18
MULTINOMIAL 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 from more than two choices, once again, making the choice that provides the greatest utility. Applications include the choices of political candidates, how to commute to work, which energy supplier to use, what health care plan to choose, where to live, or what brand of car, appliance, or food product to buy.

Ordered Choice: The individual reveals the strength of his or her preferences with respect to a single outcome. Familiar cases involve survey questions about strength of feelings regarding 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 prevalence 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, the number of customers that arrive at a service point per unit of time, or the number of messages that arrive at a switch 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 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. Hensher and Greene (2007b) analyze commuting between Sydney and Melbourne by a sample of individuals who choose from 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 (2009) studied the choice of electricity supplier by a sample of California electricity customers.
5. Michelsen and Madlener (2012) studied homeowners' choice of type of heating appliance to install in a new home.
6. Hensher, Rose, and Greene (2015) analyzed choices of automobile models by a sample of consumers offered a hypothetical menu of features.
7. Lagarde (2013) examined the choice of different sets of guidelines for preventing malaria by a sample of individuals in Ghana.

In each of these cases, there is a single decision based on two or more alternatives. In this and the next section, we will encounter two broad types of multinomial choice sets, **unordered choices** and **ordered choices**. All of the choice sets listed above 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 examine 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 (2015).

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

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

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) distributions,

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

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**. (lt 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 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.

Consider a model of 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 the 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 $\mathbf{w}$. 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

$$\mathbf{Z}_i = \begin{bmatrix} S_{i1} & D_{i1} & 1 & 0 & I_i & 0 \\ S_{i2} & D_{i2} & 0 & 1 & 0 & I_i \\ S_{i3} & D_{i3} & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The probabilities for this model would be

$$\text{Prob}(Y_i = j|\mathbf{Z}_i) = \frac{\exp\begin{pmatrix} Stores_{ij}\,\beta_1 + Distance_{ij}\,\beta_2 + \\ A_j\alpha_j + A_j\,Income_i\,\gamma_j \end{pmatrix}}{\sum_{j=1}^{3}\exp\begin{pmatrix} Stores_{ij}\,\beta_1 + Distance_{ij}\,\beta_2 + \\ A_j\alpha_j + A_j\,Income_i\gamma_j \end{pmatrix}}, \alpha_3 = \gamma_3 = 0.$$

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 multinomial logit model[1] for occupational choice is

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

(The binomial logit model in Section 17.3 is conveniently produced as the special case of $J = 1$.) The 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 nonzero vector $\mathbf{q}$, then recomputing the probabilities in (18-4) 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

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

[1]Nerlove and Press (1973) is a pioneering study in this literature, also about labor market choices.

The form of the binary choice model examined in Section 17.2 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 and identical distributions of the random terms in the original model. From a behavioral viewpoint, this fact turns out not to be 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,

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

The derivatives have the characteristically simple form

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

The exact second derivatives matrix has J^2 $K \times K$ 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 characteristics.

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. Note that all of the $\boldsymbol{\alpha}_j$'s appear in the denominator of P_{ij}. By differentiating (18-5), we find that the partial effects of the characteristics on the probabilities are

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

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}.

[2] If the data were in the form of proportions, such as market shares, then the appropriate log likelihood and derivatives are $\Sigma_i\Sigma_j n_i \ln p_{ij}$ 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.

Standard errors can be estimated using the delta method. (See Section 4.6.) 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 $\boldsymbol{\alpha}_0 = \mathbf{0}$, $\boldsymbol{\delta}_{i0} = -P_{i0}\overline{\boldsymbol{\alpha}}$, which is not $\mathbf{0}$. Note as well that $\text{Asy.Cov}[\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}}_j] = \mathbf{0}$ for $j = 1, \ldots, J$. Then

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

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

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 terms, 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.[4] 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

[3]It is common for this rule to predict all observations 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.

[4]See, also Bornstein and Bradley (2003).

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
		(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
		(−0.73)	(−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
		(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
		(−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)

by some viewpoint, the outcome for this experiment is ordered so the model in Section 18.3 might be more appropriate.) 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 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. Note that the rows of partial effects sum to zero. The interpretation of this result is that when a characteristic such as age changes, the probabilities change in turn. But they sum to one before and after the change.

Example 18.2 Home Heating Systems

Michelsen and Madlener (2012) studied the preferences of homeowners for adoption of innovative residential heating systems. The analysis was based on a survey of 2,240 German homeowners who installed one of four types of new heating systems*: GAS-ST* = gas-fired condensing boiler with solar thermal support, *OIL-ST* = oil-fired condensing boiler with solar thermal support, *HEAT-P* = heat pump, and *PELLET* = wood pellet-fired boiler. Variables in the model included sociodemographics such as age, income and gender; home characteristics such as size, age, and previous type of heating system; location and some specific characteristics, including preference for energy savings (on a five-point scale), preference for more independence from fossil fuels and, also on a five-point scale, preference for environmental protection. The authors reported only the average partial effects for the many variables (not the estimated coefficients). Two, in particular, were the survey data on

environmental protection and energy independence. They reported the following average partial effects for these two variables:

	GAS-ST	*OIL-ST*	*HEAT-P*	*PELLET*
Environment	0.002	−0.003	−0.022	0.024
Independence	−0.150	−0.043	0.100	0.093

The precise meaning of the changes in the two variables are unclear, as they are five-point scales treated as if they were continuous. Nonetheless, the substitution of technologies away from fossil fuels is suggested in the results. The desire to reduce CO_2 emissions is less obvious in the environmental protection results.[5]

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

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

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 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

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

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. There is a striking result in the elasticity; $\partial \ln P_{ij}/\partial \ln x_{mk}$ is not a function of P_{ij}. This is a strong implication of the particular functional form assumed at the outset. It implies the rather peculiar substitution pattern that can be seen in the top panel of Table 18.8, below. We will explore this result in Section 18.2.4. Much of the research on multinomial choice modeling over the past several decades has focused on more general forms (including several that we will examine here) that provide more realistic behavioral results. Some applications are developed in Example 18.3.

[5]The results were extracted from their Table 6, p. 1279.

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

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

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.

We have maintained a distinction between the multinomial logit model based on characteristics of the individual and the conditional logit model based on the attributes of the choices). The distinction is completely artificial. Applications of multinomial choice modeling usually mix the two forms—our example below related to travel mode choice includes attributes of the modes as well as household income. The general form of the multinomial logit model that appears in applications, based on (18-3), would be

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

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 for estimation, but it is not a particularly appealing restriction to place on consumer behavior. An additional consequence, also unattractive, is the peculiar pattern of substitution elasticities that is implied by the multinomial logit form. The property of the logit model whereby P_{ij}/P_{im} is independent of the remaining probabilities, and $\partial \ln P_{ij}/\partial \ln x_{im}$ is not a function of P_{ij}, is called the **independence from irrelevant alternatives (IIA)**.

The independence assumption follows from the initial assumption that the random components of the utility functions 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. The unconditional probability of choice j in the MNL model is

$$\text{Prob}(Y_i = j) = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}{\sum_{m=1}^{J}\exp(\mathbf{x}_{im}'\boldsymbol{\beta}_m)}.$$

Consider the probability of choice j in a reduced choice set, say in alternatives 1 to $J-1$. This would be

$$\frac{\text{Prob}[Y = j \text{ and } j \in (1, \ldots, J-1)]}{\text{Prob}(j \in (1, \ldots, J-1))} = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}{\sum_{m=1}^{J}\exp(\mathbf{x}_{im}'\boldsymbol{\beta}_m)} \Big/ \frac{\sum_{j=1}^{J-1}\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}{\sum_{m=1}^{J}\exp(\mathbf{x}_{im}'\boldsymbol{\beta}_m)}$$
$$= \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}{\sum_{m=1}^{J-1}\exp(\mathbf{x}_{im}'\boldsymbol{\beta}_m)}.$$

This is the same model, with the denominator summed from 1 to $J-1$, instead. The MNL model survives the restriction of the choice set—that is, the parameters of the model would be the same. Hausman and McFadden (1984) suggest that if a subset of the choice set truly is irrelevant, then omitting it from the model altogether will not change parameter estimates systematically. Exclusion of these choices (and the observations that choose them) 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 estimators 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. We will examine an application in Example 18.3.

18.2.5 ALTERNATIVE CHOICE MODELS

The multinomial logit form imposes some unattractive restrictions on the pattern of behavior in the choice process. A large variety of alternative models in a long thread of research have been developed that relax the restrictions of the MNL model.[6] Two specific restrictions are the homoscedasticity across choices and individuals of the utility functions and the lack of correlation across the choices. We consider three alternatives to the MNL model. Note it is not simply the distribution at work. Changing the model to a *multinomial probit* model based on the normal distribution, but still independent and homoscedastic, does not solve the problem.

[6]One of the earliest contributions to this literature is Gaudry and Dagenais's (1979) "DOGIT" model that "[D]odges the researcher's dilemma of choosing a priori between a format which commits to IIA restrictions and one which excludes them …." (p. 105.) The DOGIT functional form is $P_j = (V_j + \lambda_j \Sigma_m V_m)/[(1 + \Sigma_m \lambda_m)\Sigma_m V_m]$, where $V_j = \exp(\mathbf{x}_{ij}'\boldsymbol{\beta})$ and $\lambda_j \geq 0$.

18.2.5.a Heteroscedastic Extreme Value Model

The variance of ε_{ij} in (18-1) is equal to $\pi^2/6$. The heteroscedastic extreme value (HEV) specification developed by Bhat (1995) allows a separate variance,

$$\sigma_j^2 = \pi^2/(6\theta_j^2), \tag{18-9}$$

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. We can allow heterogeneity across individuals as well as across choices by specifying

$$\theta_{ij} = \theta_j \times \exp(\boldsymbol{\phi}'\mathbf{h}_i). \tag{18-10}$$

[See Salisbury and Feinberg (2010) and Louviere and Swait (2010) for applications of this type of HEV model.] The heteroscedasticity alone interrupts the IIA assumption.

18.2.5.b 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

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

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 17.2.3.) For more than two choices, this specification is far more general than the MNL model, which assumes that $\boldsymbol{\Sigma} = (\pi^2/6)\mathbf{I}$. (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 greater generality of the multinomial probit is produced by the correlations across the alternatives (and, to a lesser extent, by the possible heteroscedasticity). The distribution itself is a lesser extension. An MNP model that simply substitutes a normal distribution with $\Sigma = \mathbf{I}$ will produce virtually the same results (probabilities and elasticities) as the multinomial logit model. An obstacle to implementation of the MNP model has been the difficulty in computing the multivariate normal probabilities for models with many alternatives.[7] Results on accurate simulation of multinormal integrals

[7]Hausman and Wise (1978) point out that 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, for example, a model with four choices requires only the evaluation of trivariate normal integrals, bivariate if only the derivatives of the log likelihood are needed.

using the GHK simulator 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*.) Computation is exceedingly time consuming. It is also necessary to ensure that $\boldsymbol{\Sigma}$ remain a positive definite matrix. One way often suggested is to construct the Cholesky decomposition of $\boldsymbol{\Sigma}$, $\mathbf{LL}'$, where $\mathbf{L}$ is a lower triangular matrix, and estimate the elements of $\mathbf{L}$. The normalizations and zero restrictions can be imposed by making the last row of the $J \times J$ matrix $\boldsymbol{\Sigma}$ 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 $\mathbf{L}_{11} = 1$.

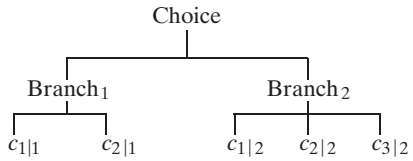
The identification restrictions in Σ needed to identify the model can appear in different places. For example, it is arbitrary which alternative provides the numeraire, and any other row of $\boldsymbol{\Sigma}$ can be normalized. One consequence is that it is not possible to compare directly the estimated coefficient vectors, $\boldsymbol{\beta}$, in the MNP and MNL models. The substantive differences between estimated models are revealed by the predicted probabilities and the estimated elasticities.

18.2.5.c The Nested Logit Model

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:



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

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

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

$$P_{ij|b} = \frac{\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})}{\sum_{j=1}^{J_b}\exp(\mathbf{x}'_{ij|b}\boldsymbol{\beta})} \quad \text{and} \quad P_b = \frac{\exp[\tau_b(\mathbf{z}'_{ib}\boldsymbol{\gamma} + IV_{ib})]}{\sum_{b=1}^{B}\exp[\tau_b(\mathbf{z}'_{ib}\boldsymbol{\gamma} + IV_{ib})]}, \qquad \textbf{(18-11)}$$

where the new parameters τ_l must equal 1 to produce the original MNL 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 and cross alternative correlation. Within each branch, the IIA restriction continues to hold. The equal variance of the disturbances within the jth branch are now[8]

$$\sigma_b^2 = \frac{\pi^2}{6\tau_b}. \qquad \textbf{(18-12)}$$

With $\tau_j = 1$, this reverts to the basic result for the multinomial logit model. The nested logit model is equivalent to a random utility model with block diagonal covariance matrix. For example, for the four-choice model examined in Example 18.3, the model is equivalent to a RUM with

$$\Sigma = \begin{bmatrix} \sigma_F^2 & 0 & 0 & 0 \\ 0 & \sigma_G^2 & \sigma_G^2\rho & \sigma_G^2\rho \\ 0 & \sigma_G^2\rho & \sigma_G^2 & \sigma_G^2\rho \\ 0 & \sigma_G^2\rho & \sigma_G^2\rho & \sigma_G^2 \end{bmatrix}.$$

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

$$\frac{\partial \ln \text{Prob}[\textit{choice} = m, \textit{branch} = b]}{\partial x_k \text{ in } \textit{choice } M \text{ and } \textit{branch } B}$$
$$= \{\mathbf{1}(b = B)[\mathbf{1}(m = M) - P_{M|B}] + \tau_B[\mathbf{1}(b = B) - P_B]P_{M|B}\}\beta_k.$$

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

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.

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.[9] For full information maximum likelihood (FIML) estimation of the model, the log likelihood is[10]

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

The information matrix is not block diagonal in $\boldsymbol{\beta}$ and $(\boldsymbol{\gamma}, \boldsymbol{\tau})$, so FIML estimation will be more efficient than two-step estimation. The FIML estimator is now available in several commercial computer packages. (It also solves the problem if efficiently mixing the B different estimators of $\boldsymbol{\beta}$ that are produced by reestimation with each branch.)

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.

Example 18.3 Multinomial Choice Model for Travel Mode

Hensher and Greene[11] 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 wage-like 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.

The model specified is

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

[9]See McFadden (1984).

[10]See Hensher (1986, 1991) and Greene (2007b).

[11]See Greene (2016).

