T_i$ will simply vanish into the constant term of the model (assuming one is included in $\mathbf{x}_i$).

The recent literature, mostly associating the result with Cameron and Trivedi's (1986, 1998) work, defines two familiar forms of the negative binomial model. The **Negbin 2** (NB2) **form** of the probability is

$$\begin{aligned} \text{Prob}(Y = y_i \mid \mathbf{x}_i) &= \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} r_i^{y_i}(1 - r_i)^{\theta}, \\ \lambda_i &= \exp(\mathbf{x}_i'\boldsymbol{\beta}), \\ r_i &= \lambda_i/(\theta + \lambda_i). \end{aligned} \tag{18-21}$$

This is the default form of the model in the received econometrics packages that provide an estimator for this model. The **Negbin 1** (NB1) **form** of the model results if θ in the preceding is replaced with $\theta_i = \theta\lambda_i$. Then, r_i reduces to $r = 1/(1 + \theta)$, and the density becomes

$$\text{Prob}(Y = y_i \mid \mathbf{x}_i) = \frac{\Gamma(\theta\lambda_i + y_i)}{\Gamma(y_i + 1)\Gamma(\theta\lambda_i)} r^{y_i}(1 - r)^{\theta\lambda_i}. \tag{18-22}$$

This is not a simple reparameterization of the model. The results in Example 18.7 demonstrate that the log-likelihood functions are not equal at the maxima, and the parameters are not simple transformations in one model versus the other. We are not aware of a theory that justifies using one form or the other for the negative binomial model. Neither is a restricted version of the other, so we cannot carry out a likelihood ratio test of one versus the other. The more general **Negbin P (NBP)** family does nest both of them, so this may provide a more general, encompassing approach to finding the right specification. [See Greene (2005, 2008).] The Negbin P model is obtained by replacing θ in the Negbin 2 form with $\theta\lambda_i^{2-P}$. We have examined the cases of $P = 1$ and $P = 2$ in (19-5) and (19-6). The full model is

$$\text{Prob}(Y = y_i \mid \mathbf{x}_i) = \frac{\Gamma(\theta\lambda_i^Q + y_i)}{\Gamma(y_i + 1)\Gamma\left(\theta\lambda_i^Q\right)} \left(\frac{\lambda_i}{\theta\lambda_i^Q + \lambda_i}\right)^{y_i} \left(\frac{\theta\lambda_i^Q}{\theta\lambda_i^Q + \lambda_i}\right)^{\theta\lambda_i^Q}, \ Q = 2 - P.$$

The conditional mean function for the three cases considered is

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

The parameter P is picking up the scaling. A general result is that for all three variants of the model,

$$\text{Var}[y_i \mid \mathbf{x}_i] = \lambda_i\left(1 + \alpha\lambda_i^{P-1}\right), \quad \text{where } \alpha = 1/\theta.$$

Thus, the NB2 form has a variance function that is quadratic in the mean while the NB1 form's variance is a simple multiple of the mean. There have been many other functional forms proposed for count data models, including the generalized Poisson, gamma, and Polya-Aeppli forms described in Winkelmann (2003) and Greene (2007a, Chapter 24).

The heteroscedasticity in the count models is induced by the relationship between the variance and the mean. The single parameter θ picks up an implicit overall scaling, so it does not contribute to this aspect of the model. As in the linear model, microeconomic data are likely to induce heterogeneity in both the mean and variance of the response variable. A specification that allows independent variation of both will be of some virtue. The result

$$\text{Var}[y_i \mid \mathbf{x}_i] = \lambda_i\left(1 + (1/\theta)\lambda_i^{P-1}\right)$$

suggests that a natural platform for separately modeling heteroscedasticity will be the dispersion parameter, θ, which we now parameterize as

$$\theta_i = \theta \exp(\mathbf{z}_i'\boldsymbol{\delta}).$$

Operationally, this is a relatively minor extension of the model. But, it is likely to introduce quite a substantial increase in the flexibility of the specification. Indeed, a heterogeneous Negbin P model is likely to be sufficiently parameterized to accommodate the behavior of most data sets. (Of course, the specialized models discussed in Section 18.4.8, for example, the zero inflation models, may yet be more appropriate for a given situation.)

Example 18.7 Count Data Models for Doctor Visits

The study by Riphahn et al. (2003) that provided the data we have used in numerous earlier examples analyzed the two count variables DocVis (visits to the doctor) and HospVis (visits to the hospital). The authors were interested in the joint determination of these two count variables. One of the issues considered in the study was whether the data contained evidence of moral hazard, that is, whether health care utilization as measured by these two outcomes was influenced by the subscription to health insurance. The data contain indicators of two levels of insurance coverage, PUBLIC, which is the main source of insurance, and ADDON, which is a secondary optional insurance. In the sample of 27,326 observations (family/years), 24,203 individuals held the public insurance. (There is quite a lot of within group variation in this. Individuals did not routinely obtain the insurance for all periods.) Of these 24,203, 23,689 had only public insurance and 514 had both types. (One could not have only the ADDON insurance.) To explore the issue, we have analyzed the DocVis variable with the count data models described in this section. The exogenous variables in our model are

$$\mathbf{x}_{it} = (1, \textit{Age}, \textit{Education}, \textit{Income}, \textit{Kids}, \textit{Public}).$$

(Variables are described in Appendix Table F7.1.)

Table 18.14 presents the estimates of the several count models. In all specifications, the coefficient on PUBLIC is positive, large, and highly statistically significant, which is consistent with the results in the authors' study. The various test statistics strongly reject the hypothesis

TABLE 18.14 Estimated Models for DOCVIS (standard errors in parentheses)

Variable	*Poisson*	*Negbin 2*	*Negbin 2 Heterogeneous*	*Negbin 1*	*Negbin P*
Constant	0.7162	0.7628	0.7928	0.6848	0.6517
	(0.03287)	(0.07247)	(0.07459)	(0.06807)	(0.07759)
Age	0.01844	0.01803	0.01704	0.01585	0.01907
	(0.0003316)	(0.0007915)	(0.0008146)	(0.0007042)	(0.0008078)
Education	−0.03429	−0.03839	−0.03581	−0.02381	−0.03388
	(0.001797)	(0.003965)	(0.004036)	(0.003702)	(0.004308)
Income	−0.4751	−0.4206	−0.4108	−0.1892	−0.3337
	(0.02198)	(0.04700)	(0.04752)	(0.04452)	(0.05161)
Kids	−0.1582	−0.1513	−0.1568	−0.1342	−0.1622
	(0.007956)	(0.01738)	(0.01773)	(0.01647)	(0.01856)
Public	0.2364	0.2324	0.2411	0.1616	0.2195
	(0.01328)	(0.02900)	(0.03006)	(0.02678)	(0.03155)
P	0.0000	2.0000	2.0000	1.0000	1.5473
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.03444)
θ	0.0000	1.9242	2.6060	6.1865	3.2470
	(0.0000)	(0.02008)	(0.05954)	(0.06861)	(0.1346)
δ (Female)	0.0000	0.0000	−0.3838	0.0000	0.0000
	(0.0000)	(0.0000)	(0.02046)	(0.0000)	(0.0000)
δ (Married)	0.0000	0.0000	−0.1359	0.0000	0.0000
	(0.0000)	(0.0000)	(0.02307)	(0.0000)	(0.0000)
ln L	−104440.3	−60265.49	−60121.77	−60260.68	−60197.15

of equidispersion. Cameron and Trivedi's (1990) semiparametric tests from the Poisson model (see Section 18.4.3 have t statistics of 22.147 for $g_i = \mu_i$ and 22.504 for $g_i = \mu_i^2$. Both of these are far larger than the critical value of 1.96. The LM statistic is 972,714.48, which is also larger than the (any) critical value. On these bases, we would reject the hypothesis of equidispersion. The Wald and likelihood ratio tests based on the negative binomial models produce the same conclusion. For comparing the different negative binomial models, note that Negbin 2 is the worst of the three by the likelihood function, although NB1 and NB2 are not directly comparable. On the other hand, note that in the NBP model, the estimate of P is more than 10 standard errors from 1.0000 or 2.000, so both NB1 and NB2 are rejected in favor of the unrestricted NBP form of the model. The NBP and the heterogeneous NB2 model are not nested either, but comparing the log-likelihoods, it does appear that the heterogeneous model is substantially superior. We computed the Vuong statistic based on the individual contributions to the log-likelihoods, with $v_i = \ln L_i(\text{NBP}) - \ln L_i(\text{NB2-H})$. (See Section 14.6.6). The value of the statistic is −3.27. On this basis, we would reject NBP in favor of NB2-H. Finally, with regard to the original question, the coefficient on PUBLIC is larger than 10 times the estimated standard error in every specification. We would conclude that the results are consistent with the proposition that there is evidence of moral hazard.

18.4.6 TRUNCATION AND CENSORING IN MODELS FOR COUNTS

Truncation and censoring are relatively common in applications of models for counts. Truncation arises as a consequence of discarding what appear to be unusable data, such as the zero values in survey data on the number of uses of recreation facilities [Shaw (1988), Bockstael et al. (1990)]. In this setting, a more common case which also gives rise to truncation is on-site sampling. When one is interested in visitation by the entire population, which will naturally include zero visits, but one draws their sample

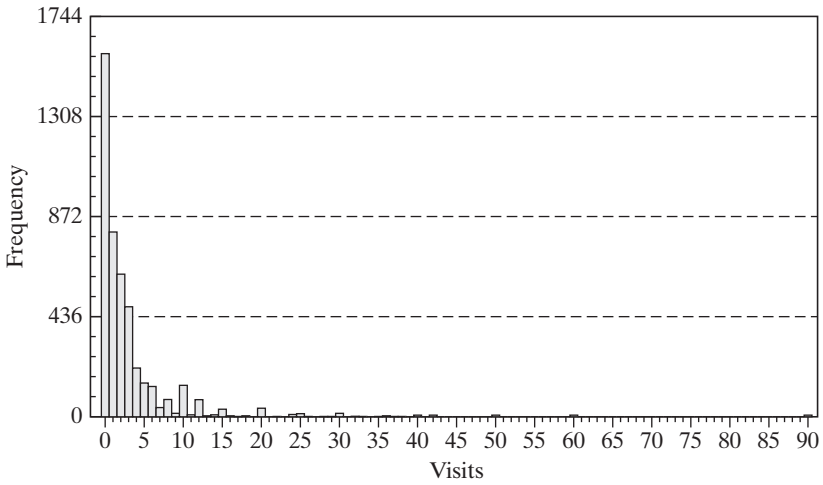


FIGURE 18.6 Number of Doctor Visits. 1988 Wave of GSOEP Data.

"on-site," the distribution of visits is truncated at zero by construction. Every visitor has visited at least once. Shaw (1988), Englin and Shonkwiler (1995), Grogger and Carson (1991), Creel and Loomis (1990), Egan and Herriges (2006) and Martinez-Espinera and Amoako-Tuffour (2008) are among a number of studies that have treated truncation due to on-site sampling in environmental and recreation applications. Truncation will also arise when data are trimmed to remove what appear to be unusual values. Figure 18.6 displays a histogram for the number of doctor visits in the 1988 wave of the GSOEP data that we have used in several examples. There is a suspiciously large spike at zero and an extremely long right tail of what might seem to be atypical observations. For modeling purposes, it might be tempting to remove these "non-Poisson" appearing observations in these tails. (Other models might be a better solution.) The distribution that characterizes what remains in the sample is a truncated distribution. Truncation is not innocent. If the entire population is of interest, then conventional statistical inference (such as estimation) on the truncated sample produces a systematic bias known as (of course) "truncation bias." This would arise, for example, if an ordinary Poisson model intended to characterize the full population is fit to the sample from a truncated population.

Censoring, in contrast, is generally a feature of the sampling design. In the application in Example 18.9, the dependent variable is the self-reported number of extramarital affairs in a survey taken by the magazine *Psychology Today*. The possible answers are 0, 1, 2, 3, 4–10 (coded as 7) and "monthly, weekly or daily" coded as 12. The two upper categories are censored. Similarly, in the doctor visits data in the previous paragraph, recognizing the possibility of truncation bias due to data trimming, we might, instead, simply censor the distribution of values at 15. The resulting variable would take values $0, \ldots, 14$, "15 or more." In both cases, applying conventional estimation methods leads to predictable biases. However, it is also possible to reconstruct the estimators specifically to account for the truncation or censoring in the data.

Truncation and censoring produce similar effects on the distribution of the random variable and on the features of the population such as the mean. For the truncation case, suppose that the original random variable has a Poisson distribution (all these results can be directly extended to the negative binomial or any of the other models considered earlier),

$$P(y_i = j \mid \mathbf{x}_i) = \exp(-\lambda_i)\lambda_i^j / j! = P_{i,j}.$$

If the distribution is truncated at value C—that is, only values $C+1, \ldots$ are observed—then the resulting random variable has probability distribution

$$P(y_i = j \mid \mathbf{x}_i, y_i > C) = \frac{P(y_i = j \mid \mathbf{x}_i)}{P(y_i > C \mid \mathbf{x}_i)} = \frac{P(y_i = j \mid \mathbf{x}_i)}{1 - P(y_i \leq C \mid \mathbf{x}_i)}.$$

The original distribution must be scaled up so that it sums to one for the cells that remain in the truncated distribution. The leading case is truncation at zero, that is, "left truncation," which, for the Poisson model produces

$$P(y_i = j \mid \mathbf{x}_i,\ y_i > 0) = \frac{\exp(-\lambda_i)\lambda_i^j}{j![1 - \exp(-\lambda_i)]} = \frac{P_{i,j}}{1 - P_{i,0}},\ j = 1, \ldots.$$

[See, e.g., Mullahy (1986), Shaw (1988), Grogger and Carson (1991), Greene (1998), and Winkelmann (1987).] The conditional mean function is

$$E(y_i \mid \mathbf{x}_i, y_i > 0) = \frac{1}{[1 - \exp(-\lambda_i)]} \sum_{j=1}^{\infty} \frac{j \exp(-\lambda_i)\lambda_i^j}{j!} = \frac{\lambda_i}{[1 - \exp(-\lambda_i)]} > \lambda_i.$$

The second equality results because the sum can be started at zero—the first term is zero—and this produces the expected value of the original variable. As might be expected, truncation "from below" has the effect of increasing the expected value. It can be shown that it decreases the conditional variance however. The partial effects are

$$\delta_i = \frac{\partial E[y_i \mid \mathbf{x}_i, y_i > 0]}{\partial \mathbf{x}_i} = \left[\frac{1 - P_{i,0} - \lambda_i P_{i,0}}{\left(1 - P_{i,0}\right)^2}\right] \boldsymbol{\lambda}_i \boldsymbol{\beta}. \qquad \textbf{(18-23)}$$

The term outside the brackets is the partial effects in the absence of the truncation while the bracketed term rises from slighter greater than 0.5 to 1.0 as λ_i increases from just above zero.

Example 18.8 Major Derogatory Reports

In Section 17.5.6 and Examples 17.9 and 17.22, we examined a binary choice model for the accept/reject decision for a sample of applicants for a major credit card. Among the variables in that model is "Major Derogatory Reports" (MDRs). This is an interesting behavioral variable in its own right that should be appropriately modeled using the count data specifications in this chapter. In the sample of 13,444 individuals, 10,833 had zero MDRs while the values for the remaining 2561 ranged from 1 to 22. This preponderance of zeros exceeds by far what one would anticipate in a Poisson model that was dispersed enough to produce the distribution of remaining individuals. As we will pursue an Example 18.11, a natural approach for these data is to treat the extremely large block of zeros explicitly in an extended model. For present purposes, we will consider the nonzero observations apart from the zeros and examine the effect of accounting for left truncation at zero on the estimated models. Estimation results are shown in Table 18.15. The first column of results compared to the second shows the

TABLE 18.15 Estimated Truncated Poison Regression Model (*t* ratios in parentheses)

	Poisson Full Sample		*Poisson*		*Truncated Poisson*	
Constant	0.8756	(17.10)	0.8698	(16.78)	0.7400	(11.99)
Age	0.0036	(2.38)	0.0035	(2.32)	0.0049	(2.75)
Income	−0.0039	(−4.78)	−0.0036	(−3.83)	−0.0051	(−4.51)
OwnRent	−0.1005	(−3.52)	−0.1020	(−3.56)	−0.1415	(−4.18)
Self Employed	−0.0325	(−0.62)	−0.0345	(−0.66)	−0.0515	(−0.82)
Dependents	0.0445	(4.69)	0.0440	(4.62)	0.0606	(5.48)
MthsCurAdr	0.00004	(0.23)	0.00005	(0.25)	0.00007	(0.30)
ln *L*	−5379.30		−5378.79		−5097.08	
			Average Partial Effects			
Age	0.0017		0.0085		0.0084	
Income	−0.0018		−0.0087		−0.0089	
OwnRent	−0.0465		−0.2477		−0.2460	
Self Employed	−0.0150		−0.0837		−0.0895	
Dependents	0.0206		0.1068		0.1054	
MthsCurAdr	0.00002		0.00012		0.00013	
Cond'l. Mean	0.4628		2.4295		2.4295	
Scale factor	0.4628		2.4295		1.7381	

suspected impact of incorrectly including the zero observations. The coefficients change only slightly, but the partial effects are far smaller when the zeros are included in the estimation. It was not possible to fit the truncated negative binomial with these data.