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	42.22			

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

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 variance $\pi^2/6$. The mean of -0.5772 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. (See Section 17.5.4.) 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 that 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, we would lose 58 observations. In addition, for every data vector left in the sample, the air-specific constant

TABLE 18.3 Parameter Estimates for Multinomial Logit Model

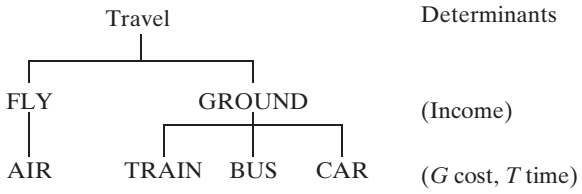
	Unweighted Sample		*Choice-Based Sample Weighting*	
	Estimate	*t Ratio*	*Estimate*	*t Ratio*
β_G	−0.01550	−3.517	−0.01333	−2.711
β_T	−0.09612	−9.207	−0.13405	−5.216
γ_H	0.01329	1.295	−0.00108	−0.097
α_{air}	5.2074	6.684	6.5940	4.075
α_{train}	3.8690	8.731	3.6190	4.317
α_{bus}	3.1632	7.025	3.3218	3.822
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 MNL 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

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.

After IIA was rejected, the authors estimated a nested logit model of the following type:



Note that one of the branches has only a single choice (this is called a "degenerate" branch), so the conditional probability, $P_{j|fly}$, $= P_{air|fly} = 1$. The estimates 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 full log likelihood for the nested logit model. In this model,

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

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	0.0000194				0.000101			
β_T	−0.0000005	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

$H = 33.3367$. Critical chi-squared[4] $= 9.488$.

TABLE 18.6 Estimates of a Nested Logit Model (standard errors in parentheses)

Parameter	*Nested Logit*		*Multinomial Logit*	
α_{air}	6.0423	(1.1989)	5.2074	(0.7791)
α_{bus}	4.0963	(0.6152)	3.1632	(0.4503)
α_{train}	5.0646	(0.6620)	3.8690	(0.4431)
β_{GC}	−0.0316	(0.0082)	−0.1550	(0.0044)
β_{TTME}	−0.1126	(0.0141)	−0.0961	(0.0104)
γ_H	0.0153	(0.0094)	0.0133	(0.0103)
τ_{fly}	0.5860	(0.1406)	1.0000	(0.0000)
τ_{ground}	0.3890	(0.1237)	1.0000	(0.0000)
σ_{fly}	2.1886	(0.5255)	1.2825	(0.0000)
σ_{ground}	3.2974	(1.0487)	1.2825	(0.0000)
$\ln L$	−193.6561		−199.1284	

The likelihood ratio statistic for the nesting against the null hypothesis of homoscedasticity is $-2[-199.1284 - (-193.6561)] = 10.945$. The 95% 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. The simplest form allows a separate variance, $\sigma_j^2 = \pi^2/(6\theta_j^2)$, for each ε_{ij} in (18-1). (One of the θs 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. But the important degree of freedom is that the HEV model does not impose the IIA assumptions anywhere in the choices, whereas the nested logit does, within each branch. Table 18.7 contains additional results for HEV specifications. In the "Restricted HEV Model," the variance of $\varepsilon_{i,Air}$ is allowed to differ from the others.

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. The multinomial probit model produces free correlations among

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

Parameter	*HEV Model*		*Restricted HEV Model*	
α_{air}	2.228	(1.047)	1.622	(1.247)
α_{train}	3.412	(0.895)	3.942	(0.489)
α_{bus}	3.286	(0.836)	2.866	(0.418)
β_{GC}	−0.026	(0.009)	−0.033	(0.006)
β_{TTME}	−0.071	(0.024)	−0.075	(0.005)
γ	0.028	(0.019)	0.039	(0.021)
θ_{air}	0.472	(0.199)	0.380	(0.095)
θ_{train}	0.886	(0.460)	1.000	(0.000)
θ_{bus}	3.143	(3.551)	1.000	(0.000)
θ_{car}	1.000	(0.000)	1.000	(0.000)
	Implied Standard Deviations			
σ_{air}	2.720	(1.149)		
σ_{train}	1.448	(0.752)		
σ_{bus}	0.408	(0.461)		
σ_{car}	1.283	(0.000)		
ln L	−199.0306		−203.2679	

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	−1.377	0.523	0.523	0.523
Train	0.377	−2.955	1.168	1.168
Bus	0.196	0.604	−3.037	0.604
Car	0.337	1.142	1.142	−1.872
Heteroscedastic Extreme Value				
Air	−1.019	0.410	0.954	0.429
Train	0.395	−3.026	3.184	0.898
Bus	0.282	0.999	−8.161	1.326
Car	0.314	0.708	2.733	−2.589
Multinomial Probit				
Air	−1.092	0.606	0.530	0.290
Train	0.591	−4.078	3.187	1.043
Bus	0.245	1.294	−7.694	1.218
Car	0.255	1.009	2.942	−2.364

TABLE 18.9 Parameter Estimates for Normal-Based Multinomial Choice Models

Parameter	*Multinomial Probit*	*Random Parameters*	
α_{air}	1.799 (1.705)	4.393 (1.698)	
σ_{air}	4.638 (2.251)	4.267 (2.224)	[4.455][a]
α_{train}	4.347 (1.789)	5.649 (1.383	
σ_{train}	1.877 (1.222)	1.097 (1.388)	[1.688][a]
α_{bus}	3.652 (1.421)	4.587 (1.260)	
σ_{bus}	1.000[b]	0.677 (0.958)	[1.450][a]
α_{car}	0.000[b]	0.000[b]	
σ_{car}	1.000[b]	0.000[b]	[1.283][a]
β_G	−0.035 (0.134)	−0.036 (0.014)	
β_T	−0.081 (0.039)	−0.118 (0.022)	
γ_H	0.056 (0.038)	0.047 (0.035)	
ρ_{AT}	0.507 (0.491)	−0.707 (1.268)[c]	
ρ_{AB}	0.457 (0.853)	−0.696 (1.619)[c]	
ρ_{BT}	0.653 (0.346)	−0.014 (2.923)[c]	
ρ_{AC}	0.000[b]	0.000[b]	
ρ_{BC}	0.000[b]	0.000[b]	
ρ_{TC}	0.000[b]	0.000[b]	
$\ln L$	−196.927	−195.646	

[a] Computed as the square root of $(\pi^2/6 + \sigma_j^2)$.
[b] Restricted to this fixed value.
[c] Computed using the delta method.

the choices, which implies an unrestricted 3 × 3 correlation matrix and two free standard deviations.

Table 18.9 reports a variant of the random parameters logit model in which the alternative specific constants are random and freely correlated. The 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 estimates of the specific parameters, θ_j, are given in the table. The estimated model allows unrestricted variation and correlation among the three intercept parameters—this parallels the general specification of the multinomial probit 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. 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 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.

18.2.6 MODELING HETEROGENEITY

Much of the recent development of choice models has been directed toward accommodating individual heterogeneity. We will consider a few of these, including the mixed logit, which has attracted most of the focus of recent research. The mixed logit model is the extension of the random parameters framework of Sections 15.6–15.10 to multinomial choice models. We will also examine the latent class MNL model.

18.2.6.a The Mixed Logit Model

The **random parameters logit model (RPL)** is also called the **mixed logit model**. [See Revelt and Train (1996); Bhat (1996); Berry, Levinsohn, and Pakes (1995); Jain, Vilcassim, and Chintagunta (1994); Hensher and Greene (2010a); and Hensher, Rose and Greene (2015).] Train's (2009) formulation of the RPL model (which encompasses the others) is a modification of the MNL model. The model is a **random coefficients** formulation. The change to the basic MNL model is the parameter specification in the distribution of the parameters across individuals, i,

$$\beta_{ik} = \beta_k + \mathbf{z}_i'\boldsymbol{\theta}_k + \sigma_k u_{ik}, \tag{18-13}$$

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). (Most of the received applications have $\boldsymbol{\theta}_k = \mathbf{0}$ – that is, homogeneous means of the random parameters.

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,

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

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

$$E_{\boldsymbol{u}}[\text{Prob}(\text{choice } j|\mathbf{u})] = \int_{u_1, \ldots, u_k} \text{Prob}[\text{choice } j|\mathbf{u}]f(\mathbf{u})d\mathbf{u},$$

which can be estimated by simulation, using

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

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 his or her 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 18.2.8),

one can formulate a random effects model simply by making the variation in the coefficients time invariant. Thus, the model is changed to

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

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.[12] A second degree of flexibility is in (18-13). 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 (such as based on the lognormal distribution). We will make use of both of these features in the application in Example 18.8.

18.2.6.b 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 (18-10). 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.]

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, serves this purpose. In this case, the observed heterogeneity enters the deterministic part of the utility functions. The heteroscedastic HEV model shown in (18-10) moves the observable heterogeneity to the scaling of the utility function instead of the mean. The mixed logit model in (18-13) accommodates both observed and unobserved heterogeneity in the preference parameters. A recent thread of research including Keane (2006), Feibig, et al. (2009), and Greene and Hensher (2010a) has considered functional forms that accommodate individual heterogeneity in both taste parameters (marginal utilities) and overall scaling of the preference structure. Feibig et al.'s **generalized mixed logit model** is

$$U_{i,j} = \mathbf{x}_{ij}'\boldsymbol{\beta}_i + \varepsilon_{ij},$$
$$\boldsymbol{\beta}_i = \sigma_i\boldsymbol{\beta} + [\gamma + \sigma_i(1 - \gamma)]\mathbf{u}_i$$
$$\sigma_i = \exp[\overline{\sigma} + \tau w_i],$$

where $0 \leq \gamma \leq 1$ and w_i is an additional source of unobserved random variation in preferences along with $\mathbf{u}_i$. In this formulation, the weighting parameter, γ, distributes the

[12] A stated choice experiment in which consumers make several choices in sequence about automobile features appears in Hensher, Rose, and Greene (2015).

individual heterogeneity in the preference weights, $\mathbf{u}_i$, and the overall scaling parameter, σ_i. Heterogeneity across individuals in the overall scaling of preference structures is introduced by a nonzero τ while $\overline{\sigma}$ is chosen so that $E_w[\sigma_i] = 1$. Greene and Hensher (2010a) 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 $\mathbf{\Gamma}$) produces a multilayered form of the generalized mixed logit model,

$$\boldsymbol{\beta}_i = \sigma_i[\boldsymbol{\beta} + \mathbf{\Delta}\mathbf{z}_i] + [\gamma + \sigma_i(1 - \gamma)]\mathbf{\Gamma}\mathbf{u}_i$$
$$\sigma_i = \exp[\overline{\sigma} + \boldsymbol{\delta}'\mathbf{h}_i + \tau w_i].$$

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

Example 18.4 Using Mixed Logit to Evaluate a Rebate Program

In 2005, Australia led OECD countries and most of the world in per capita greenhouse gas emissions. Among the many federal and state programs aimed at promoting energy efficiency was a water heater rebate program for the New South Wales residential sector. Wasi and Carson (2013) sought to evaluate the impact of the program on Sydney area homeowners' demand for efficient water heaters. The study assessed the effect of the rebate program in shifting existing stocks of electric (primarily coal generated) heaters toward more climate-friendly technologies. Two studies were undertaken: a "revealed preference" (RP) analysis of choices made by recent purchasers of new water heaters and a "stated preference" (SP) study of households that had not replaced their water heaters in the past ten years (and were likely to be in the market in the near future). Broad conclusions drawn from the study included:

> *Our results suggest that households who do not have access to natural gas are more responsive to the rebate program.Without incentive, these households are more likely to replace their electric heater with another electric heater. For those with access to natural gas, many of them would have chosen to replace their electric heater with a gas heater even if the rebate programs had not been in place. These findings are consistent in both ex-post and ex-ante evaluation. From actual purchase data, we also find that the rebate programs appear to work largely on households that deliberately set out to replace their water heater rather than on households that replaced their water heater on an emergency/urgent basis.* (p. 646.)

Data for the study were obtained through a web-based panel by a major survey research firm. A total of 3,322 respondents out of 9,400 invitees were interested in participating. Access to natural gas is a key determinant of the technology choices that households make. The RP (ex-post) sample included 408 with gas access and 504 without; the SP (ex-ante) sample included 547 with access and 354 without.

Modeling the RP respondents was complicated by the fact that many did not remember the available choice set or could not accurately provide data for the installation cost and running cost. The authors opted for a difference in differences approach based on a simple logit model, as shown in Table 18.10 (which is extracted from their Table 3).[13] (Results are based on a binary logit model for households with no gas access and trinomial logit for those with gas access.)

The SP choice model was based on a mixed logit framework: Attributes of the choices included setup cost net of the rebate, running cost, and a dummy variable for a mail-in rebate.

[13]Wasi and Carson (2013).

TABLE 18.10 Results from Table 3

Estimated Policy Effects on Probability of Switching from Electric for Households with Gas Access

Probability of Switching to	Before Policy	After Policy	Change in Shares
Electric	0.28**	0.19**	−0.09
Gas	0.69**	0.55**	−0.14**
Solar/Heat Pump	0.03**	0.26**	0.23**
Probability of Switching to	Before Policy 2004-2005	2006-Sep 2007	Change in Shares
Electric	0.39**	0.22**	−0.17*
Gas	0.61**	0.74**	0.13
Solar/Heat Pump	0.00	0.04*	0.04*
Effects of Policy on Probability of Switching to			Difference of Changes in Shares
Electric			0.08
Gas			−0.27**
Solar/Heat Pump			0.19**

**,* = Statistically significant at 1%, 5%, respectively.

TABLE 18.11 Results from Table 6

Estimated SP Choice Models

	MNL	***GMNL***		***MM-MNL***			
				Class 1		***Class 2***	
		Mean	StdDev	Mean	StdDev	Mean	StdDev
Cost after rebate/10000	−8.62**	−27.13**	12.53**	−27.3**	14.66**	−16.93**	12.9**
1 if mail-in rebate	0.002	0.01	0.61**	0.01	0.07	−0.28	1.33**
Annual running cost/1000	−3.99**	−17.66**	9.21**	−22.02**	15.42**	−9.35**	6.94**
Class probability				0.66**		0.34**	
τ		0.75**					
γ		−0.81					

**,* = Statistically significant at 1%, 5%, respectively.

The choice experiment included 16 repetitions. The choice set for new installations included electric, gas storage, gas instantaneous, solar, and heat pump. A variety of models were considered: multinomial logit (MNL), mixed logit (MXL), generalized mixed logit (GMXL), latent class logit (LCM), and a mixture of two normals (MM), which is a latent class model in which each class is defined by a mixed logit model. Based on the BIC values, it was determined that the GMXL and MM models were preferred. Some of the results are shown in Table 18.11, which is extracted from their Table 6.