Censoring is handled similarly. The usual case is "right censoring," in which realized values greater than or equal to C are all given the value C. In this case, we have a two-part distribution [see Terza (1985b)]. The observed random variable, y_i is constructed from an underlying random variable, y_i^* by

$$y_i = \text{Min}(y_i^*, C).$$

Probabilities are constructed using the axioms of probability. This produces

$$\text{Prob}(y_i = j \mid \mathbf{x}_i) = P_{i,j},\ j = 0, 1, \ldots, C-1,$$

$$\text{Prob}(y_i = C \mid \mathbf{x}_i) = \sum_{j=C}^{\infty} P_{i,j} = 1 - \sum_{j=0}^{C-1} P_{i,j}.$$

In this case, the conditional mean function is

$$\begin{aligned} E[y_i \mid \mathbf{x}_i] &= \sum_{j=0}^{C-1} j P_{i,j} + \sum_{j=C}^{\infty} C P_{i,j} \\ &= \sum_{j=0}^{\infty} j P_{i,j} - \sum_{j=C}^{\infty} (j - C) P_{i,j} \\ &= \lambda_i - \sum_{j=C}^{\infty} (j - C) P_{i,j} < \lambda_i. \end{aligned}$$

The infinite sum is computed by using the complement. Thus,

$$E[y_i|\mathbf{x}_i] = \lambda_i - \left[\sum_{j=0}^{\infty}(j-C)P_{i,j} - \sum_{j=0}^{C-1}(j-C)P_{i,j}\right]$$
$$= \lambda_i - (\lambda_i - C) + \sum_{j=0}^{C-1}(j-C)P_{i,j}$$
$$= C - \sum_{j=0}^{C-1}(C-j)P_{i,j}.$$

Example 18.9 Extramarital Affairs

In 1969, the popular magazine *Psychology Today* published a 101-question survey on sex and asked its readers to mail in their answers. The results of the survey were discussed in the July 1970 issue. From the approximately 2,000 replies that were collected in electronic form (of about 20,000 received), Professor Ray Fair (1978) extracted a sample of 601 observations on men and women then currently married for the first time and analyzed their responses to a question about extramarital affairs. Fair's analysis in this frequently cited study suggests several interesting econometric questions. [In addition, his 1977 companion paper in *Econometrica* on estimation of the tobit model contributed to the development of the EM algorithm, which was published by and is usually associated with Dempster, Laird, and Rubin (1977).]

Fair used the tobit model that we discuss in Chapter 19 as a platform The nonexperimental nature of the data (which can be downloaded from the Internet at http://fairmodel.econ.yale.edu/rayfair/work.ss.htm and are given in Appendix Table F18.1). provides a laboratory case that we can use to examine the relationships among the tobit, truncated regression, and probit models. Although the tobit model seems to be a natural choice for the model for these data, given the cluster of zeros, the fact that the behavioral outcome variable is a count that typically takes a small value suggests that the models for counts that we have examined in this chapter might be yet a better choice. Finally, the preponderance of zeros in the data that initially motivated the tobit model suggests that even the standard Poisson model, although an improvement, might still be inadequate. We will pursue that aspect of the data later. In this example, we will focus on just the censoring issue. Other features of the models and data are reconsidered in the exercises.

The study was based on 601 observations on the following variables (full details on data coding are given in the data file and Appendix Table F18.1):

y = number of affairs in the past year, 0, 1, 2, 3, 4–10 coded as 7
"monthly, weekly, or daily," coded as 12. Sample mean = 1.46
Frequencies = (451, 34, 17, 19, 42, 38)
z_1 = sex = 0 for female, 1 for male. Sample mean = 0.476
z_2 = age. Sample mean = 32.5
z_3 = number of years married. Sample mean = 8.18
z_4 = children, 0 = no, 1 = yes. Sample mean = 0.715
z_5 = religiousness, 1 = anti, . . . , 5 = very. Sample mean = 3.12
z_6= education, years, 9 = grade school, 12 = high school, . . . , 20 = Ph.D or other Sample mean = 16.2
z_7 = occupation, "Hollingshead scale," 1–7. Sample mean = 4.19
z_8 = self-rating of marriage, 1 = very unhappy, . . . , 5 = very happy. Sample mean = 3.93

TABLE 18.16 Censored Poisson and Negative Binomial Distributions

	Poisson Regression			*Negative Binomial Regression*		
Variable	*Estimate*	*Standard Error*	*Marginal Effect*	*Estimate*	*Standard Error*	*Marginal Effect*
			Based on Uncensored Poisson Distribution			
Constant	2.53	0.197	—	2.19	0.664	—
z_2	−0.0322	0.00585	−0.0470	−0.0262	0.0192	−0.00393
z_3	0.116	0.00991	0.168	0.0848	0.0350	0.127
z_5	−0.354	0.0309	−0.515	−0.422	0.111	−0.632
z_7	0.0798	0.0194	0.116	0.0604	0.0702	0.0906
z_8	−0.409	0.0274	−0.0596	−0.431	0.111	−0.646
α				7.01	0.786	
ln L	−1427.037			−728.2441		
			Based on Poisson Distribution Right Censored at $y = 4$			
Constant	1.90	0.283	—	4.79	1.16	—
z_2	−0.0328	0.00838	−0.0235	−0.0166	0.0250	−0.00428
z_3	0.105	0.0140	0.0754	0.174	0.0568	0.045
z_5	−0.323	0.0437	−0.232	−0.723	0.198	−0.186
z_7	0.0798	0.0275	0.0521	0.0900	0.116	0.0232
z_8	−0.390	0.0391	−0.279	−0.854	0.216	−0.220
α				9.40	1.35	
ln L	−747.7541			−482.0505		

The tobit model was fit to y using a constant term and all eight variables. A restricted model was fit by excluding z_1, z_4, and z_6, none of which was individually statistically significant in the model. We are able to match exactly Fair's results for both equations. The tobit model should only be viewed as an approximation for these data. The dependent variable is a count, not a continuous measurement. The Poisson regression model, or perhaps one of the many variants of it, should be a preferable modeling framework. Table 18.16 presents estimates of the Poisson and negative binomial regression models. There is ample evidence of overdispersion in these data; the t ratio on the estimated overdispersion parameter is $7.014/0.945 = 7.42$, which is strongly suggestive. The large absolute value of the coefficient is likewise suggestive.

Responses of 7 and 12 do not represent the actual counts. It is unclear what the effect of the first recoding would be, because it might well be the mean of the observations in this group. But the second is clearly a censored observation. To remove both of these effects, we have recoded both the values 7 and 12 as 4 and treated this observation (appropriately) as a censored observation, with 4 denoting "4 or more." As shown in the third and fourth sets of results in Table 18.16, the effect of this treatment of the data is greatly to reduce the measured effects. Although this step does remove a deficiency in the data, it does not remove the overdispersion; at this point, the negative binomial model is still the preferred specification.

18.4.7 PANEL DATA MODELS

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.