Column 1 of Table 18.11 reports the estimates from the MNL model for the gas access sample.[14] The two cost variables have negative coefficients as expected. The coefficient of

[14]Ibid.

the rebate dummy is positive but not statistically different from zero. The coefficient is large and negative in one of the two classes, suggesting that in this segment, there is substantial disutility attached to filing for the rebate. The average WTP for $1 saved annually is $-3.99 \times 10/-8.62 = 4.62$. Assuming the durability of 15 years, this implies a discount rate of 20%. Column 2 presents the result from the GMNL (generalized mixed logit) model using the full covariance matrix version. The average WTP for $1 saved annually from this model is $6.55, implying a discount rate of 12.8%. Policy evaluations were carried out by simulating the market shares of the different water heater technologies and evaluating the implied impacts on emissions. For households with gas access, the share of electric and gas heaters would reduce by 8% and 11%, respectively. The share of solar/heat pump would increase by 19%. Households with no access to natural gas, while still possessing more electric heaters, are more responsive to the rebate policy (38% reduction in the share of electric heaters). The final step is the evaluation of the cost of the rebate for emission reduction. It was determined that the average costs of carbon reduction from the SP data are $254/ton using a gas access sample and $105/ton from a sample with no access to natural gas. These values were significantly higher than U.S results ($47/ton) but similar to other results from Mexico. Notably, they are much larger than provided for by the NSW climate change fund ($26/ton).

18.2.6.c Latent Classes

We examined the latent class model in Sections 14.15 and 17.7.6. The framework has been used in a number of choice experiments to model heterogeneity semiparametrically. The base framework is

$$\begin{aligned}\text{Prob}(\text{choice}_{it} = j|\mathbf{X}_{it}, \text{class} = c) &= \frac{\exp(\mathbf{x}_{ijt}'\boldsymbol{\beta}_c)}{\Sigma_{m=1}^{J}\exp(\mathbf{x}_{imt}'\boldsymbol{\beta}_c)},\\ \text{Prob}(\text{class} = c) &= \pi_{\mathbf{c}}, c = 1, \ldots, C.\end{aligned}$$

The latent class model can usefully be cast as a random parameters specification in which the support of the parameter space is a finite set of points. By this hierarchical structure, the parameter vector, $\boldsymbol{\beta}$, has a discrete distribution, such that

$$\text{Prob}(\boldsymbol{\beta}_i = \boldsymbol{\beta}_c) = \pi_c, 0 \leq \pi_{\mathbf{c}} \leq 1, \Sigma_c\pi_c = 1.$$

The unconditional choice probability is

$$\text{Prob}(\text{choice}_{it} = j|\mathbf{X}_{it}) = \sum_{c=1}^{C}\pi_c\frac{\exp(\mathbf{x}_{ijt}'\beta_c)}{\Sigma_{m=1}^{J}\exp(\mathbf{x}_{imt}'\beta_c)}.$$

Wasi and Carson (2013), in Example 18.4, settled on a latent class specification in which each class defined a mixed logit model. (In Wasi and Carson's specification, $\boldsymbol{\beta}_{i|c} \sim \text{N}[\boldsymbol{\beta}_c, \boldsymbol{\Sigma}_c]$.)

Example 18.5 Latent Class Analysis of the Demand for Green Energy

Ndebele and Marsh (2014) examined preferences for Green Energy among electricity consumers in New Zealand. The study was motivated by a New Zealand study by the Electricity Commission (2008) that reported that nearly 50% of respondents indicated that they would consider the environment when choosing an electricity retailer whilst 17% indicated they would "very seriously" consider switching to a retailer which promotes itself for using renewable resources.

Ndebele and Marsh used a latent class choice modeling framework in which the integration of Environmental Attitude (EA) with stated choices is either direct via the utility function as interactions with the attribute levels of alternatives or as a variable in the class membership

probability model. They identified three latent classes with different preferences for the attributes of electricity suppliers. A typical respondent with a high New Ecological Paradigm (NEP) scale score is willing to pay on average $12.80 more per month on his or her power bill to secure a 10% increase in electricity generated from renewable energy sources compared to respondents with low NEP scores.

An online survey questionnaire was developed to collect the data required for this research. The first part of the survey questionnaire elicited socio-demographic and EA. EA was measured using the 15 items of the NEP scale. The NEP scale is a measure of environmental attitude.[15] The NEP scale is a five-point Likert-type scale consisting of 15 items or statements about the human-environment relationship. The design for the SP experiment is shown in Table 18.12, which is extracted from their Table 2.[16]

An online survey was administered by a market research company in January 2014 to a sample of 224 New Zealand residential electricity bill payers. Stratification was based on age group, gender, and income group. The NEP scores were obtained through online interview. As part of the debriefing, respondents were asked to state the attributes they ignored in choosing their preferred supplier. Attitudinal questions also included questions measuring *awareness of the consequences (AC)* of switching to a supplier that generates most of its electricity from renewables and how far they felt personally responsible—that is, ascription of responsibility (AR)—for reducing CO_2 emissions by switching to a supplier that generates electricity from renewable energy sources. The authors report that "[t]o account for attribute non-attendance in model estimation we coded our data to reflect stated serial non-attendance to specific attributes." Attribute nonattendance is examined in Section 18.2.6d and Example 18.6.

Estimated models are shown in Table 18.13, which is extracted from their Table 13. Based on the MNL model, consumers with moderate NEP scale scores are willing to pay ($10 × 0.0066/0.0255) ≈ $2.60 more per month to secure a 10% increase in electricity generated from renewable sources compared to consumers with a low NEP scale score or low EA. Consumers with strong EA (high NEP scale score) are willing to pay ($10 × 0.0105/0.0255) ≈ $4.10 more per month to secure a 10% increase in electricity generated from renewables compared with customers with low EA. A supplier that is offering a 10% higher prompt payment discount may charge $3.80 more per month than other suppliers *ceteris paribus* and still retain its customers.

TABLE 18.12 Experimental Design: Attributes in Stated Choice Experiment

Attribute	*Description*
Time	= Average wait time for customer service calls (minutes)
Fixed	= Amount of time prices are guaranteed (months)
Discount	= Percent discount for paying bills on time
Rewards	= Presence of a loyalty program (yes/no)
Renewable	= Proportion of electricity generated by green technologies
Ownership	= Proportion of supplier New Zealand owned
Supplier Type	= New or well known company (yes/no)
Bill	= Average monthly bill

[15]See (Dunlap (2008) and Hawcroft and Milfont (2010).

[16]From Ndebele and Marsh (2014).

TABLE 18.13 Estimated Models

Selected Estimates of MNL and Latent Class Model Parameters

Variables	*MNL*	*Latent Class*		
		Class 1	Class 2	Class 3
ASC_{QC}	0.5766***	0.5213***	0.0953	3.2544***
Time (Minutes)	−0.0430***	−0.0378***	−0.0340***	−0.0420
Fixed Term (Months)	0.0046**	0.0057	0.0103**	−0.0033
Discount	0.0096***	0.0054	0.0157***	0.0516***
Loyalty Rewards	0.3691***	0.2698*	0.3607***	0.4891
%Renewable	0.0031	0.0019	0.0079	−0.0042
MNEP × Renewable	0.0066**	0.0075	0.0056	0.0230*
SNEP × Renewable	0.0105***	0.0145*	0.0099**	−0.0003
%NZ Ownership	0.0082***	0.0135***	0.0122***	0.0057
Monthly Power Bill	−0.0255***	−0.0572***	−0.0139***	−0.0147***
Class Probability		0.5374***	0.3479***	0.1147***
Log Likelihood	−2153.4	−1748.41		

*,**, *** Significant at 0.10, 0.05, 0.01, respectively.

18.2.6.d Attribute Nonattendance

In the choice model,

$$U_{ijt} = \alpha_j + \beta_1 x_{ijt,1} + \beta_2 x_{ijt,2} + \ \ldots \ + \varepsilon_{ijt},$$

and the familiar multinomial logit probability, the presence of a nonzero part worth (β) on attribute k suggests a nonzero marginal utility (or disutility) of that attribute for individual i. One possible misspecification of the model would be an assumption of homogeneous attendance. In a given population, one form of heterogeneity might be attribute nonattendance for some (or all) of the attributes.[17] **Attribute nonattendance** (ANA) can represent a rational result of zero marginal utility or it can result from a deliberate strategy to simplify the choice process. These outcomes might be directly observable in a choice experiment in which respondents are specifically queried about them. In Example 18.5, we noted that Ndebele and Marsh solicited this information in the debriefing interview. Nonattendance might only be indirectly observable by behavior that seems to suggest its presence. Consider, for example, a stated choice experiment in which large variation in an attribute such as price appears not to induce switching behavior.

Attribute nonattendance represents a form of individual heterogeneity. Consider the utility function suggested above, which suggests full attendance of both attributes. In a heterogeneous population, there could be (at least) four types of individuals

$$\begin{aligned}
&\text{(Type 1, 2)} \quad U_{ijt} = \alpha_j + \beta_1 x_{ijt,1} + \beta_2 x_{ijt,2} + \ \ldots \ + \varepsilon_{ijt,}\\
&\text{(Type 0, 2)} \quad U_{ijt} = \alpha_j + 0 \qquad + \beta_2 x_{ijt,2} + \ \ldots \ + \varepsilon_{ijt,}\\
&\text{(Type 1, 0)} \quad U_{ijt} = \alpha_j + \beta_1 x_{ijt,1} + \quad 0 \quad + \ \ldots \ + \varepsilon_{ijt,}\\
&\text{(Type 0, 0)} \quad U_{ijt} = \alpha_j + 0 \qquad + \quad 0 \quad + \ \ldots \ + \varepsilon_{ijt.}
\end{aligned}$$

[17]See, for example, Alemu et al. (2013), Hensher, Rose, and Greene (2005, 2012), Hensher and Greene (2010), Hess and Hensher (2012), Hole (2011), and Scarpa, Thiene, and Hensher (2010). The first of these is an extensive survey of the subject.

If the partitioning of the population is observed—Ndebele and Marsh note "we coded our data to reflect stated serial non-attendance to specific attributes"—then the appropriate estimation strategy is to impose the implied zero constraints on $\boldsymbol{\beta}$ selectively, observation by observation. The indicator of which attributes are nonttended by each individual, d_{Type}, becomes part of the "coding" of the data. The log likelihood to be maximized would be

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\Bigg[d_{i,Type1,2} \ln L_i\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} + d_{i,Type0,2} \ln L_i\begin{pmatrix}0\\ \beta_2\end{pmatrix} + d_{i,Type1,0} \ln L_i\begin{pmatrix}\beta_1\\ 0\end{pmatrix} + d_{i,Type0,0} \ln L_i\begin{pmatrix}0\\ 0\end{pmatrix}\Bigg].$$

(Only one of the indicators, $d_{i,Type}$, equals one.)

One framework for analyzing attribute nonattendance when it is only indirectly observed is a form of latent class model. If the analyst has not directly observed the types, then this suggests a latent class approach to modeling attribute nonattendance. In the model above, this case is simply a missing data application. Since d_{Type} is unobserved, it is replaced in the log likelihood with the probabilities, π_{Type} (which are to be estimated as well) and the model becomes a familiar latent class model,

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{i=1}^{n}\Bigg[\pi_{Type1,2} \ln L_i\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} + \pi_{Type0,2} \ln L_i\begin{pmatrix}0\\ \beta_2\end{pmatrix} + \pi_{Type1,0} \ln L_i\begin{pmatrix}\beta_1\\ 0\end{pmatrix} + \pi_{Type0,0} \ln L_i\begin{pmatrix}0\\ 0\end{pmatrix}\Bigg].$$

For the example above, the latent class structure would have four classes. For reasons apparent in the listing above, Hensher and Greene (2010) label this the "2^K model." Note that the implied latent class model has two types of restrictions. There is only a single parameter vector in the model — there are cross-class restrictions on the parameters — and there are fixed zeros at different positions in the parameter vector.[18] We will examine an application in Example 18.6.

Example 18.6 Malaria Control During Pregnancy

Lagarde (2013) used the 2^K approach to model attribute nonattendance in a choice experiment about adoption of guidelines for malaria control during pregnancy. The discrete choice experiment was administered to health care providers in Ghana to evaluate their potential resistance to changes in clinical guidelines. The choice task involved whether or not to accept a new set of clinical guidelines. Results showed that less than 3% of the respondents considered all six attributes when choosing between the two hypothetical scenarios proposed, with a majority looking at only one or two attributes. Accounting for ANA strategies affected the magnitude of some of the coefficients and willingness-to-pay estimates.

Guidelines involved six attributes, hence 64 combinations of attendance: The attributes were

1. **Approach:** preventive or curative,
2. **Antimalarial drugs:** SP (Fansidar) or SS-AQ Artesunate-amodiaquine,
3. **Prevalence of anemia for mothers treated with protocol:** 1% or 15%,

[18] A natural extension would be to relax the restriction of equal coefficients across the classes. This is testable.

4. **Prevalence of low birth weight among infants of mothers treated:** 10% or 15%,
5. **Staffing level for the SN clinic:** Under-staffed or adequately staffed,
6. **Salary supplement included in the protocol:** GH. C10, GH. C20.

The author devised a stepwise simplification in the estimation strategy to allow analysis of the excessively large number of classes (64) in the base case model. Accounting for ANA produced fairly large changes in model estimates and estimates of WTP. For examples the estimated coefficients on Anemia Risk and Treatment changed from −0.127 (0.086) to −0.214 (0.016) and from −0.096 (0.077) to −1.840 (0.540). The main results suggested that WTP measures were very sensitive to the presence of ANA. The estimated WTP for the SP drug rose from 8.75 to 24.59 when ANA was considered.[19]

18.2.7 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

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

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

$$\widehat{\text{WTP}}_{time} = -\hat{\beta}_{TIME}/\hat{\gamma}_{Income}.$$

(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 4.6 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 $-(-0.09612)/0.01329 = 7.23$. The units of measurement must be resolved in this computation, since terminal time is measured in minutes while income is in \$1,000/year. Multiplying this result by 60 minutes/hour and dividing by the equivalent hourly income times 8,760/1,000 gives \$49.52 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

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

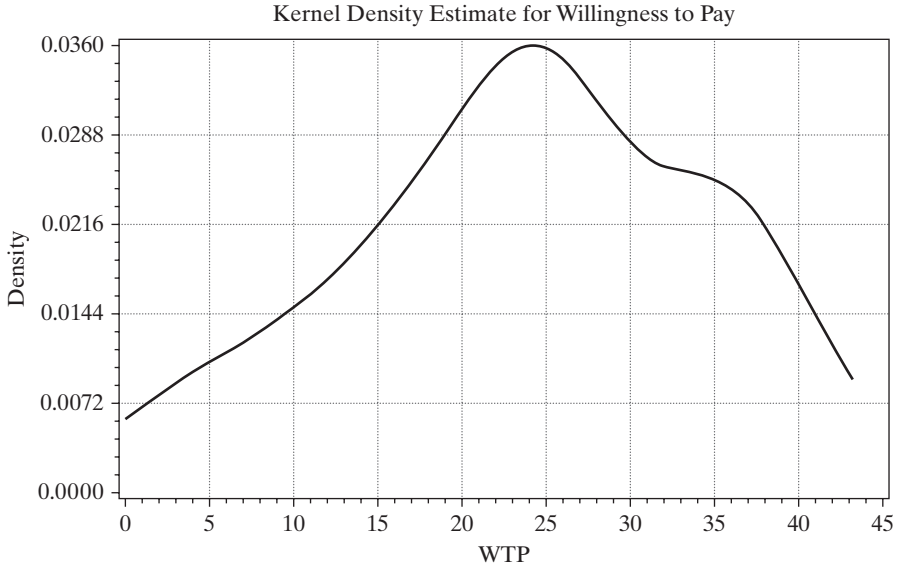
[19]Figures from Lagarde (2013) Tables IV and V.

The derivatives of $\widehat{\text{WTP}}_{TIME} = -\hat{\beta}_{TIME}/\hat{\gamma}_{HINC}$ are $-1/\hat{\gamma}_{HINC}$ for $\hat{\beta}_{TTME}$ and $-\widehat{\text{WTP}}/\hat{\gamma}_{HINC}$ for $\hat{\gamma}_{HINC}$. This provides an estimator of 38.8304 for the standard error. The confidence interval for this parameter would be −26.55 to +125.66. 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. (See Example 18.8.) Using our mode choice application, we refit the model with $\hat{\beta}_{TTME,i} = \hat{\beta}_{TTME} + \hat{\sigma}_{TTME}v_i$ and all other coefficients nonrandom. We then used the method described in Section 15.10 to estimate the mixed logit model and $\text{E}[\hat{\beta}_{TTME,i}|\mathbf{X}_i, choice_i]/\hat{\gamma}_H$ 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. 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 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

FIGURE 18.1 Estimated Willingness to Pay for Decreased Terminal Time.



the multinomial logit case, this amounts to a trivial reparameterization of the model. Using our application as an example, we would write

$$\begin{aligned} U_{ij} &= \alpha_j + \beta_{GC} GC_i + \gamma_{HINC}[(\beta_{TTME}/\gamma_{HINC})TTME_i + (A_{AIR} HINC_i)] + \varepsilon_{ij} \\ &= \alpha_j + \beta_{GC} GC_i + \gamma_{HINC}[\lambda_{TTME} TTME_i + (A_{AIR} HINC_i)] + \varepsilon_{ij}. \end{aligned}$$

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; $-\hat{\lambda}_{TTME}$ 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 $\hat{\lambda}_{TTME,i} = \hat{\lambda}_{TTME} + \hat{\theta}_{TTME} v_{TTME,i}$. Greene and Hensher (2010b) apply this method to the generalized mixed logit model in Section 18.2.8.

Example 18.7 Willingness to Pay for Renewable Energy

Scarpa and Willis (2010) examined the willingness to pay for renewable energy in the UK with a stated choice experiment. A sample of 1,279 UK households were interviewed about their preferences for heating systems. One analysis in the study considered answers to the following question:

> *"Please imagine that your current heating system needs replacement. I would like you to think about some alternative heating systems for your home. All of the following systems would fully replace your current system. For example, if you had a gas boiler, it would be taken out and replaced by the new system. The rest of your heating system, such as the radiators, would not need to be changed."*

This *primary* experiment included alternative systems such as biomass boilers and supplementary heat pumps with their associated attributes (with space requirements for fuel storage and hot water storage tanks), compared to combi-gas boilers, which deliver central heating and hot water on-demand without the need for hot water storage or fuel storage or the inconvenience associated with tending solid fuel boilers. Notably, in this experiment, the authors did not suggest an opt-out choice. The experiment assumed that the heating system had failed and needed to be replaced. A second experiment, the one discussed below, was based on the *discretionary* case, *"Now I would like you to imagine that your current heating system is functioning completely normally, and to think about supplementing your existing system with an additional system."*

Respondents were asked to choose the type of heating system they would prefer between two alternatives, in four different scenarios. Results for multinomial logit models estimated in preference space and WTP space are shown in Table 18.14 in the results extracted from their Table 5.[20] In addition to the MNL models, they estimated a nested logit model (not shown) and a mixed logit model in WTP space. (We will examine a stated choice experiment based on a mixed logit model in the next application.) Note the two MNL models produce the same log likelihood and related statistics. This is a result of the fact that the WTP space model is a 1:1 transformation of the preference space model. (This is an application of the invariance principle in Section 14.4.5.d.) We can deduce the second model from the first. For example, the numeraire coefficient is the capital cost, equal to −0.3288. Thus, in the WTP space model, the coefficient on solar energy is 0.9312/0.3288 = 2.8316. The coefficient on energy savings is 0.0973/0.3288 = 0.2957 (plus some rounding error) and likewise for the other coefficients in the WTP space model. (This leaves a loose end. The coefficient on capital costs should

[20]Scarpa and Willis (2009).

TABLE 18.14 Estimated Models

Estimated Multiomial Logit Models (1,241 Individuals, 7,280 observations)

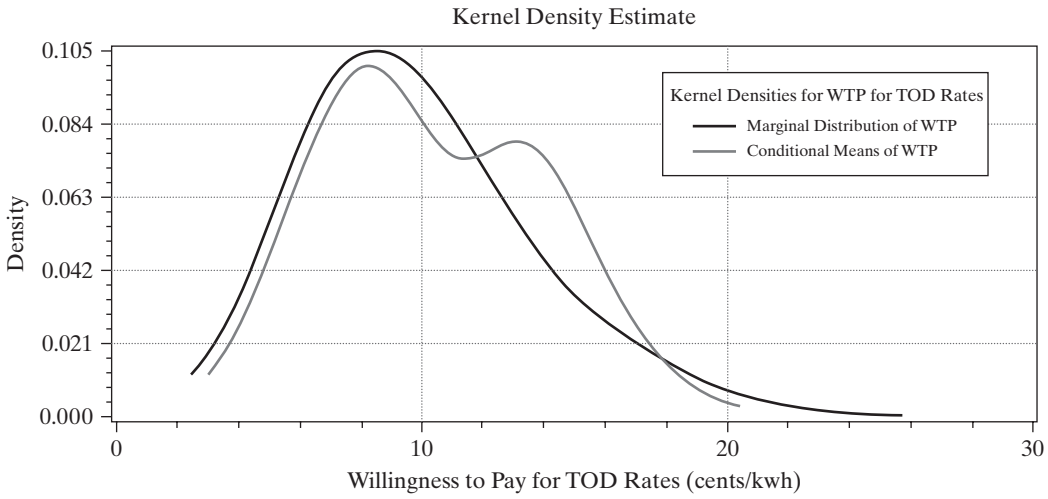
	MNL Preference Space		*MNL WTP-Space*	
	Coefficient	$\|t\|$	***Coefficient***	***Std. Error***
Solar electricity	0.9312	11.01	2.8316	0.2441
Solar hot water	0.9547	10.84	2.90322	0.2555
Wind turbine	0.4236	5.15	1.2882	0.2408
Capital cost/mean $\ln(\lambda)$	−0.3288	24.13	−1.1122	0.0415
Friend	−0.0698	1.31	−0.2120	0.1627
Heating engineer	0.0864	1.43	0.2626	0.1834
Both	0.1820	3.52	0.5534	0.1575
Maintenance cost	−0.0303	5.08	−0.0922	0.0184
Energy savings	0.0973	5.20	0.2957	0.0590
Log likelihood	−7328.88		−7328.88	
Rho-square	0.08091		0.08091	

be 1.0000. The authors do not make clear where the 1.1122 comes from.) By adjusting for the units of measurement, the 2.3816 for solar energy translates to a value of 2381.6 GBP. The average installation costs for a 2 kWh solar PV unit in 2008 was 10,638 GBP, 3,904 GBP for a 2 kWh solar hot water unit, and 4,998 GBP for a 1 kWh micro-wind unit. The implied WTP values from the model in Table 5 are 2,381 GBP, 2,903 GBP and 1,288 GBP, respectively. The estimates from the CE data also permitted the evaluation of the relative importance consumers attached to capital in relation to ongoing energy savings. Consumers were WTP 2.91 $\pm$ 0.30 GBP in capital costs to reduce annual fuel bills by 1 GBP. The authors conclude that "whilst renewable energy adoption is significantly valued by households, this value is not sufficiently large, for the vast majority of households, to cover the higher capital costs of micro-generation energy technologies, and in relation of annual savings in energy running costs." (p. 135)

18.2.8 PANEL DATA AND STATED CHOICE EXPERIMENTS

The counterpart to panel data in the multinomial choice context is usually the "stated choice experiment," such as the study discussed in Example 18.7. In a stated choice experiment, the analyst (typically) hypothesizes several variations on a general scenario and requests the respondent's preferences among several alternatives each time. In Example 18.8, the sampled individuals are offered a choice of four different electricity suppliers. Each alternative supplier is a specific bundle of rate structure types, contract length, familiarity, and other attributes. The respondent is presented with from 8 to 12 such scenarios, and makes a choice each time. The panel data aspect of this setup is that the same individual makes the choice each time. Any chooser-specific feature, including the underlying preference, is repeated and carried across from scenario to scenario. The MNL model (whether analyzed in preference or WTP space) does not explicitly account for the common underlying characteristics of the individual. The analogous case in the regression and binary choice cases we have already examined would be the pooled model. Several modeling approaches have been used to accommodate the underlying individual heterogeneity in the choice model. The mixed logit model is the most common. Note the third set of results in Figure 18.2 is based on a mixed logit model,

FIGURE 18.2 WTP for Time of Day Rates.



$$\text{Prob}(choice_{it} = j|\mathbf{X}_{it}) = \frac{\exp(\mathbf{x}_{ijt}'\boldsymbol{\beta}_i)}{\sum_{m=1}^{J}\exp(\mathbf{x}_{imt}'\boldsymbol{\beta}_i)}, \boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i; i = 1, \ldots, n; t = 1, \ldots, T_i.$$

The random elements in the coefficients are analogous to random effects in the settings we have already examined.

18.2.8.a The Mixed Logit Model

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. Rossi and Allenby (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 scaling matrix, $\boldsymbol{\Gamma}$, allows the random elements of $\boldsymbol{\beta}$ to be correlated; a diagonal $\boldsymbol{\Gamma}$ returns the more familiar case.

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 (2015) 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 proposed 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) suggested a nested logit model that groups the hypothetical choices in one branch of a tree and the observed choices in another.

18.2.8.b Random Effects and the Nested Logit Model

The mixed logit model in a stated choice experiment setting can be restricted to produce a random effects model. Consider the four-choice example below. The corresponding formulation would be

$$\begin{aligned}
U_{i1,t} &= (\alpha_1 + u_{i1}) + \mathbf{x}'_{i1,t}\boldsymbol{\beta} + \varepsilon_{i1,t},\\
U_{i2,t} &= (\alpha_2 + u_{i2}) + \mathbf{x}'_{i2,t}\boldsymbol{\beta} + \varepsilon_{i2,t},\\
U_{i3,t} &= (\alpha_3 + u_{i3}) + \mathbf{x}'_{i3,t}\boldsymbol{\beta} + \varepsilon_{i3,t},\\
U_{i4,t} &= \qquad\qquad\quad \mathbf{x}'_{i4,t}\boldsymbol{\beta} + \varepsilon_{i4,t}.
\end{aligned}$$

This is simply a restricted version of the random parameters model in which the constant terms are the random parameters. This formulation also provides a way to specify the nested logit model by imposing a further restriction. For example, the nested logit model in the mode choice in Example 18.3 results from an error components model,

$$\begin{aligned}
U_{i,air} &= \qquad u_{i,fly} \qquad + \mathbf{x}'_{i,air}\boldsymbol{\beta} + \varepsilon_{i,air},\\
U_{i,train} &= (\alpha_{train} + u_{i,ground}) + \mathbf{x}'_{i,train}\boldsymbol{\beta} + \varepsilon_{i,train},\\
U_{i,bus} &= (\alpha_{bus} + u_{i,ground}) \;+ \mathbf{x}'_{i,bus}\boldsymbol{\beta} + \varepsilon_{i,bus},\\
U_{i,car} &= (\alpha_{car} + u_{i,ground}) \;+ \mathbf{x}'_{i,car}\boldsymbol{\beta} + \varepsilon_{i,car}.
\end{aligned}$$

This is the model suggested after (18-12). The implied covariance matrix for the four utility functions would be

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_F^2 & 0 & 0 & 0 \\ 0 & \sigma_G^2 & \sigma_G^2\rho & \sigma_G^2\rho \\ 0 & \sigma_G^2\rho & \sigma_G^2 & \sigma_G^2\rho \\ 0 & \sigma_G^2\rho & \sigma_G^2\rho & \sigma_G^2 \end{bmatrix}.$$

FIML estimates of the nested logit model from Table 18.6 in Example 18.3 are reported in Table 18.15 below. We have refit the model as an error components model with the two components shown above. This is a model with random constant terms. The estimated

parameters in Table 18.15 are similar as would be expected. The estimated standard deviations for the FIML estimated model are 2.1886 and 3.2974 for *Fly* and *Ground*, respectively. For the random parameters model, we would calculate these using $v = (\pi^2/6 + \sigma_b^2)^{1/2} = 3.48$ for *Fly* and 1.3899 for *Ground*. The similarity of the results carries over to the estimated elasticities, some of which are shown in Table 18.16.