18.4.7.a Robust Covariance Matrices for Pooled Estimators

The standard asymptotic covariance matrix estimator for the Poisson model is

$$\text{Est. Asy. Var}[\hat{\boldsymbol{\beta}}] = \left[-\frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\beta}} \partial \hat{\boldsymbol{\beta}}'}\right]^{-1} = \left[\sum_{i=1}^{n} \hat{\lambda}_i \mathbf{x}_i \mathbf{x}_i'\right]^{-1} = [\mathbf{X}'\hat{\boldsymbol{\Lambda}}\mathbf{X}]^{-1},$$

where $\hat{\boldsymbol{\Lambda}}$ is a diagonal matrix of predicted values. The BHHH estimator is

$$\begin{aligned}\text{Est. Asy. Var}[\hat{\boldsymbol{\beta}}] &= \left[\sum_{i=1}^{n} \left(\frac{\partial \ln P_i}{\partial \hat{\boldsymbol{\beta}}}\right)\left(\frac{\partial \ln P_i}{\partial \hat{\boldsymbol{\beta}}}\right)'\right]^{-1} \\ &= \left[\sum_{i=1}^{n} \left(y_i - \hat{\lambda}_i'\right)^2 \mathbf{x}_i \mathbf{x}_i'\right]^{-1} = [\mathbf{X}'\hat{\mathbf{E}}^2\mathbf{X}]^{-1},\end{aligned}$$

where $\hat{\mathbf{E}}$ is a diagonal matrix of residuals. The Poisson model is one in which the MLE is robust to certain misspecifications of the model, such as the failure to incorporate latent heterogeneity in the mean (i.e., one fits the Poisson model when the negative binomial is appropriate). In this case, a robust covariance matrix is the "sandwich" estimator,

$$\text{Robust Est. Asy. Var}[\hat{\boldsymbol{\beta}}] = [\mathbf{X}'\hat{\boldsymbol{\Lambda}}\mathbf{X}]^{-1}[\mathbf{X}'\hat{\mathbf{E}}^2\mathbf{X}][\mathbf{X}'\hat{\boldsymbol{\Lambda}}\mathbf{X}]^{-1},$$

which is appropriate to accommodate this failure of the model. It has become common to employ this estimator with all specifications, including the negative binomial. One might question the virtue of this. Because the negative binomial model already accounts for the latent heterogeneity, it is unclear what *additional* failure of the assumptions of the model this estimator would be robust to. The questions raised in Section 14.8.3 and 14.8.4 about robust covariance matrices would be relevant here.

A related calculation is used when observations occur in groups that may be correlated. This would include a random effects setting in a panel in which observations have a common latent heterogeneity as well as more general, stratified, and clustered data sets. The parameter estimator is unchanged in this case (and an assumption is made that the estimator is still consistent), but an adjustment is made to the estimated asymptotic covariance matrix. The calculation is done as follows: Suppose the n observations are assembled in G clusters of observations, in which the number of observations in the ith cluster is n_i. Thus, $\sum_{i=1}^{G} n_i = n$. Denote by $\boldsymbol{\beta}$ the full set of model parameters in whatever variant of the model is being estimated. Let the observation-specific gradients and Hessians be $\mathbf{g}_{ij} = \partial \ln L_{ij}/\partial \boldsymbol{\beta} = (y_{ij} - \lambda_{ij})\mathbf{x}_{ij}$ and $\mathbf{H}_{ij} = \partial^2 \ln L_{ij}/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}' = -\lambda_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}'$. The uncorrected estimator of the asymptotic covariance matrix based on the Hessian is

$$\mathbf{V}_H = -\mathbf{H}^{-1} = \left(-\sum_{i=1}^{G}\sum_{j=1}^{n_i} \mathbf{H}_{ij}\right)^{-1}.$$

The corrected asymptotic covariance matrix is

$$\text{Est. Asy. Var}[\hat{\boldsymbol{\beta}}] = \mathbf{V}_H \left(\frac{G}{G-1}\right)\left[\sum_{i=1}^{G}\left(\sum_{j=1}^{n_i} \mathbf{g}_{ij}\right)\left(\sum_{j=1}^{n_i} \mathbf{g}_{ij}\right)'\right]\mathbf{V}_H.$$

Note that if there is exactly one observation per cluster, then this is $G/(G-1)$ times the sandwich (robust) estimator.

18.4.7.b Fixed Effects

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

$$\log \lambda_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i, \tag{18-24}$$

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 17.4.3. 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

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

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

$$\hat{\alpha}_i = \ln\left(\frac{(1/T_i)\sum_{t=1}^{T_i} y_{it}}{(1/T_i)\sum_{t=1}^{T_i} \hat{\mu}_{it}}\right) = \ln\left(\frac{\bar{y}_i}{\bar{\hat{\mu}}_i}\right). \tag{18-25}$$

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, and 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 multinomial logit model discussed earlier is produced:

$$p\left(y_{i1}, y_{i2}, \ldots, y_{iT_i} \,\middle|\, \sum_{i=1}^{T_i} y_{it}\right) = \frac{\left(\sum_{t=1}^{T_i} y_{it}\right)!}{\left(\prod_{t=1}^{T_i} y_{it}!\right)} \prod_{t=1}^{T_i} p_{it}^{y_{it}}, \tag{18-26}$$

where

$$p_{it} = \frac{e^{\mathbf{x}_{it}'\boldsymbol{\beta}+\alpha_i}}{\sum_{t=1}^{T_i} e^{\mathbf{x}_{it}'\boldsymbol{\beta}+\alpha_i}} = \frac{e^{\mathbf{x}_{it}'\boldsymbol{\beta}}}{\sum_{t=1}^{T_i} e^{\mathbf{x}_{it}'\boldsymbol{\beta}}}. \tag{18-27}$$

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.

Hausman, Hall, and Griliches (1984) (HHG) report the following conditional density for the fixed effects negative binomial (FENB) model:

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

which is free of the fixed effects. This is the default FENB formulation used in popular software packages such as SAS, Stata, and LIMDEP. Researchers accustomed to the admonishments that fixed effects models cannot contain overall constants or time-invariant covariates are sometimes surprised to find (perhaps accidentally) that this fixed effects model allows both. [This issue is explored at length in Allison (2000) and Allison and Waterman (2002).] The resolution of this apparent contradiction is that the HHG FENB model is not obtained by shifting the conditional mean function by the fixed effect, $\ln\lambda_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i$, as it is in the Poisson model. Rather, the HHG model is obtained by building the fixed effect into the model as an individual specific θ_i in the Negbin 1 form in (18-22). The conditional mean functions in the models are as follows (we have changed the notation slightly to conform to our earlier formulation):

$$\text{NB1(HHG): } E[y_{it} \mid \mathbf{x}_{it}] = \theta_i\phi_{it} = \theta_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta}),$$

$$\text{NB2: } \qquad E[y_{it} \mid \mathbf{x}_{it}] = \exp(\alpha_i)\phi_{it} = \lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i).$$

The conditional variances are

$$\text{NB1(HHG): } \mathrm{Var}[y_{it} \mid \mathbf{x}_{it}] = \theta_i\phi_{it}[1+\theta_i],$$

$$\text{NB2: } \qquad \mathrm{Var}[y_{it} \mid \mathbf{x}_{it}] = \lambda_{it}[1+\theta\lambda_{it}].$$

Letting $\mu_i = \ln\theta_i$, it appears that the HHG formulation does provide a fixed effect in the mean, as now, $E[y_{it} \mid \mathbf{x}_{it}] = \exp(\mathbf{x}_{it}'\beta + \mu_i)$. Indeed, by this construction, it appears (as the authors suggest) that there are separate effects in both the mean and the variance. They make this explicit by writing $\theta_i = \exp(\mu_i)\gamma_i$ so that in their model,

$$\mathrm{E}[y_{it} \mid \mathbf{x}_{it}] = \gamma_i \exp(\mathbf{x}_{it}'\beta + \mu_i),$$

$$\mathrm{Var}[y_{it} \mid \mathbf{x}_{it}] = \gamma_i \exp(\mathbf{x}_{it}'\beta + \mu_i)/[1 + \gamma_i \exp(\mu_i)].$$

The contradiction arises because the authors assert that μ_i and γ_i are separate parameters. In fact, they cannot vary separately only θ_i can vary autonomously. The firm-specific effect in the HHG model is still isolated in the scaling parameter, which falls out of the conditional density. The mean is homogeneous, which explains why a separate constant, or a time-invariant regressor (or another set of firm-specific effects) can reside there. [See Greene (2007d) and Allison and Waterman (2002) for further discussion.]

18.4.7.c Random Effects

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 11), 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 a negative binomial distribution,

$$p(y_{i1}, \ldots, y_{iT_i}) = \frac{\left[\prod_{t=1}^{T_i} \lambda_{it}^{y_{it}}\right] \Gamma\left(\theta + \sum_{t=1}^{T_i} y_{it}\right)}{\left[\Gamma(\theta) \prod_{t=1}^{T_i} y_{it}!\right] \left[\left(\sum_{t=1}^{T_i} \lambda_{it}\right)^{\sum_{t=1}^{T_i} y_{it}}\right]} Q_i^{\theta} (1 - Q_i)^{\sum_{t=1}^{T_i} y_{it}}, \qquad \textbf{(18-28)}$$

where

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

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

Like the fixed effects model, introducing random effects into the negative binomial model adds some additional complexity. We do note, because the negative binomial model derives from the Poisson model by adding latent heterogeneity to the conditional mean, adding a random effect to the negative binomial model might well amount to introducing the heterogeneity a second time. However, one might prefer to interpret the negative binomial as the density for y_{it} in its own right and treat the common effects in the familiar fashion. Hausman et al.'s (1984) random effects negative binomial (RENB) model is a hierarchical model that is constructed as follows. The heterogeneity is assumed to enter λ_{it} additively with a gamma distribution with mean 1, $\Gamma(\theta_i, \theta_i)$. Then, $\theta_i/(1+\theta_i)$ is assumed to have a beta distribution with parameters a and b

[see Appendix B.4.6)]. The resulting unconditional density after the heterogeneity is integrated out is

$$p(y_{i1}, y_{i2}, \ldots, y_{iT_i}) = \frac{\Gamma(a+b)\Gamma\left(a+\sum_{t=1}^{T_i}\lambda_{it}\right)\Gamma\left(b+\sum_{t=1}^{T_i}y_{it}\right)}{\Gamma(a)\Gamma(b)\Gamma\left(a+\sum_{t=1}^{T_i}\lambda_{it}+b+\sum_{t=1}^{T_i}y_{it}\right)}.$$

As before, the relationship between the heterogeneity and the conditional mean function is unclear, because the random effect impacts the parameter of the scedastic function. An alternative approach that maintains the essential flavor of the Poisson model (and other random effects models) is to augment the NB2 form with the random effect,

$$\text{Prob}(Y = y_{it} \mid \mathbf{x}_{it}, \varepsilon_i) = \frac{\Gamma(\theta + y_{it})}{\Gamma(y_{it}+1)\Gamma(\theta)} r_{it}^{y_{it}}(1-r_{it})^{\theta},$$
$$\lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_i),$$
$$r_{it} = \lambda_{it}/(\theta + \lambda_{it}).$$

We then estimate the parameters by forming the conditional (on ε_i) log-likelihood and integrating ε_i out either by quadrature or simulation. The parameters are simpler to interpret by this construction. Estimates of the two forms of the random effects model are presented in Example 18.10.2 for a comparison.

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. To compute the probabilities or the marginal effects, it is necessary 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), Riphahn et al. (2003), 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 a 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 14.10 and the random parameters model in Section 15.9.5 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.]

Example 18.10 Panel Data Models for Doctor Visits

The German health care panel data set contains 7,293 individuals with group sizes ranging from 1 to 7. Table 18.17 presents the fixed and random effects estimates of the equation for DocVis. The pooled estimates are also shown for comparison. Overall, the panel data treatments bring large changes in the estimates compared to the pooled estimates. There is also a considerable amount of variation across the specifications. With respect to the parameter of interest, *Public,* we find that the size of the coefficient falls substantially with all panel data treatments. Whether using the pooled, fixed, or random effects specifications, the test statistics (Wald, LR) all reject the Poisson model in favor of the negative binomial. Similarly, either common effects specification is preferred to the pooled estimator. There is no simple basis for choosing between the fixed and random effects models, and we have further blurred the distinction by suggesting two formulations of each of them. We do note that the two random effects estimators are producing similar results, which one might hope for. But, the two fixed effects estimators are producing very different estimates. The NB1 estimates include two coefficients, *Income* and *Education,* which are positive, but negative in every other case. Moreover, the coefficient on *Public*, which is large and significant throughout the table, has become small and less significant with the fixed effects estimators.

We also fit a three-class latent class model for these data. (See Section 14.10.) The three class probabilities were modeled as functions of *Married* and *Female,* which appear from the results to be significant determinants of the class sorting. The average prior probabilities for the three classes are 0.09212, 0.49361, and 0.41427. The coefficients on *Public* in the three classes, with associated *t* ratios are 0.3388 (11.541), 0.1907 (3.987), and 0.1084 (4.282). The qualitative result concerning evidence of moral hazard suggested at the outset of Example 18.7 appears to be supported in a variety of specifications (with FE-NB1 the sole exception).

18.4.8 TWO-PART MODELS: ZERO INFLATION AND HURDLE MODELS

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.[15] 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

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

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

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

$$\text{Prob}(z_i = 0 \mid \mathbf{w}_i) = F(\mathbf{w}_i, \boldsymbol{\gamma}),\ (\text{Regime 1} : y \text{ will equal zero.})$$

$$\text{Prob}(y_i = j \mid \mathbf{x}_i, z_i = 1) = \frac{\exp(-\lambda_i)\lambda_i^j}{j!}.(\text{Regime 2} : y \text{ will be a count outcome.})$$

[15] 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)]

TABLE 18.17 Estimated Panel Data Models for Doctor Visits (standard errors in parentheses)

	Poisson				*Negative Binomial*			
					Fixed Effects		*Random Effects*	
Variable	*Pooled (Robust S.E.)*	*Fixed Effects*	*Random Effects*	*Pooled NB2*	*FE-NB1*	*FE-NB2*	*HHG-Gamma*	*Normal*
Constant	0.7162 (0.1319)	0.0000	0.4957 (0.05463)	0.7628 (0.07247)	−1.2354 (0.1079)	0.0000	−0.6343 (0.07328)	0.1169 (0.06612)
Age	0.01844 (0.001336)	0.03115 (0.001443)	0.02329 (0.0004458)	0.01803 (0.0007916)	0.02389 (0.001188)	0.04479 (0.002769)	0.01899 (0.0007820)	0.02231 (0.0006969)
Educ	−0.03429 (0.007255)	−0.03803 (0.01733)	−0.03427 (0.004352)	−0.03839 (0.003965)	0.01652 (0.006501)	−0.04589 (0.02967)	−0.01779 (0.004056)	−0.03773 (0.003595)
Income	−0.4751 (.08212)	−0.3030 (0.04104)	−0.2646 (0.01520)	−0.4206 (0.04700)	0.02373 (0.05530)	−0.1968 (0.07320)	−0.08126 (0.04565)	−0.1743 (0.04273)
Kids	−0.1582 (0.03115)	−0.001927 (0.01546)	−0.03854 (0.005272)	−0.1513 (0.01738)	−0.03381 (0.02116)	−0.001274 (0.02920)	−0.1103 (0.01675)	−0.1187 (0.01582)
Public	0.2365 (0.04307)	0.1015 (0.02980)	0.1535 (0.01268)	0.2324 (0.02900)	0.05837 (0.03896)	0.09700 (0.05334)	0.1486 (0.02834)	0.1940 (0.02574)
θ	0.0000	0.0000	1.1646 (0.01940)	1.9242 (0.02008)	0.0000	1.9199 (0.02994)	0.0000	1.0808 (0.01203)
a	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	2.1463 (0.05955)	0.0000
b	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	3.8011 (0.1145)	0.0000
σ	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9737 (0.008235)
ln L	−104440.3	−47703.34	−71763.13	−60265.49	−34016.16	−49476.36	−58182.52	−58177.66

The zero inflation model can also be viewed as a type of latent class model. The two class probabilities are $F(\mathbf{w}_i, \boldsymbol{\gamma})$ and $1 - F(\mathbf{w}_i, \boldsymbol{\gamma})$, and the two regimes are $y = 0$ and the Poisson or negative binomial data generating process.[16] The extension of the ZIP formulation to the negative binomial model is widely labeled the ZINB model.[17] [See Zaninotti and Falischetti (2010) for an application.]

The mean of this random variable in the Poisson case is

$$E[y_i|\mathbf{x}_i, \mathbf{w}_i] = F_i \times 0 + (1 - F_i) \times E[y_i^*|\mathbf{x}_i, z_i = 1] = (1 - F_i)\lambda_i.$$

Lambert (1992) and Greene (1994) consider a number of alternative formulations, including logit and probit models discussed in Sections 17.2 and 17.3, for the probability of the two regimes.

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 $\text{Prob}[z = 0] = 0$. 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. Because 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. (See Section 14.6.6.) Let $f_j(y_i|\mathbf{x}_i)$ denote the predicted probability that the random variable Y equals y_i under the assumption that the distribution is $f_j(y_i|\mathbf{x}_i)$, for $j = 1, 2$, and let

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

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

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

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 the zero inflation phenomenon. The logic of the testing procedure is to allow for overdispersion by specifying a negative binomial count data process and then examine whether, *even allowing for the overdispersion*, there still appear to be excess zeros. In his application, that appears to be the case.

[16] Harris and Zhao (2007) applied this approach to a survey of teenage smokers and nonsmokers in Australia, using an ordered probit model. (See Section 18.3.)

[17] Greene (2005) presents a survey of two-part models, including the zero inflation models.