18.2.8.c A Fixed Effects Multinomial Logit Model

A fixed effects multinomial logit model can be formulated as

$$\text{Prob}(y_{it} = j) = \frac{\exp(\alpha_{ij} + \mathbf{x}_{it,j}'\boldsymbol{\beta})}{\sum_{m=1}^{J}\exp(\alpha_{im} + \mathbf{x}_{it,m}'\boldsymbol{\beta})}.$$

Because the probabilities are based on comparisons, one of the utility functions must be normalized at zero. We take that to be the last (Jth) alternative, so the normalized model is

$$\text{Prob}(y_{it} = j) = \frac{\exp(\alpha_{ij} + \mathbf{x}_{it,j}'\boldsymbol{\beta})}{1 + \sum_{m=1}^{J-1}\exp(\alpha_{im} + \mathbf{x}_{it,m}'\boldsymbol{\beta})}, j = 1, \ldots, J-1.$$

We examined the binary logit model with fixed effects in Section 17.7.3. The model here is a direct extension. The Rasch/Chamberlain method for the fixed effects logit model can be used, in principle, for this multinomial logit case. [Chamberlain (1980) mentions this possibility briefly.] However, the amount of computation involved in doing so increases vastly with J. Part of the complexity stems from the difficulty of constructing

TABLE 18.15 Estimated Nested Logit Models

	FIML Nested Logit		*Mixed Logit*	
	Estimate	***Std. Error***	***Estimate***	***Std. Error***
Air	6.04234	(1.19888)	4.65134	(1.26475)
Train	5.06460	(0.66202)	5.13427	(0.67043)
Bus	4.09632	(0.61516)	4.15790	(0.62631)
GC	−0.03159	(0.00816)	−0.03228	(0.00689)
TTME	−0.11262	(0.01413)	−0.11423	(0.01183)
HINC	0.02616	(0.01761)	0.03571	(0.02468)
Fly	0.58601	(0.14062)	3.24032	(1.71679)
Ground	0.38896	(0.12367)	0.53580	(10.65887)
ln L	−193.65615		−195.72711	

TABLE 18.16 Elasticities with Respect to Generalized Cost

	AIR		***TRAIN***		***BUS***		***CAR***	
	NL	***MXL***	***NL***	***MXL***	***NL***	***MXL***	***NL***	***MXL***
AIR	−1.3772	−1.1551	0.5228	0.4358	0.5228	0.4358	0.5228	0.4358
TRAIN	0.3775	0.4906	−2.9452	−3.0467	1.1675	1.1562	1.1675	1.1562
BUS	0.1958	0.2502	0.6039	0.5982	−3.0368	−3.1223	0.6039	0.5982
CAR	0.3372	0.3879	1.1424	1.1236	1.1424	1.1236	−1.8715	−1.9564

the denominator of the conditional probability. The terms in the sum are the different ways that the sequence of $J \times T$ outcomes can sum to T *including the constraint that within each block of J, the outcomes sum to one*. The amount of computation is potentially prohibitive. For our example below, with $J = 4$ and $T = 12$, the number of terms is roughly 6×10^{10}. The Krailo and Pike algorithm is less useful here due to the need to impose the constraint that only one choice be made in each period. However, there is a much simpler approach available based on the minimum distance principle that uses the same information.[21] (See Section 13.3.) For each of outcomes 1 to $J - 1$, the choice between observation j and the numeraire, alternative J, produces a fixed effects binary logit. For each of the $J - 1$ outcomes, then, the $\Sigma_{i=1}^{n} T_i$ observations that chose either outcome j or outcome J can be used to fit a binary logit model to estimate $\boldsymbol{\beta}$. This produces $J - 1$ estimates, $\hat{\boldsymbol{\beta}}_j$, each with estimated asymptotic covariance matrix $\mathbf{V}_j$. The minimum distance estimator of the single $\boldsymbol{\beta}$ would then be

$$\hat{\boldsymbol{\beta}} = \left[\sum_{j=1}^{J-1} \mathbf{V}_j^{-1}\right]^{-1} \sum_{j=1}^{J-1} (\mathbf{V}_j^{-1} \hat{\boldsymbol{\beta}}_j).$$

The estimated asymptotic covariance matrix would be the first term. Each of the binary logit estimates and the averaging at the last step require an insignificant amount of computation.

It does remain true that, like the binary choice estimator, the post-estimation analysis is severely limited because the fixed effects are not actually estimated. It is not possible to compute probabilities and partial effects, etc.

Example 18.8 Stated Choice Experiment: Preference for Electricity Supplier

Revelt and Train (2000) studied the preferences for different prices of a sample of California electricity customers.[22] The authors were particularly interested in individual heterogeneity and used a mixed logit approach. The choice experiment examines the choices among electricity suppliers in which a supplier is defined by a set of attributes. The choice model is based on

$$U_{ijt} = \beta_1 PRICE_{ijt} + \beta_2 TOD_{ijt} + \beta_3 SEAS_{ijt} + \beta_4 CNTL_{ijt} + \beta_5 LOCAL_{ijt} + \beta_6 KNOWN_{ijt} + \varepsilon_{ijt},$$

where

PRICE	= Fixed rates, cents/kwh = 7 or 9, or 0 if seasonal or time of day rates,
TOD	= Dummy for time of day rates, 11 cents 8AM-8PM, 5 cents 8PM – 8AM,
SEAS	= Dummy for seasonal rates, 10 summer, 8 winter, 6 spring and fall,
CNTL	= Fixed term contract with exit penalty, length 0, 1 year, 5 years,
LOCAL, KNOWN	= Dummies for familiarity: local utility, known but not local, unknown.

Data were collected in 1997 by the Research Triangle Institute for the Electric Power Research Institute.[23] The sample contains 361 individuals, each asked to make 12 choices from a set of 4 candidate firms.[24] There were a total of 4,308 choice situations analyzed.

[21] Pforr (2011) reports results for a moderate-sized problem with 4,344 individuals, about six periods and only two outcomes with four attributes. Using the brute force method takes over 100 seconds. The minimum distance estimator for the same problem takes 0.2 seconds to produce the identical results. The time advantage would be far greater for the four-choice model analyzed in Example 18.8.

[22] See also Train (2009, Chapter 11).

[23] Professor Train has generously provided the data for this experiment for us (and readers) to replicate, analyze, and extend the models in this example.

[24] A handful of the 361 individuals answered fewer than 12 choice tasks: two each answered 8 or 9; one answered 10 and eight answered 11.

This is an **unlabeled choice** experiment. There is no inherent distinction between the firms in the choice set other than the attributes. Firm 1 in the choice set is only labeled Firm 1 because it is first in the list. The choice situations we have examined in this chapter have varied in this dimension:

Example 18.2 Heating system types	labeled,
Example 18.3 Travel mode	labeled,
Example 18.4 Water heating type	labeled,
Example 18.5 Green energy	unlabeled,
Example 18.6 Malaria control guidelines	unlabeled,
Example 18.7 Heating systems	labeled,
Example 18.8 Electricity pricing	unlabeled.

One of the main uses of choice models is to analyze substitution patterns. In Example 18.3, we estimated elasticities of substitution among travel modes. Unlabeled choice experiments generally do not provide information about substitution between alternatives. They do provide information about willingness to pay. That will be the focus of the study in this example. When the utility function is based on price, rather than income, the marginal disutility of an increase in price is treated as a surrogate for the marginal utility of an increase in income for purposes of measuring willingness to pay. In general, the interpretation of the sign of the WTP is context specific. In the example below, we are interested in the perceived value of time of day rates, measured by the *TOD*/*PRICE* coefficients. Both coefficients are negative in the MNL model. But the negative of the price change is the surrogate for income. We interpret the WTP of approximately 10 cents/kwh as the amount the customer would accept as a fixed rate if he or she could avoid the *TOD* rates. But, the *LOCAL* brand value of the utility is positive, so the positive WTP is interpreted as the extra amount the customer would be willing to pay to be supplied by the local utility as opposed to an unknown supplier.

Table 18.17 reports estimates of the choice models for rate structures and utility companies. The MNL model shows marginal valuations of contract length, time, and seasonal rates relative to the fixed rates and the brand value of the utility. The WTP results are shown in Table 18.18. The negative coefficient on *Contract Length* implies that the average customer is willing to pay a premium of (0.17 cents/kwh)/year to avoid a fixed length contract. The offered contracts are one and five years, so customers appear to be willing to pay up to 0.85 cents/kwh to avoid a long-term contract. The brand value of the local utility compared to a new and unknown supplier is 2.3 cents/kwh. Since the average rate across the different scenarios is about 9 cents, this is quite a large premium. The value is somewhat less for a known, but not the local, utility. The coefficients on time of day and seasonal rates suggest the equivalent valuations of the rates compared to the fixed rate schedule. Based on the MNL model, the average customer would value the time of day rates as equivalent to a fixed rate schedule of 8.74 cents. The fixed rate offer was 7 or 9 cents/kwh, so this is on the high end.

The mixed logit model allows heterogeneity in the valuations. A normal distribution is used for the contract length and brand value coefficients. These allow the distributions to extend on both sides of zero so that, for example, some customers prefer the local utility while others do not. With an estimated mean of 2.16117 and standard deviation of 1.50097, these results suggest that $(1 - \Phi(2.16117/1.50097)) = 7.5\%$ of customers actually prefer an unknown outside supplier to their local utility. The coefficients on *TOD* and seasonal rates have been specified to have lognormal distributions. Because they are assumed to be negative, the specified coefficient is $-\exp(\beta + \sigma v)$. (The negative sign is attached to the variable and the coefficient on $-TOD$ is then specified with a positive lognormal distribution.) The mean value of this coefficient in the population distribution is then $E[\beta_{TOD}] = -\exp(2.11304 + 0.38651^2/2) = 8.915.$, so the average customer is roughly indifferent between the *TOD* rates and the fixed rate schedule. Figure 18.2 shows a kernel

TABLE 18.17 Estimated Choice Models for Electricity Supplier (Standard errors in parentheses)

		Mixed Logit[b]				
Variable	***MNL***[a]	***Mean β***	***Std. Dev. σ***	***FEM***	***REM***[c]	***ANA***[d]
Price	−0.62523	−0.86814	0.00000	−0.38841	−0.63762	−0.54713
	(0.03349)	(0.02273)	(0.00000)	(0.02039)	(0.07432)	(0.03962)
Contract	−0.10830	−0.21831	0.36379	−0.05586	−0.10940	−0.10937
	(0.01402)	(0.01659)	(0.01736)	(0.00682)	(0.00964)	(0.00862)
Time of Day[e]	−5.46276	2.11304[e]	0.38651	−3.46145	−5.57917	−5.11061
	(0.27815)	(0.02693)	(0.01847)	(0.16622)	(0.59680)	(0.30446)
Seasonal[e]	−5.84003	2.13564[e]	0.27607	−3.59727	−5.95563	−5.34035
	(0.27272)	(0.02571)	(0.01589)	(0.16596)	(0.61004)	(0.30811)
Local	1.44224	2.16117	1.50097	0.83266	1.47522	1.44016
	(0.07887)	(0.08915)	(0.08985)	(0.04106)	(0.09103)	(0.05510)
Known	0.99550	1.46173	0.97705	0.47649	1.02153	0.97419
	(0.06387)	(0.06538)	(0.07272)	(0.03319)	(0.07962)	(0.04944)
ln L	−4958.65	−3959.73		−4586.93	−4945.98	−4882.34

[a] Robust standard errors are clustered over individuals. Conventional standard errors for MNL are 0.02322, 0.00824, 0.18371, 0.18668, 0.05056, 0.04478, respectively.
[b] Train (2009) reports point estimates (b,s) = (−0.8827,0), (−0.2125, 0.3865), (2.1328, 0.4113), (2.1577, 0.2812), (2.2297, 1.7514), (1.5906, 0.9621) for Price, Cntl, TOD, Seas, Local, Known, respectively.
[c] Estimated Standard Deviations in RE Model are 0.00655 (0.02245), 0.47463 (0.06049), 0.016062 (0.04259).
[d] Class probabilities are 0.93739, 0.06261.
[e] Lognormal coefficient in mixed logit model is $\exp(\beta + \sigma v)$.

TABLE 18.18 Estimated WTP Based on Different Models

	Contract	***Local***	***Known***	***TOD***	***Seasonal***
Multinomial Logit Fixed Parameter					
Estimate	0.17322	2.30675	1.59223	8.73723	9.34065
Standard Error	0.02364	0.18894	0.13870	0.15126	0.15222
Lower Confidence Limit	0.12689	1.93643	1.32038	8.44076	9.04230
Upper Confidence Limit	0.21955	2.67707	1.86407	9.03370	9.63899
Mixed Logit WTP for Rates					
Lognormal					
Estimated Mean = $\exp(\beta + \sigma^2/2)$				8.91500	8.79116
Estimated Std. Dev. = Mean × $[\exp(\sigma^2) - 1]^{1/2}$				3.57852	2.47396
5% Lower Limit				1.90110	3.94220
95% Upper Limit				15.92900	13.64012
Triangular					
Estimated Mean = β				7.83937	8.19676
Estimated Spread = $\beta \pm \sigma$				5.90744	4.15295
Estimated Std. Dev. = $[\sigma^2/6]^{1/2}$				2.41170	1.69543
5% Lower Limit				3.11244	4.87370
95% Upper Limit				12.56630	11.51981

density estimator of the estimated population distribution of marginal valuations of the *TOD* rates. The bimodal distribution shows the sample of estimated values of $E[-\beta_{TOD} | \text{choices made}]$. Train notes, if the model is properly specified and the estimates appropriate, the means of these two distributions should be the same. The sample mean of the estimated conditional means is 10.4 cents/kwh while the estimated population mean is 9.9. The estimated standard deviation of the population distribution is $8.915 \times [\exp(0.38651^2) - 1]^{1/2} = 3.578$. Thus, about 95% of the population is estimated to value the *TOD* rates in the interval $9.9 +/- 7.156$. Note that a very high valuation of the *TOD* rates suggests a strong aversion to *TOD* rates. The lognormal distribution tends to produce implausibly large values such as those here in the thick tail of the distribution. We refit the model using triangular distributions that have fixed widths $\beta \pm \sigma$. The estimated distributions have range 7.839 ± 5.907 for *TOD* and 8.197 ± 4.152 for *Seasonal*. Computation of 95% probability intervals (based on a normal approximation, m $\pm$ 1.96s) are shown in Table 18.18.

Results are also shown for simple fixed and random effects estimates. The random effects results are essentially identical to the MNL results while the fixed effects results depart substantially from both the MNL and mixed logit results. The ANA model relates to whether, in spite of the earlier findings, there are customers who do not consider the brand value of the local utility in choosing their suppliers. The ANA model specifies two classes, one with full attendance and one in which coefficients on *LOCAL* and *KNOWN* are both equal to zero. The results suggest that 6.26% of the population ignores the brand value of the supplier in making their choices.

18.2.9 AGGREGATE MARKET SHARE DATA—THE BLP RANDOM PARAMETERS MODEL

The structural demand model of Berry, Levinsohn, and Pakes (BLP) (1995) is an important application of the mixed logit model. 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.[25] A basic structure is defined for

$$\begin{aligned}&\text{Markets, denoted } t = 1, \ldots, T,\\ &\text{Consumers in the markets, denoted } i = 1, \ldots, n_t,\\ &\text{Products, denoted } j = 1, \ldots, J.\end{aligned}$$

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

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

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

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

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 α, $\boldsymbol{\beta}$, $\boldsymbol{\pi}$, $\boldsymbol{\Pi}$, γ, 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),

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

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,

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

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

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

These would be simulated using the method of Section 18.2.7. The algorithm developed by BLP for estimation of the model is famously intricate and complicated. Several authors have proposed faster, less complicated methods of estimation. Lee and Seo (2011) proposed a useful device that is straightforward to implement.

Example 18.9 Health Insurance Market

Tamm, Tauchmann, Wasem, and Greβ (2007) analyzed the German health insurance market in this framework. The study was motivated by the introduction of competition into the German social health insurance system in 1996. The authors looked for evidence of competition in estimates of the price elasticities of the market shares of the firms using an extensive panel data set spanning 2001–2004. The starting point is a model for the market shares,

$$s_{it} = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{it} + \gamma_i + \varepsilon_{it})}{\sum_{i=1}^{N}\exp(\boldsymbol{\beta}'\mathbf{x}_{it} + \gamma_i + \varepsilon_{it})}, i = 1, \ldots, N.$$

Taking logs produces

$$\ln(s_{it}) = \boldsymbol{\beta}'\mathbf{x}_{it} + \delta_t + \gamma_i + \varepsilon_{it},$$

where δ_t is the log of the denominator, which is the same for all firms, and γ_i is an endogenous firm effect. Since consumers do not change their insurer every period, the model is augmented to account for persistence,

$$\ln(s_{it}) = \alpha \ln(s_{i,t-1}) + \boldsymbol{\beta}'\mathbf{x}_{it} + \delta_t + \gamma_i + \varepsilon_{it}.$$

The limiting cases of $\alpha = 0$ (the static case) and $\alpha = 1$ (random walk) are examined in the study, as well as the intermediate cases. GMM estimators are formulated for the three cases. The preferred estimate of the premium elasticity (from their Table VII) is -1.09, with a confidence interval of (-1.43 to -0.75), which suggests the influence of price competition in this market.