TABLE 18.18 Estimated Zero Inflated Count Models

	Poisson			*Negative Binomial*		
		Zero Inflation			*Zero Inflation*	
	Poisson Regression	*Regression*	*Zero Regime*	*Negative Binomial*	*Regression*	*Zero Regime*
Constant	−1.33276	0.75483	2.06919	−1.54536	−0.39628	4.18910
Age	0.01286	0.00358	−0.01741	0.01807	−0.00280	−0.14339
Income	−0.02577	−0.05127	−0.03023	−0.02482	−0.05502	−0.33903
OwnRent	−0.17801	−0.15593	−0.01738	−0.18985	−0.28591	−0.50026
Self Employment	0.04691	−0.01257		0.07920	0.06817	
Dependents	0.13760	0.06038	−0.09098	0.14054	0.08599	−0.32897
Cur. Add.	0.00195	0.00046		0.00245	0.00257	
α				6.41435	4.85653	
ln L	−15467.71	−11569.74		−10582.88	−10516.46	
Vuong		20.6981			4.5943	

Example 18.11 Zero Inflation Models for Major Derogatory Reports

In Example 18.8, we examined the counts of major derogatory reports for a sample of 13,444 credit card applicants. It was noted that there are over 10,800 zeros in the counts. One might guess that among credit card users, there is a certain (probably large) proportion of individuals who would never generate an MDR, and some other proportion who might or might not, depending on circumstances. We propose to extend the count models in Example 10.8 to accommodate the zeros. The extensions to the ZIP and ZINB models are shown in Table 18.18. Only the coefficients are shown for purpose of the comparisons. Vuong's diagnostic statistic appears to confirm intuition that the Poisson model does not adequately describe the data; the value is 20.6981. 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. For the simple Poisson model, the average probability of zero times the sample size gives a prediction of 8609. For the ZIP model, the value is 10914.8, which is a dramatic improvement. By the likelihood ratio test, the negative binomial is clearly preferred; comparing the two zero inflation models, the difference in the log-likelihood functions is over 1,000. As might be expected, the Vuong statistic falls considerably, to 4.5943. However, the simple model with no zero inflation is still rejected by the test.

In some settings, the zero outcome of the data generating process is qualitatively different from the positive ones. The zero or nonzero value of the outcome is the result of a separate decision whether or not to "participate" in the activity. On deciding to participate, the individual decides separately how much, that is, how intensively. Mullahy (1986) argues that this fact constitutes a shortcoming of the Poisson (or negative binomial) model and suggests a **hurdle model** as an alternative.[18] In his formulation, a binary probability model determines whether a zero or a nonzero outcome occurs and then, in the latter case, a (truncated) Poisson distribution describes the positive outcomes. The model is

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

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

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

This formulation changes the probability of the zero outcome and scales the remaining probabilities so that they sum to one. Mullahy suggests some formulations and applies they model to a sample of observations on daily beverage consumption. Mullahy's formulation adds a new restriction that $\text{Prob}(y_i = 0|\mathbf{x}_i)$ no longer depends on the covariates, however. The natural next step is to parameterize this probability. This extension of the hurdle model would combine a binary choice model like those in Section 17.2 and 17.3 with a truncated count model as shown in Section 18.4.6. This would produce, for example, for a logit participation equation and a Poisson intensity equation,

$$\text{Prob}(y_i = 0|\mathbf{w}_i) = \Lambda(\mathbf{w}_i'\boldsymbol{\gamma})$$

$$\text{Prob}(y_i = j|\mathbf{x}_i, \mathbf{w}_i, y_i > 0) = \frac{[1 - \Lambda(w_i'\gamma)]\exp(-\lambda_i)\lambda_i^j}{j![1 - \exp(-\lambda_i)]}.$$

The conditional mean function in the hurdle model is

$$E[y_i|\mathbf{x}_i, \mathbf{w}_i] = \frac{[1 - F(\mathbf{w}_i'\boldsymbol{\gamma})]\lambda_i}{[1 - \exp(-\lambda_i)]}, \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}),$$

where $F(.)$ is the probability model used for the participation equation (probit or logit). The partial effects are obtained by differentiating with respect to the two sets of variables separately,

$$\frac{\partial E[y_i|\mathbf{x}_i, \mathbf{w}_i]}{\partial \mathbf{x}_i} = [1 - F(\mathbf{w}_i'\boldsymbol{\gamma})]\boldsymbol{\delta}_i,$$

$$\frac{\partial E[y_i|\mathbf{x}_i, \mathbf{w}_i]}{\partial \mathbf{w}_i} = \left\{\frac{-f(\mathbf{w}_i'\gamma)\lambda_i}{[1 - \exp(-\lambda_i)]}\right\}\gamma,$$

where $\boldsymbol{\delta}_i$ is defined in (18-23) and $f(.)$ is the density corresponding to $F(.)$. For variables that appear in both $\mathbf{x}_i$ and $\mathbf{w}_i$, the effects are added. For dummy variables, the preceding would be an approximation; the appropriate result would be obtained by taking the difference of the conditional mean with the variable fixed at one and zero.

It might be of interest to test for hurdle effects. The hurdle model is similar to the zero inflation model in that a model without hurdle effects is not nested within the hurdle model; setting $\boldsymbol{\gamma} = \mathbf{0}$ produces either $F = \alpha$, a constant, or $F = 1/2$ if the constant term is also set to zero. Neither serves the purpose. Nor does forcing $\boldsymbol{\gamma} = \boldsymbol{\beta}$ in a model with $\mathbf{w}_i = \mathbf{x}_i$ and $F = \Lambda$ with a Poisson intensity equation, which might be intuitively appealing. A complementary log log model with

$$\text{Prob}(y_i = 0|\mathbf{w}_i) = \exp[-\exp(\mathbf{w}_i'\boldsymbol{\gamma})]$$

does produce the desired result if $\mathbf{w}_i = \mathbf{x}_i$. In this case, "hurdle effects" are absent if $\boldsymbol{\gamma} = \boldsymbol{\beta}$. The strategy in this case, then, would be a test of this restriction. But, this formulation is otherwise restrictive, first in the choice of variables and second in its unconventional functional form. The more general approach to this test would be the Vuong test used earlier to test the zero inflation model against the simpler Poisson or negative binomial model.

The hurdle model bears some similarity to the zero inflation model; however, the behavioral implications are different. The zero inflation model can usefully be viewed as a latent class model. The splitting probability defines a regime determination. In the hurdle model, the splitting equation represents a behavioral outcome on the same

TABLE 18.19 Estimated Hurdle Model for Doctor Visits

	Participation Equation		*Intensity Equation*		*Total Partial Effect*
	Parameter	*Partial Effect*	*Parameter*	*Partial Effect*	*(Poisson Model)*
Constant	−0.0598		1.1203		
Age	0.0221	0.0244	0.0113	0.0538	0.0782 (0.0625)
Income	0.0725	0.0800	−0.5152	−2.4470	−2.3670 (−1.8130)
Kids			−0.0842	−0.4000	−0.4000 (−0.4836)
Public	0.2411	0.2663	0.1966	0.9338	1.2001 (0.9744)
Education	−0.0291	−0.0321			−0.0321
Married	−0.0233	−0.0258			−0.0258
Working	−0.3624	−0.4003			−0.4003

level as the intensity (count) equation. 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 may misinterpret the excess zeros as due to the splitting process instead of the heterogeneity.

Example 18.12 Hurdle Model for Doctor Visits

The hurdle model is a natural specification for models of utilization of the health care system, and has been used in a number of studies. Table 18.19 shows the parameter estimates for a hurdle model for doctor visits based on the entire pooled sample of 27,326 observations. The decomposition of the partial effects shows that the participation and intensity decisions each contribute substantively to the effects of Age, Income, and Public insurance. The value of the Vuong statistic is 51.16, strongly in favor of the hurdle model compared to the pooled Poisson model with no hurdle effects. The effect of the hurdle model on the partial effects is shown in the last column where the results for the Poisson model are shown in parentheses.

18.4.9 ENDOGENOUS VARIABLES AND ENDOGENOUS PARTICIPATION

As in other situations, one would expect to find endogenous variables in models for counts. For example, in the study on which we have relied for our examples of health care utilization, Riphahn, Wambach, and Million (RWM, 2003), the authors were interested in the role of insurance (specifically the *Add-On* insurance) in the usage variable. One might expect the choice to buy insurance to be at least partly influenced by some of the same factors that motivate usage of the health care system. Insurance purchase might well be endogenous in a model such as the hurdle model in Example 18.12.

The Poisson model presents a complication for modeling endogeneity that arises in some other cases as well. For simplicity, consider a continuous variable, such as *Income*, to continue our ongoing example. A model of income determination and doctor visits might appear

$$Income = \mathbf{z}_i'\boldsymbol{\gamma} + u_i,$$

$$\text{Prob}(DocVis_i = j|\mathbf{x}_i, Income_i) = \exp(-\lambda_i), \lambda_i^j/j!, \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \delta\, Income_i).$$

Endogeneity as we have analyzed it, for example, in Chapter 8 and Sections 17.3.5 and 17.5.5, arises through correlation between the endogenous variable and the unobserved omitted factors in the main equation. But, the Poisson model does not contain any unobservables. This is a major shortcoming of the specification as a "regression" model; all of the regression variation of the dependent variable arises through variation of the observables. There is no accommodation for unobserved heterogeneity or omitted factors. This is the compelling motivation for the negative binomial model or, in RWM's case, the Poisson-normal mixture model. [See Terza (2010, pp. 555–556) for discussion of this issue.] If the model is reformulated to accommodate heterogeneity, as in

$$\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\delta}\ Income_i + \varepsilon_i),$$

then $Income_i$ will be endogenous if u_i and ε_i are correlated.

A bivariate normal model for (u_i, ε_i) with zero means, variances σ_u^2 and σ_ε^2 and correlation ρ provides a convenient (and the usual) platform to operationalize this idea. By projecting ε_i on u_i, we have

$$\varepsilon_i = (\rho\sigma_\varepsilon/\sigma_u)u_i + v_i,$$

where v_i is normally distributed with mean zero and variance $\sigma_\varepsilon^2(1-\rho^2)$. It will prove convenient to parameterize these based on the regression and the specific parameters as follows:

$$\begin{aligned}\varepsilon_i &= \rho\sigma_\varepsilon(Income_i - \mathbf{z}_i'\boldsymbol{\gamma})/\sigma_u + v_i,\\ &= \tau[(Income_i - \mathbf{z}_i'\boldsymbol{\gamma})/\sigma_u] + \theta w_i.\end{aligned}$$

where w_i will be normally distributed with mean zero and variance one while $\tau = \rho\sigma_\varepsilon$ and $\theta^2 = \sigma_\varepsilon^2(1-\rho^2)$. Then, combining terms,

$$\varepsilon_i = \tau\ u_i^* + \theta w_i.$$

With this parameterization, the conditional mean function in the Poisson regression model is

$$\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \delta\ Income_i + \tau u_i^* + \theta w_i).$$

The parameters to be estimated are $\boldsymbol{\beta}, \boldsymbol{\gamma}, \delta, \sigma_\varepsilon, \sigma_u$, and ρ. There are two ways to proceed. A two-step method can be based on the fact that $\boldsymbol{\gamma}$ and σ_u can consistently be estimated by linear regression of *Income* on $\mathbf{z}$. After this first step, we can compute values of u_i^* and formulate the Poisson regression model in terms of

$$\hat{\lambda}_i(w_i) = \exp[\mathbf{x}_i'\boldsymbol{\beta} + \delta\ Income_i + \tau\hat{u}_i + \theta w_i].$$

The log-likelihood to be maximized at the second step is

$$\ln L(\boldsymbol{\beta},\ \delta,\ \tau, \theta|\mathbf{w}) = \sum_{i=1}^{n} -\hat{\lambda}_i(w_i) + y_i \ln \hat{\lambda}_i(w_i) - \ln y_i!.$$

A remaining complication is that the unobserved heterogeneity, w_i remains in the equation so it must be integrated out of the log-likelihood function. The unconditional log-likelihood function is obtained by integrating the standard normally distributed w_i out

of the conditional densities.

$$\ln L(\boldsymbol{\beta},\ \boldsymbol{\gamma},\ \tau,\ \theta) = \sum_{i=1}^{n} \ln \left\{ \int_{-\infty}^{\infty} \left[\frac{\exp\left(-\hat{\lambda}_i(w_i)\right)\left(\hat{\lambda}_i(w_i)\right)^{y_i}}{y_i!} \right] \phi(w_i) dw_i \right\}.$$

The method of Butler and Moffitt or maximum simulated likelihood that we used to fit a probit model in Section 17.4.2 can be used to estimate $\boldsymbol{\beta}$, δ, τ, and θ. Estimates of ρ and σ_ε can be deduced from the last two of these; $\sigma_\varepsilon^2 = \theta^2 + \tau^2$ and $\rho = \tau/\sigma_\varepsilon$. This is the control function method discussed in Section 17.3.5 and is also the "residual inclusion" method discussed by Terza, Basu, and Rathouz (2008).

The full set of parameters can be estimated in a single step using **full information maximum likelihood**. To estimate all parameters simultaneously and efficiently, we would form the log-likelihood from joint density of *DocVis* and *Income* as $P(DocVis|\ Income)\ f(Income)$. Thus,

$$f(DocVis,\ Income) = \frac{\exp\left[-\lambda_i(w_i)\right]\left[\lambda_i(w_i)\right]^{y_i}}{y_i!} \frac{1}{\sigma_u} \phi\left(\frac{Income - \mathbf{z}_i'\boldsymbol{\gamma}}{\sigma_u}\right)$$

$$\lambda_i(w_i) = \exp\left(\mathbf{x}_i'\boldsymbol{\beta} + \delta\ Income_i + \tau(Income_i - \mathbf{z}_i'\boldsymbol{\gamma})/\sigma_u + \theta w_i\right)$$

As before, the unobserved w_i must be integrated out of the log-likelihood function. Either quadrature or simulation can be used. The parameters to be estimated by maximizing the full log-likelihood are $(\beta, \gamma, \delta, \sigma_u, \sigma_\varepsilon, \rho)$. The invariance principle has been used to simplify the estimation a bit by parameterizing the log-likelihood function in terms of τ and θ. Some additional simplification can also be obtained by using the Olsen (1978) [and Tobin (1958)] transformations, $\eta = 1/\sigma_u$ and $\alpha = (1/\sigma_u)\boldsymbol{\gamma}$.

An endogenous binary variable, such as *Public* or *AddOn* in our *DocVis* example is handled similarly but is a bit simpler. The structural equations of the model are

$$\begin{aligned} T^* &= \mathbf{z}_i'\boldsymbol{\gamma} + u_i, & u &\sim N[0, 1], \\ T &= 1(T^* > 0), & & \\ \lambda &= \exp(\mathbf{x}'\boldsymbol{\beta} + \delta T + \varepsilon) & \varepsilon &\sim N[0, \sigma_\varepsilon^2], \end{aligned}$$

with $\text{Cov}(u, \varepsilon) = \rho\sigma_\varepsilon$. The endogeneity of T is implied by a nonzero ρ. We use the bivariate normal result

$$u = (\rho/\sigma_\varepsilon)\varepsilon + v$$

where v is normally distributed with mean zero and variance $1 - \rho^2$. Then, using our earlier results for the probit model (Section 17.2),

$$P(T|\varepsilon) = \Phi\left[(2T-1)\left(\frac{\mathbf{z}'\boldsymbol{\gamma} + (\rho/\sigma_\varepsilon)\varepsilon}{\sqrt{1-\rho^2}}\right)\right],\ T = 0, 1.$$

It will be convenient once again to write $\varepsilon = \sigma_\varepsilon w$ where $w \sim \text{N}[0, 1]$. Making the substitution, we have

$$P(T|w) = \Phi\left[(2T-1)\left(\frac{\mathbf{z}'\boldsymbol{\gamma} + \rho w}{\sqrt{1-\rho^2}}\right)\right],\ T = 0, 1.$$

The probability density function for $y|T, w$ is Poisson with $\lambda(w) = \exp(\boldsymbol{x}'\boldsymbol{\beta} + \delta T + \sigma_\varepsilon w)$. Combining terms,

$$P(y, T|w) = \frac{\exp[-\lambda(w)][\lambda(w)]^y}{y!}\Phi\left[(2T-1)\left(\frac{\boldsymbol{z}'\boldsymbol{\gamma} + \rho w}{\sqrt{1-\rho^2}}\right)\right].$$

This last result provides the terms that enter the log-likelihood for $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \delta, \rho, \sigma_\varepsilon)$. As before, the unobserved heterogeneity, w, must be integrated out of the log-likelihood, so either the quadrature or simulation method discussed in Chapter 17 is used to obtain the parameter estimates. Note that this model may also be estimated in two steps, with $\boldsymbol{\gamma}$ obtained in the first-step probit. The two-step method will not be appreciably simpler, since the second term in the density must remain to identify ρ. The residual inclusion method is not fesible here since T^* is not observed.