18.3 RANDOM UTILITY MODELS FOR ORDERED CHOICES

The analysts at bond rating agencies such as Moody's and Standard & Poor's 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 quality	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, among other activities, streams movies to subscribers. After a subscriber streams a movie, the next time

he or she logs onto the Web site, he or she is invited to rate that movie on a five-point scale, where five is the highest, most favorable rating. The ratings of the many thousands of subscribers who streamed 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 & 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) (www.iser.essex.ac.uk/bhps), the Australian HILDA data (www.melbourneinstitute.com/hilda/), and the German Socioeconomic Panel (GSOEP) (www.diw.de/en/soep), all 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 0 to 10 scale of the question about health satisfaction in the GSOEP.[26] Ratings such as these provide applications of the models and methods that interest us in this section.[27]

For an individual respondent, we hypothesize that there is a continuously varying strength of preferences that underlies the rating he or she submits. 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 are 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 for 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

[26]The original survey used a 0–10 scale for self-assessed health. It is currently based on a five-point scale.

[27]Greene and Hensher (2010a) provide a survey of ordered choice modeling. Other textbook and monograph treatments include DeMaris (2004), Long (1997), Johnson and Albert (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).

two levels of a rating scale (for example, 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 his or her own set of characteristics to the utility function, such as age, income, education, gender, where he or she lives, family situation, and so on, which we denote $x_{i1}, x_{i2}, \ldots, x_{iK.}$ They also bring their own aggregates 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,

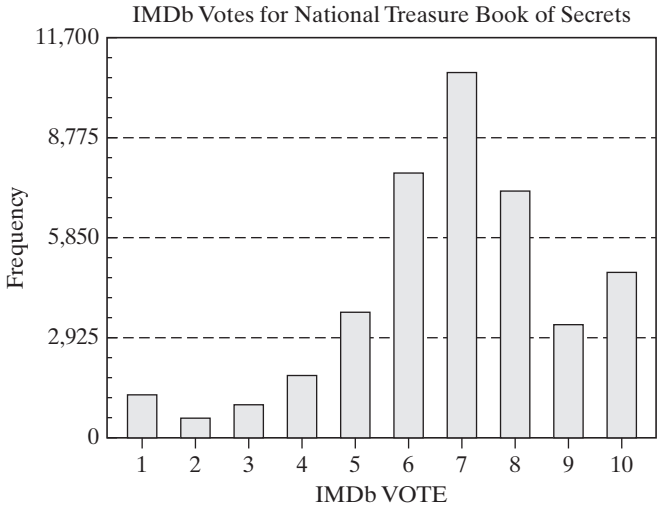
$$U_{im}^{*} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_{im}.$$

Example 18.10 Movie Ratings

The Web site www.IMDb.com invites visitors to rate movies that they have seen. This site uses a 10-point scale. It reported the results in Figure 18.3 for the movie *National Treasure: Book of Secrets* for 41,771 users of the site.[28] The figure 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

$$\begin{aligned}
R_{im} &= 1 \text{ if } -\infty < \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{im} \leq \mu_1,\\
R_{im} &= 2 \text{ if } \mu_1 < \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{im} \leq \mu_2,\\
&\cdots\\
R_{im} &= 9 \text{ if } \mu_8 < \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{im} \leq \mu_9,\\
R_{im} &= 10 \text{ if } \mu_9 < \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{im} < \infty.
\end{aligned}$$

FIGURE 18.3 IMDb.com Ratings.



	Votes	Mean
Males	33,644	6.5
Females	5,464	6.8
Aged under 18	2,492	6.6
Males under 18	1,795	6.5
Females under 18	695	7.2
Aged 18–29	26,045	6.6
Males Aged 18–29	22,603	6.6
Females Aged 18–29	3,372	6.8
Aged 30–44	8,210	6.4
Males Aged 30–44	7,216	6.4
Females Aged 30–44	936	6.7
Aged 45+	2,258	6.5
Males Aged 45+	1,814	6.4
Females Aged 45+	420	6.9
US Users	14,792	6.6
Non-US Users	24,283	6.5

[28]The data are as of December 1, 2008. A rating for the same movie as of August 1, 2016 at www.imdb.com/title/tt0465234/ratings?ref_=tt_ov_rt shows essentially the same pattern for 182,780 viewers.

Relying on a central limit theorem 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

$$\text{Prob}(R_{im} = 10 \,|\, \mathbf{x}_i) = \text{Prob}[\varepsilon_{im} > \mu_9 - \mathbf{x}_i'\boldsymbol{\beta}].$$

The structure provides the framework for an econometric model of how individuals rate movies (that they stream 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 simple probit model by focusing on the variable

$$E_{im} = 1 \text{ if } R_{im} = 10$$
$$E_{im} = 0 \text{ if } R_{im} < 10.$$

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.

The model described here is an *ordered choice model*. (The use 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 just 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)],
- Job skill rating [Marcus and Greene (1985)],
- Life satisfaction [Clark, Georgellis, and Sanfey (2001), Groot and ven den Brink (2003), Winkelmann (2002)],
- 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)], and
- 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 $\boldsymbol{\beta}$.

We assume that ε is normally distributed across observations.[29] For the same reasons as in the binomial probit model (which is the special case with $J = 1$), we normalize the mean and variance of ε to zero and one. We then have the following probabilities:

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

For all the probabilities to be positive, we must have

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

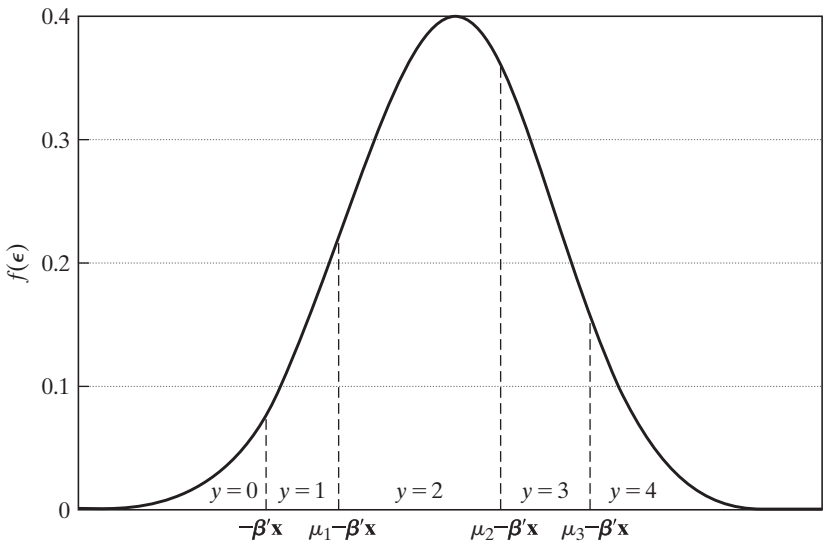
Figure 18.4 shows the implications of the structure. This is an extension of the univariate probit model we examined in Chapter 17. The log-likelihood function and its derivatives can be obtained readily, and optimization can be done by the usual means.

As usual, the partial 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 probabilities are

$$\begin{aligned} \text{Prob}(y = 0|\mathbf{x}) &= 1 - \Phi(\mathbf{x}'\boldsymbol{\beta}), \\ \text{Prob}(y = 1|\mathbf{x}) &= \Phi(\mu - \mathbf{x}'\boldsymbol{\beta}) - \Phi(-\mathbf{x}'\boldsymbol{\beta}), \\ \text{Prob}(y = 2|\mathbf{x}) &= 1 - \Phi(\mu - \mathbf{x}'\boldsymbol{\beta}). \end{aligned}$$

[29] 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.

FIGURE 18.4 Probabilities in the Ordered Probit Model.



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

$$\frac{\partial \text{Prob}(y = 0|\mathbf{x})}{\partial \mathbf{x}} = -\phi(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta},$$

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

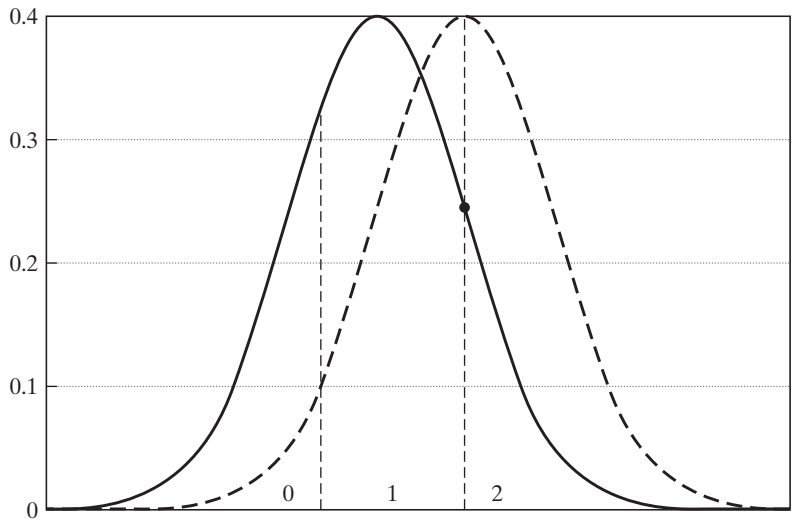
$$\frac{\partial \text{Prob}(y = 2|\mathbf{x})}{\partial \mathbf{x}} = \phi(\mu - \mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}.$$

Figure 18.5 illustrates the effect. The probability distributions of y and y^* are shown in the solid curve. Increasing one of the x's while holding $\boldsymbol{\beta}$ and μ constant is equivalent to shifting the distribution slightly to the right, which is shown as the dashed curve. The effect of the shift is unambiguously to shift some mass out of the leftmost cell. Assuming that $\boldsymbol{\beta}$ is positive (for this x), $\text{Prob}(y = 0|\mathbf{x})$ must decline. Alternatively, from the previous expression, it is obvious that the derivative of $\text{Prob}(y = 0|\mathbf{x})$ has the opposite sign from $\boldsymbol{\beta}$. By a similar logic, the change in $\text{Prob}(y = 2|\mathbf{x})$ [or $\text{Prob}(y = J|\mathbf{x})$ 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 $\text{Prob}(y = 0|\mathbf{x})$ and $\text{Prob}(y = J|\mathbf{x})$ are unambiguous! The upshot is that we must be very careful 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 18.11 Rating Assignments

Marcus and Greene (1985) estimated an ordered probit model for the job assignments of new Navy recruits. The Navy attempts to direct recruits into job classifications in which they will be

FIGURE 18.5 Effects of Change in x on Predicted Probabilities.



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) EDYR = 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 18.19. 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

TABLE 18.19 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 18.20 Partial Effect of a Binary Variable

	$-\hat{\boldsymbol{\beta}}'\mathbf{x}$	$\hat{\mu} - \hat{\boldsymbol{\beta}}'\mathbf{x}$	$Prob[y = 0]$	$Prob[y = 1]$	$Prob[y = 2]$
$MARR = 0$	−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

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 18.20.

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,

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

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.[30] 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

[30] See, for example, Long (1997, p. 141).

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

where $\hat{\boldsymbol{\beta}}^*$ is obtained by stacking the individual binary logit or probit estimates of $\boldsymbol{\beta}$ (without the constant terms).[31]

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 (that is, 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 18.12 Brant Test for an Ordered Probit Model of Health Satisfaction

In Examples 17.6 through 17.10 and several others, 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 to 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 first (1984) wave of the data set, which is a cross section of 4,483 observations. We then collapsed the 11 cells into 5 [(0–2), (3–5), (6–8), (9), (10)] for this example. The utility function is

$$\begin{aligned} HSAT_i^* = {} & \beta_1 + \beta_2\, AGE_i + \beta_3\, INCOME_i + \beta_4\, KIDS_i \\ & + \beta_5\, EDUC_i + \beta_6\, MARRIED_i\, \beta_7\, WORKING_i + \varepsilon_i. \end{aligned}$$

Variables *KIDS*, M*ARRIED*, 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 18.14.) 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% 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 17.9. A direct analog to the base case two-equation model is used in the study in Example 18.13.

Example 18.13 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 University 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

[31] See Brant (1990), Long (1997), or Greene and Hensher (2010a, p. 187) for details on computing the statistic.

ordered probit model of Section 18.3.1 to the level of calculus achievement, which is coded 0,1, . . . , 6:

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

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,

$$\textit{Grade}_i = \mathbf{z}_i'\delta + u_i, \quad \text{where} \quad u_i|\mathbf{z}_i \sim \text{N}[0, \sigma_u^2].$$

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

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

They thus adopt a "control function" approach to accommodate the endogeneity of the math attainment dummy variables. [See Sections 17.6.2d and 17.6.2e) 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.[32] 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

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

[32] A precise statement of the form of this variable is given in Li and Tobias (2006).

where

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

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.[33] 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 appear. Whether the results in Section 17.9.5 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 (1995)], 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.[34]
- The British Household Panel Survey (BHPS) and the Australian HILDA data include questions about health status.[35]
- The German Socioeconomic Household Panel (GSOEP) includes questions about subjective well-being[36] and subjective assessment of health satisfaction.[37]

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-Carbonell et al. (2004), analyzed survey data on job satisfaction using the Danish

[33]See Greene (2007b).

[34]See D'Addio (2004).

[35]See Contoyannis et al. (2004).

[36]See Winkelmann (2005).

[37]See Riphahn et al. (2003) and Example 18.4.

component of the European Community Household Panel (ECHP). Their estimator for an ordered logit model is built around the logic of Chamberlain's estimator for the binary logit model. [See Section 17.7.3.] 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 18.3.1.[38] 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.[39] Consider the base case ordered logit model with fixed effects,

$$y_{it}^{*} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \varepsilon_{it}|\mathbf{X}_i \sim \text{logistic}[0, \pi^2/3],$$

$$y_{it} = j \quad \text{if} \quad \mu_{j-1} < y_{it}^{*} < \mu_j, j = 0, 1, \ldots, J \quad \text{and} \quad \mu_{-1} = -\infty, \mu_0 = 0, \mu_J = +\infty.$$

The model assumptions imply that

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

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

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

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 17.7.3 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 and 18.2.8c.) The minimum distance estimator at the second step is chosen to minimize

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

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 $\mathbf{V}$.

[38]Cross-section versions of the ordered probit model with individual specific thresholds appear in Terza (1985a), Pudney and Shields (2000), and Greene (2009a).

[39]See Long's (1997) discussion of the "parallel regressions assumption," which employs this device in a cross-section framework.

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, so despite its superficial complexity, the estimation is straightforward.[40] (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 (2004a) (see, as well, Section 15.5.2), suggest that biases comparable to those in the binary choice models persist in the ordered probit model as well. (See, also, Bester and Hansen (2009) and Carro (2007).) 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 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 (2005), 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 Butler and Moffitt (1982) quadrature method (Section 14.9.6.c) can be extended to this model.

Winkelmann (2005) 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 14.14.2), 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.

18.14 Example Health Satisfaction

The GSOEP German Health Care data that we have used in Examples 11.16, 17.4, and others includes a self-reported measure of health satisfaction, *HSAT*, that takes values 0, 1, . . . ,10.[41] 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:

[40] See Pratt (1981).

[41] 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.

HSAT	*Frequency*	*Proportion (%)*
0	447	1.6
1	255	0.9
2	642	2.3
3	1,173	4.2
4	1,390	5.0
5	4,233	15.4
6	2,530	9.2
7	4,231	15.4
8	6,172	22.5
9	3,061	11.2
10	3,192	11.6

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

$$y_{it}^{*} = \beta_1 + \beta_2\, Age_{it} + \beta_3\, Income_{it} + \beta_4\, Kids_{it} + \beta_5\, Education_{it} + \beta_6\, Married_{it} + \beta_7\, Working_{it} + \varepsilon_{it} + c_i,$$

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 18.21 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 *j*th block is the inverse of the *j*th 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 15.6.3) is considerably lower 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 18.3, the components of the ordered choice model must be interpreted with some care. By construction, the partial effects of the variables on the probabilities of the outcomes must change sign, so the simple coefficients do not show the complete picture implied by the estimated model. Table 18.22 shows the partial effects for the pooled model to illustrate the computations.

Example 18.15 A Dynamic Ordered Choice Model:

Contoyannis, Jones, and Rice (2004) analyzed a self-assessed health (SAH) scale that ranged from 1 (very poor) to 5 (excellent) in the British Household Panel Survey. The data set examined consisted of the first eight waves of the data set, from 1991 to 1999, roughly 5,000

TABLE 18.21 Estimated Ordered Probit Models for Health Satisfaction

	(1)	(2)	(3)	(4)	(5)	
		Fixed Effects	Fixed Effects	Random	Random Effects Mundlak	
Variable	Pooled	Uncond.	Conditional	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)	(0.00244)
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	−0.00797	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	
$\ln L$	−56,813.52	−41,875.63		−53,215.54	−53,070.43	

TABLE 18.22 Estimated Partial 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

households. 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.[42] 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 (2010) 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.

The authors were interested in transitions in the reported health status, especially to and from the highest level. Based on the balanced panel for women, the authors estimated the unconditional probabilities of transition to Excellent Health from (Excellent, Good, Fair, Poor, and Very Poor) to be (0.572, 0.150, 0.040, 0.021, 0.014).[43]

The presence of attrition complicates the analysis. The authors examined the issue in a set of tests, and found evidence of nonrandom attrition for men in the sample, but not women. (See Example 11.2 in Section 11.2.5, where we have examined their study.) Table 18.23, extracted from their Table XII, displays a few of the partial effects of most interest, the implications for the probability of reporting the highest value of SAH.[44] Several specifications were considered. Model (4) in the results includes the IPW treatment for possible attrition (see Section 17.7.7). Model (6) is the most general specification considered. Surprisingly, the income effect is extremely small. However, given the considerable inertia suggested by the transition probabilities, one might expect that it would require a large change in the covariates to induce switching out of the top cell. The mean log income in the data is about 0.5 and the proportion of responders who report *EX* is roughly 4884/23,408 = 0.2086. If log income rises by 0.1, or 20%, the average probability for *EX* would rise by only 0.1 × 0.008 = 0.0008, which is trivial. Having reported *EX* in the previous period is expected to raise the probability by 0.074 compared to the value if SAH were *GOOD* (the omitted cell is the second one), which is substantial.

TABLE 18.23 Average Partial Effects on Probability of Reporting Excellent Health

	Pooled Model (4)	***Random Effects Model (6)***
ln Income	0.004 (0.002)	0.008 (0.004)
SAH EX(*t*-1)	0.208 (0.092)	0.074 (0.035)
SAH FAIR(*t*-1)	−0.127 (0.074)	−0.061 (0.033)

[42]This is the same device that was used by Butler et al. (1994) in Example 18.13. Van Ooijen, Alessie, and Knoef (2015) also analyzed self-assessed health in the context of a dynamic ordered choice model, using the Dutch Longitudinal Internet Study in the Social Sciences.

[43]Figures from Contoyannis, Jones, and Rice (2004), Table II.

[44]Contoyannis et al. (2004).

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[45] and heterogeneity in the preferences (i.e., random parameters and latent classes).[46] 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 Greene and Hensher (2010a).

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 (2010a) all present applications that include individual variation in the thresholds of the ordered choice model.

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

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

With three outcomes, the probabilities are formed from

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

and

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

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

[45] See Section 17.5.2, Keele and Park (2005), and Wang and Kockelman (2005), for an application.

[46] An extensive study of heterogeneity in health satisfaction based on 22 waves of the GSOEP is Jones and Schurer (2010).

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

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

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 (2007b), Eluru, Bhat, and Hensher (2008), and Greene and Hensher (2010a) employ a hierarchical ordered probit, or HOPIT model,

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

Case 2 is the Terza (1985a) and Pudney and Shields's (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,

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

and in case 2,

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

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

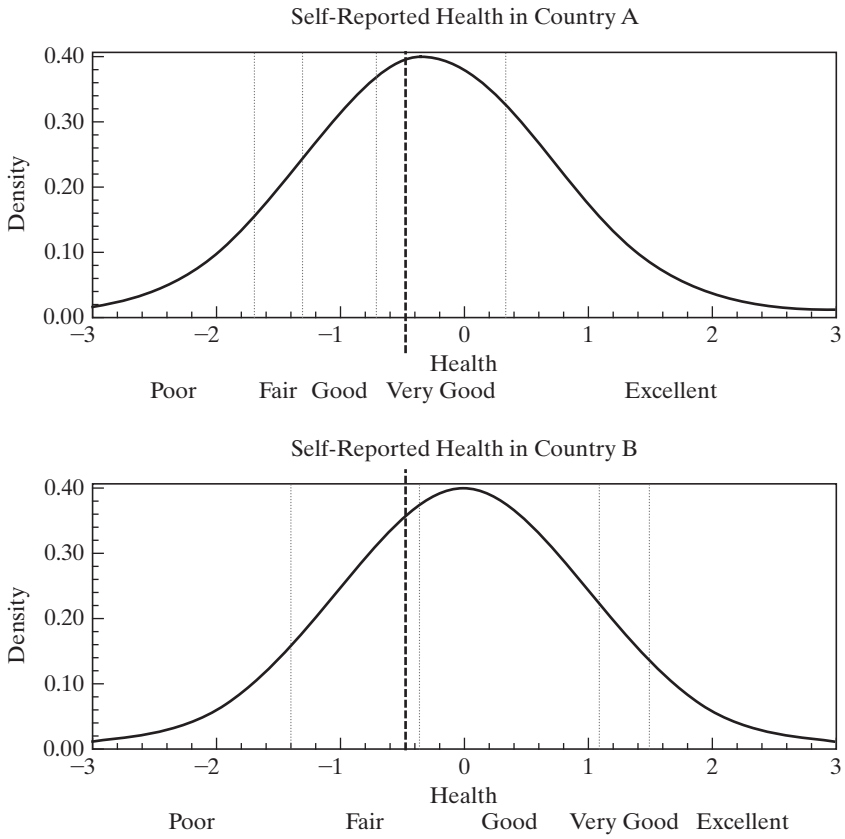
The introduction of observed heterogeneity into the threshold parameters attempts to deal with a fundamentally restrictive assumption of the ordered choice model. Survey respondents rarely view the survey questions exactly the same way. This is certainly true in surveys of health satisfaction or subjective well-being.[47] 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 et al. (2007) suggest the results in Figure 18.6 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 he or she is 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.[48] 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

[47]See Boes and Winkelmann (2006b) and Ferrer-i-Carbonell and Frijters (2004).

[48]See also Kristensen and Johansson (2008).

FIGURE 18.6 Differential Item Functioning in Ordered Choices.



new; KMST cite a number of earlier uses. The innovation is their method for incorporating the approach in a formal model for ordered choices. The bivariate and multivariate probit models that they develop combine the elements described in Sections 18.3.1 through 18.3.3 and the HOPIT model in Section 18.3.5.

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 14.13 and 17.33, we examined the number of times an individual visited the physician using the GSOEP data. The credit default data that we used in Example 17.21 also includes another behavioral variable, the number of derogatory reports in an individual's credit history. Finally, in Example 17.36, we analyzed data on firm innovation. Innovation is often analyzed in terms of the number of patents that the firm obtains (or applies for).[49] In each of these cases, the variable of interest is a count

[49]For example, by Hausman, Hall, and Griliches (1984) and many others.

of events. This obviously differs from the discrete dependent variables we have 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 through 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 Sections 18.4.8 and 18.4.9.

18.4.1 THE POISSON REGRESSION MODEL

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

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

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 or per unit of space is given by

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

so

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

With the parameter estimates in hand, this vector can be computed using any data vector desired or averaged across the sample to estimate the average partial effects. Because the model to this point is a straightforward regression, computation of treatment effects (at this point) is simple as well. For *exogenous* treatment indicator, T,

$$E[y|\mathbf{x}, T] = \exp(\mathbf{x}'\boldsymbol{\beta} + \gamma T).$$

So, average treatment effects can be estimated with

$$\text{ATE} = \frac{1}{n}\sum_{i=1}^{n}[\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}} + \hat{\gamma}) - \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})].$$

ATET is computed by averaging over only those observations with T = 1. The case of endogenous treatment is more complicated, as usual, and is examined in Section 18.4.9.

In principle, the Poisson model is simply a nonlinear regression. But it is 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, $\left[\sum_{i=1}^{n}\hat{\lambda}_i\mathbf{x}_i\mathbf{x}_i'\right]^{-1}$ provides an estimator of the asymptotic covariance matrix for the parameter estimator.

There are a variety of extensions of the Poisson model—some considered later in Section 18.4.5—that introduce heterogeneity or relax the assumption of equidispersion. In general, the implication of these extensions is upon the (heteroscedastic) variance of the random variable. The conditional mean function remains the same; $E[y|\mathbf{x}] = \lambda(\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta})$. A consequence is that the Poisson log likelihood will provide a consistent ML estimator of $\boldsymbol{\beta}$ even in the presence of a wide variety of failures of the Poisson model assumptions. Thus, the Poisson MLE is one of the fundamental examples of a QMLE. In these settings, it is generally appropriate to adjust the estimated asymptotic covariance matrix of the estimator. For this case, a robust covariance matrix is computed using

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

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 the delta method (see Section 4.6).

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. Characteristically, the LM statistic can be computed as nR^2 in the regression of a column of ones on $\mathbf{g}_i = e_i \mathbf{x}_i$.

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.[50] A measure based on the standardized residuals is

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

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:* because y_i is an integer while the

[50] See the surveys by Cameron and Windmeijer (1993), Gurmu and Trivedi (1994), and Greene (2005).

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 $\ell(y_i, y_i)$. 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.[51] We note, finally, the passing resemblance of R_d^2 to the "*pseudo*-R^2," or "likelihood ratio index" reported by some statistical packages (for example, *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

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

[51]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.[52] The LM statistic is

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

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:

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

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.[53] 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:

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

[52]See Cameron and Trivedi (1986, p. 41).

[53]See Hausman, Hall, and Griliches (1984), Cameron and Trivedi (1986, 1998), Gurmu and Trivedi (1994), Johnson and Kotz (1993), and Winkelmann (2005) for discussion.

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

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

The choice of a density for u_i defines the unconditional distribution. For mathematical convenience, a gamma distribution is usually assumed for $u_i = \exp(\varepsilon_i)$.[54] 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

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

which is one form of the **negative binomial distribution**. The distribution has conditional mean λ_i and conditional variance $\lambda_i(1 + (1/\theta)\lambda_i)$.[55] 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, $E[y_i|\mathbf{x}_i] = \text{Var}[y_i|\mathbf{x}_i]$, 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 or space*. Thus, there is a presumption in the model formulation, for example, the Poisson, that the same amount of time is observed for each i. In a spatial

[54] An alternative approach based on the normal distribution is suggested in Terza (1998), Greene (1995b, 1997, 2005), Winkelmann (2003), 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 14.14.4 and 17.7.2.

[55] This model is Negbin 2 in Cameron and Trivedi's (1986) presentation.

context, such as measurements of the prevalence 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:

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

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

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

This is the default form of the model in the standard 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

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

This is not a simple reparameterization of the model. The results in Example 18.15 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, 2008b).] 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 (18-21) and (18-22). The full model is

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

The conditional mean function for the three cases considered is

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

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

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

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 (2016).

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,

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

suggests that a convenient 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 18.16 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.[56] 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

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

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

Table 18.24 presents the estimates of the several count models. In all specifications, the coefficient on *ADDON* is positive but not statistically significant, which is consistent with the results in the authors' study. They found evidence of moral hazard in a simple model,

[56]Munkin and Trivedi (2007) is a similar application to dental insurance.