This same set of methods is used to allow for endogeneity of the participation equation in the hurdle model in Section 18.4.8. Mechanically, the hurdle model with endogenous participation is essentially the same as the endogenous binary variable. [See Greene (2005, 2007).]

18.5 SUMMARY AND CONCLUSIONS

The analysis of individual decisions in microeconometrics is largely about discrete decisions such as whether to participate in an activity or not, whether to make a purchase or not, or what brand of product to buy. This chapter and Chapter 17 have developed the four essential models used in that type of analysis. Random utility, the binary choice model, and regression-style modeling of probabilities developed in Chapter 17 are the three fundamental building blocks of discrete choice modeling. This chapter extended those tools into the three primary areas of choice modeling, unordered choice models, ordered choice models, and models for counts. In each case, we developed a core modeling framework that provides the broad platform and then developed a variety of extensions.

In the analysis of unordered choice models, such as brand or location, the multinomial logit (MNL) model has provided the essential starting point. The MNL works well to provide a basic framework, but as a behavioral model in its own right, it has some important shortcomings. Much of the recent research in this area has focused on relaxing these behavioral assumptions. The most recent research in this area, on the mixed logit model, has produced broadly flexible functional forms that can match behavioral modeling to empirical specification and estimation.

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. We began this analysis with the ordered probit and logit model pioneered by Zavoina and McKelvey (1975). Recent developments of this model have produced the same sorts of extensions to panel data and modeling heterogeneity that we considered in Chapter 17 for binary choice. We also examined some multiple-equation specifications. For all its versatility, the familiar ordered choice models have an important shortcoming in the assumed constancy underlying preference behind the rating scale. The current work on differential item

functioning, such as King et al. (2004), has produced significant progress on filling this gap in the theory.

Finally, we examined probability models for counts of events. Here, the Poisson regression model provides the broad framework for the analysis. The Poisson model has two shortcomings that have motivated the current stream of research. The functional form binds the mean of the random variable to its variance, producing an unrealistic regression specification. Second, the basic model has no component that accommodates unmeasured heterogeneity. (This second feature is what produces the first.) Current research has produced a rich variety of models for counts, such as two-part behavioral models that account for many different aspects of the decision-making process and the mechanisms that generate the observed data.

Key Terms and Concepts

- Bivariate ordered probit
- Censoring
- Choice based sample
- Conditional logit model
- Count data
- Deviance
- Differential item functioning (DIF)
- Event count
- Exposure
- Full information maximum likelihood (FIML)
- Heterogeneity
- Hurdle model
- Identification through functional form
- Inclusive value
- Independence from irrelevant alternatives (IIA)
- Lagrange multiplier test
- Limited information
- Log-odds
- Loglinear model
- Method of simulated moments
- Mixed logit model
- Multinomial choice
- Multinomial logit model
- Multinomial probit model (MNP)
- Negative binomial model
- Negbin 1 (NB1) form
- Negbin 2 (NB2) form
- Negbin P(NBP) model
- Nested logit model
- Nonnested models
- Ordered choice model
- Overdispersion
- Parallel regression assumption
- Poisson regression model
- Random coefficients
- Random parameters logit model (RPL)
- Revealed preference data
- Specification error
- Stated choice experiment
- Subjective well-being
- Unordered choice model
- Willingness to pay space
- Zero inflated Poisson model (ZIP)

Exercises

1. We are interested in the ordered probit model. Our data consist of 250 observations, of which the responses are

$\boldsymbol{y}$	0	1	2	3	4
$\boldsymbol{n}$	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.)
2. For the zero-inflated Poisson (ZIP) model in Section 18.4.8, we derived the conditional mean function, $E[y_i|\mathbf{x}_i, \mathbf{w}_i] = (1 - F_i)\lambda_i$.
   a. For the same model, now obtain $Var[y_i|\mathbf{x}_i, \mathbf{w}_i]$. Then, obtain $\tau_i = \text{Var}[y_i|\mathbf{x}_i, \mathbf{w}_i]/E[y_i|\mathbf{x}_i, \mathbf{w}_i]$. Does the zero inflation produce overdispersion? (That is, is the ratio greater than one?)
   b. Obtain the partial effect for a variable z_i that appears in both $\mathbf{w}_i$ and $\mathbf{x}_i$.

3. Consider estimation of a Poisson regression model for $y_i|x_i$. The data are truncated on the left—these are on-site observations at a recreation site, so zeros do not appear in the data set. The data are censored on the right—any response greater than 5 is recorded as a 5. Construct the log-likelihood for a data set drawn under this sampling scheme.

Applications

1. Appendix Table F18.1 provides Fair's (1978) *Redbook Magazine* survey on extramarital affairs. The variables in the data set are as follows:

   id = an identification number
   C = constant, value = 1
   yrb = a constructed measure of time spent in extramarital affairs
   v_1 = a rating of the marriage, coded 1 to 5
   v_2 = age, in years, aggregated
   v_3 = number of years married
   v_4 = number of children, top coded at 5
   v_5 = religiosity, 1 to 4, 1 = not, 4 = very
   v_6 = education, coded 9, 12, 14, 16, 17, 20
   v_7 = occupation
   v_8 = husband's occupation

   and three other variables that are not used. The sample contains a survey of 6,366 married women. For this exercise,we will analyze, first, the binary variable $A = 1$ if $yrb > 0$,0 otherwise. The regressors of interest are v_1 to v_8; however, not necessarily all of them belong in your model. Use these data to build a binary choice model for *A*. Report all computed results for the model. Compute the marginal effects for the variables you choose. Compare the results you obtain for a probit model to those for a logit model. Are there any substantial differences in the results for the two models?
2. Continuing the analysis of the first application, we now consider the self-reported rating, v_1. This is a natural candidate for an ordered choice model, because the simple four-item coding is a censored version of what would be a continuous scale on some subjective satisfaction variable. Analyze this variable using an ordered probit model. What variables appear to explain the response to this survey question? (*Note:* The variable is coded 1, 2, 3, 4, 5. Some programs accept data for ordered choice modeling in this form, for example, *Stata*, while others require the variable to be coded 0, 1, 2, 3, 4, for example, *LIMDEP*. Be sure to determine which is appropriate for the program you are using and transform the data if necessary.) Can you obtain the partial effects for your model? Report them as well. What do they suggest about the impact of the different independent variables on the reported ratings?
3. Several applications in the preceding chapters using the German health care data have examined the variable Doc Vis, the reported number of visits to the doctor. The data are described in Appendix Table F7.1. A second count variable in that data set that we have not examined is Hosp Vis, the number of visits to hospital. For this application, we will examine this variable. To begin, we treat the full sample (27,326) observations as a cross section.

a. Begin by fitting a Poisson regression model to this variable. The exogenous variables are listed in Appendix Table F7.1. Determine an appropriate specification for the right-hand side of your model. Report the regression results and the partial effects.
b. Estimate the model using ordinary least squares and compare your least squares results to the partial effects you computed in part a. What do you find?
c. Is there evidence of over dispersion in the data? Test for overdispersion. Now, reestimate the model using a negative binomial specification.What is the result? Do your results change? Use a likelihood ratio test to test the hypothesis of the negative binomial model against the Poisson.

4. The GSOEP data are an unbalanced panel, with 7,293 groups. Continue your analysis in Application 3 by fitting the Poisson model with fixed and with random effects and compare your results. (Recall, like the linear model, the Poisson fixed effects model may not contain any time-invariant variables.) How do the panel data results compare to the pooled results?
5. Appendix Table F18.2 contains data on ship accidents reported in McCullagh and Nelder (1983). The data set contains 40 observations on the number of incidents of wave damage for oceangoing ships. Regressors include "aggregate months of service", and three sets of dummy variables, Type (1, . . . , 5), operation period (1960–1974 or 1975–1979), and construction period (1960–1964, 1965–1969, or 1970–1974). There are six missing values on the dependent variable, leaving 34 usable observations.
   a. Fit a Poisson model for these data, using the log of service months, four types of dummy variables, two construction period variables, and one operation period dummy variable. Report your results.
   b. The authors note that the rate of accidents is supposed to be per period, but the exposure (aggregate months) differs by ship. Reestimate your model constraining the coefficient on log of service months to equal one.
   c. The authors take overdispersion as a given in these data. Do you find evidence of over dispersion? Show your results.