TABLE 18.24 Estimated Models for *DocVis* (standard errors in parentheses)

Variable	*Poisson*	*Negbin 2*	*Negbin 2 Heterogeneous*	*Negbin 1*	*Negbin P*	*Poisson Normal*
Constant	1.05266	1.10083	1.14129	0.93184	0.97164	0.09302
	(0.11395)	(0.05970)	(0.06175)	(0.05630)	(0.06389)	(0.04364)
Age	0.01838	0.01789	0.01689	0.01571	0.01888	0.02267
	(0.00134)	(0.00079)	(0.00081)	(0.00070)	(0.00081)	(0.00051)
Education	−0.04355	−0.04797	−0.04450	−0.03127	−0.04282	−0.04595
	(0.00699)	(0.00378)	(0.00386)	(0.00355)	(0.00414)	(0.00276)
Income	−0.52502	−0.46285	−0.45443	−0.23198	−0.37774	−0.45804
	(0.08240)	(0.04600)	(0.04654)	(0.04451)	(0.05122)	(0.03235)
Kids	−0.16109	−0.15656	−0.16266	−0.13658	−0.16521	−0.18450
	(0.03118)	(0.01735)	(0.01769)	(0.01648)	(0.01855)	(0.01217)
AddOn	0.07282	0.07134	0.06839	0.17879	0.16107	0.27067
	(0.07801)	(0.07205)	(0.07142)	(0.05493)	(0.06969)	(0.04068)
	[0.06548]					
	{0.02534}					
P	0.0000	2.0000	2.0000	1.0000	1.52377	
	—	—	—	—	(0.03485)	
α	0.0000	1.92971	2.61217	6.19585	3.34512	
σ	—	(0.02009)	(0.05965)	(0.06867)	(0.13995)	1.31484
	—	—	—	—	—	(0.00425)
	—	—	—	—	—	
δ *(Female)*	—	—	−0.38157	—	—	
	—	—	(0.02040)	—	—	
δ *(Married)*	—	—	−0.13661	—	—	
	—	—	(0.02305)	—	—	
ATE	0.24018	0.23491	0.22070	0.62105	0.55460	0.42961
	(0.26637)	(0.24561)	(0.23850)	(0.20782)	(0.25929)	(0.07399)
ATET	0.21945	0.21482	0.21781	0.59304	0.51528	0.39914
	(0.24317)	(0.22454)	(0.25055)	(0.19813)	(0.24066)	(0.06810)
ln *L*	−104,603.0	−60,291.50	−60,149.00	−60,274.94	−60,219.19	−60,619.11

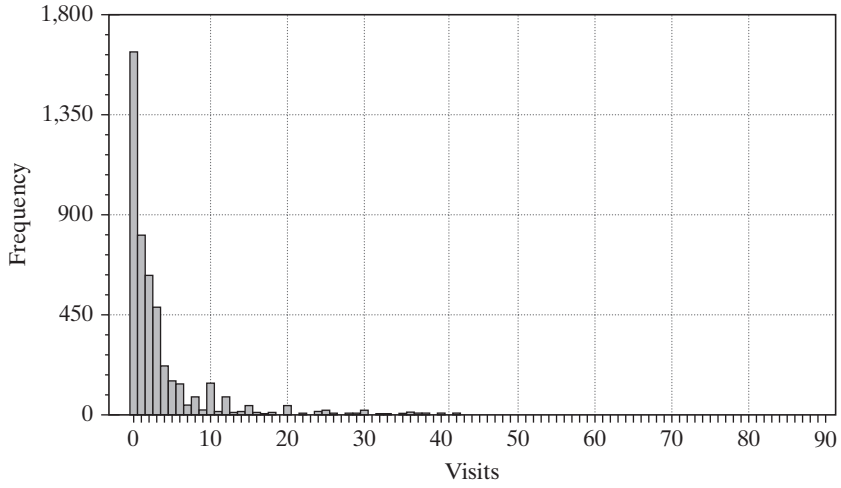
but none when their model was expanded. The various test statistics strongly reject the hypothesis of equidispersion. Cameron and Trivedi's (1990) semiparametric tests from the Poisson model (see Section 18.4.3) have *t* statistics of 22.151 for $g_i = \mu_i$ and 22.440 for $g_i = \mu_i^2$. Both of these are far larger than the critical value of 1.96. The LR statistic comparing to the NB model is over 80,000, 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 four 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 2.0000, 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 ATE and ATET computed for *ADDON* are generally quite small with the Poisson and NB models—the mean of *DocVis* is about 3.2 and the effect is about 0.2 and insignificant. The effect is larger in the less restrictive NBP and normal mixture models. The evidence here, as in RWM, is mixed.

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.[57] 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 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 Martínez-Espinera and Amoako-Tuffour (2008) are 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 18.7 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 the 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

FIGURE 18.7 Number of Doctor Visits, 1988 Wave of GSOEP Data.



[57]Shaw (1988) and Bockstael et al. (1990).

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 18.18, 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 to 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—with

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

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

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

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[58]

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

The conditional mean function is

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

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

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

[58]See, for example, Mullahy (1986), Shaw (1988), Grogger and Carson (1991), Greene (1995a,b), and Winkelmann (2003).

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 18.17 Major Derogatory Reports

In Examples 17.17 and 17.21, 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 can 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 2,561 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 in Example 18.18, 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 18.25. The first column of results compared to the second shows the 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 a 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.[59] The observed random variable, y_i, is constructed from an underlying random variable, y_i^*, by $y_i = \text{Min}(y_i^*, C)$. Wang and Zhou (2015) applied this specification with a negative binomial count model to a study of the number of deliveries to online shoppers. The dependent variable, deliveries, ranging from 0 to 200, was censored at 10 for the analysis.

TABLE 18.25 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)	0.0001	(0.25)	0.0001	(0.30)
ln L	−5,379.30		−5,378.79		−5,097.08	
			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		0.0001		0.0001	
Cond'l. Mean	0.4628		2.4295		2.4295	
Scale factor	0.4628		2.4295		1.7381	

[59] See Terza (1985b).

Probabilities in the presence of censoring are constructed using the axioms of probability. This produces

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

In this case, the conditional mean function is

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

The infinite sum can be computed by using the complement. Thus,

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

Example 18.18 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.[60]

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.

[60]In addition, his 1977 companion paper in *Econometrica* on estimation of the tobit model proposed a variant of the EM algorithm, developed by Dempster, Laird, and Rubin (1977).

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) = 7, (monthly, weekly, or daily) = 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.

A 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 18.26 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 7.015/0.945 = 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 lower panel of results in Table 18.26, 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.[61] We will examine them for the Poisson model. Hausman, Hall and Griliches (1984) 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

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

[61] Hausman, Hall, and Griliches (1984) give full details for these models.

TABLE 18.26 Censored Poisson and Negative Binomial Distributions

	Poisson Regression			*Negative Binomial Regression*		
Variable	*Estimate*	*Std. Error*	*Partial Effect*	*Estimate*	*Std. Error*	*Partial Effect*
			Based on Uncensored Poisson Distribution			
Constant	2.53	0.197	—	2.19	0.859	—
z_2	−0.0322	0.0059	−0.047	−0.0262	0.0180	−0.0039
z_3	0.116	0.0099	0.168	0.0848	0.0401	0.127
z_5	−0.354	0.0309	−0.515	−0.422	0.171	−0.632
z_7	0.0798	0.0194	0.116	0.0604	0.0909	0.0906
z_8	−0.409	0.0274	−0.596	−0.431	0.167	−0.646
α				7.015	0.945	
ln L	−1,427.037			−728.2441		
			Based on Poisson Distribution Right Censored at y = 4			
Constant	1.90	0.283	—	4.79	1.16	—
z_2	−0.0328	0.0084	−0.0235	−0.0166	0.0250	−0.0043
z_3	0.105	0.0140	0.0755	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	0.0572	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		

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 (that is, 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 about robust covariance matrices would be relevant here. However, if the model is, indeed, complete, then the robust estimator does no harm.

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

With fixed effects, the Poisson distribution will 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 17.7.3. 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,

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

Unlike the regression or the probit model, this estimator 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 also free of the fixed effects. This is the default FENB formulation used in popular software packages such as *SAS* and *Stata*. 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.[62] 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):

$$\begin{aligned}
&\text{NB1(HHG):} \quad E[y_{it}|\mathbf{x}_{it}] = \theta_i \phi_{it} = \theta_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta}),\\
&\text{NB2:} \qquad\quad E[y_{it}|\mathbf{x}_{it}] = \exp(\alpha_i)\phi_{it} = \lambda_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i).
\end{aligned}$$

The conditional variances are

$$\begin{aligned}
&\text{NB1(HHG):} \quad \text{Var}[y_{it}|\mathbf{x}_{it}] = \theta_i \phi_{it}[1 + \theta_i],\\
&\text{NB2:} \qquad\quad \text{Var}[y_{it}|\mathbf{x}_{it}] = \lambda_{it}[1 + \theta\lambda_{it}].
\end{aligned}$$

Letting $\mu_i = \ln \theta_i$, it appears that the HHG formulation does provide a fixed effect in the mean, as now, $E[y_{it}|\mathbf{x}_{it}] = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \mu_i)$. 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,

$$\begin{aligned}
E[y_{it}|\mathbf{x}_{it}] &= \gamma_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \mu_i),\\
\text{Var}[y_{it}|\mathbf{x}_{it}] &= \gamma_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \mu_i)/[1 + \gamma_i \exp(\mu_i)].
\end{aligned}$$

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

[62] This issue is explored at length in Allison (2000) and Allison and Waterman (2002).

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.[63]

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 Section 11.5), 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

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

Then the random effect is swept out by obtaining

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

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 $p(y_{it} | u_i)$ 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,

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

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,

[63] See Greene (2005) and Allison and Waterman (2002) for further discussion.

adding a random effect to the negative binomial model might well amount to introducing the heterogeneity a second time—the random effects NB model is a Poisson regression with $E[y_{it}|\mathbf{x}_{it}, \varepsilon_i, w_{it}] = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + w_{it} + \varepsilon_i)$. 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, i.e., $G(\theta_i, \theta_i)$. 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,

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

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 18.19 for a comparison.

There is a 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, 1995b, 1997), 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 14.15.4 and the random parameters model in Section 15.9 extend naturally to the Poisson model. Indeed, most of the received applications of the latent class structure have been in the Poisson or negative binomial regression framework.[64]

Example 18.19 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 18.27 presents the fixed and random effects estimates of the equation. 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

TABLE 18.27 Estimated Panel Data Models for Doctor Visits (standard errors in parentheses)

	Poisson				*Negative Binomial*			
					Fixed Effects		*Random Effects*	
Variable	*Pooled Robust Std. Error*	*Fixed Effects*	*Random Effects*	*Pooled NB2*	*FE NB1*	*FE NB2*	*HHG Gamma*	*Normal*
Constant	1.05266	—	0.69553	1.10083	−1.14543	—	−0.41087	0.37764
	(0.11395)		(0.05266)	(0.05970)	(0.09392)		(0.06062)	(0.05499)
Age	0.01838	0.03127	0.02331	0.01789	0.02383	0.04476	0.01886	0.02230
	(0.00134)	(0.00144)	(0.00045)	(0.00079)	(0.00119)	(0.00277)	(0.00078)	0.00070)
Educ	−0.04355	−0.03934	−0.03938	−0.04797	0.01338	−0.04788	−0.02469	−0.04536
	(0.00699)	(0.01734)	(0.00434)	(0.00378)	(0.00630)	(0.02963)	(0.00386)	(0.00345)
Income	−0.52502	−0.30674	−0.27282	−0.46285	0.01635	−0.20085	−0.10785	−0.18650
	(.08240)	(0.04103)	(0.01519)	(0.04600)	(0.05541)	(0.07321)	(0.04577)	(0.04267)
Kids	−0.16109	0.00153	−0.03974	−0.15656	−0.03336	−0.00131	−0.11181	−0.12013
	(0.03118)	(0.01534)	(0.00526)	(0.01735)	(0.02117)	(0.02921)	(0.01677)	(0.01583)
AddOn	0.07282	−0.07946	−0.05654	0.07134	0.11224	−0.02158	0.15086	0.05637
	(0.07801)	(0.03568)	(0.01605)	(0.07205)	(0.06622)	(0.06739)	(0.05836)	(0.05699)
α	—	—	1.16959	1.92971	—	1.91953	—	1.08433
			(0.01949)	(0.02009)		(0.02993)		(0.01210)
a	—	—	—	—	—	—	2.13948	—
							(0.05928)	
b	—	—	—	—	—	—	3.78252	—
							(0.11377)	
σ	—	—	—	—	—	—	—	0.96860
								(0.00828)
ln L	−104,603.0	−60,327.8	−71,779.6	−60,291.5	34,015.4	−49,478.0	−58,189.5	−58,170.5

[64]See Greene (2001) for a survey.

considerable amount of variation across the specifications. With respect to the parameter of interest, *AddOn,* we find that the size of the coefficient falls substantially with all panel data treatments and it becomes negative in the Poisson models. 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 *AddOn*, varies in sign, and is insignificant in nearly all cases. As before, the data do not suggest the presence of moral hazard, at least as measured here.

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 0.09027, 0.49332, and 0.41651. The coefficients on *AddOn* in the three classes, with associated *t* ratios, are −0.02191 (0.45), 0.36825 (5.60), and 0.01117 (0.26). The qualitative result concerning evidence of moral hazard suggested here is that there might be a segment of the population for which we have some evidence, but more generally, we find relatively little.

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.[65] 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

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

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:

$$\text{Prob}(z_i = 0|\mathbf{w}_i) = F(\mathbf{w}_i, \boldsymbol{\gamma}), \text{(Regime 1: } y \text{ will equal zero)};$$
$$\text{Prob}(y_i = j|\mathbf{x}_i, z_i = 1) = \frac{\exp(-\lambda_i)\lambda_i^j}{j!}, \text{(Regime 2: } y \text{ will be a count outcome)}.$$

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

[65]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)].

Poisson or negative binomial data-generating process.[66] The extension of the ZIP formulation to the negative binomial model is widely labeled the ZINB model.[67] [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 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

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

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 2, 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.

Example 18.20 Zero-Inflation Models for Major Derogatory Reports

In Example 18.17, 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

[66] 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.)

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

TABLE 18.28 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	−15,467.71	−11,569.74		−10,582.88	−10,516.46	
Vuong		20.6981			4.5943	

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 18.17 to accommodate the zeros. The extensions to the ZIP and ZINB models are shown in Table 18.28. 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 8,609. For the ZIP model, the value is 10,914.8, 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.[68] 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

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

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

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

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

the 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,

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

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,

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

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 means 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 level

as the intensity (count) equation.[69] Both of these modifications substantially alter the Poisson formulation. First, note that the equality of the mean and variance of the distribution no longer follow; 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 18.21 Hurdle Models for Doctor Visits

Jones and Schurer (2009) used the hurdle framework to study physician visits in several countries using the ECHP panel data set. The base model was a negative binomial regression, with a logit hurdle equation. The main interest was the cross-country variation in the income elasticity of health care utilization. A few of their results for general practitioners are shown in Table 18.29, which is extracted from their Table 8.[70] (Corresponding results are computed for specialists.) Note that individuals are classified as high or low users. The latent classes have been identified as a group of heavy users of the system and light users, which would seem to suggest that the classes are not latent. The class assignments are done using the method described in Section 14.15.4. The posterior (conditional) class probabilities, $\hat{\pi}_{i1}$ and $\hat{\pi}_{i2}$, are computed for each person in the sample. An individual is classified as coming from class 1 if $\hat{\pi}_{i1} \geq 0.5$ and class 2 if $\hat{\pi}_{i1} < 0.5$. With this classification, the average within group utilization is computed. The group with the higher group mean is labeled the "High users."

In Examples 18.16 and 18.21, we fit Poisson regressions with means

$$E[DocVis|\mathbf{x}] = \exp(\beta_1 + \beta_2 Age + \beta_3 Education + \beta_4 Income + \beta_5 Kids + \beta_6 AddOn).$$

TABLE 18.29 Income Elasticities

Estimated Income Coefficients and Elasticities for GP and Specialist Visits—Country-Specific LC Hurdle Models (Asymptotic t ratios in parentheses)

		GPs			
		Low Users		*High Users*	
Country		*Estimated Coefficient*	*Estimated Elasticity*	*Estimated Coefficient*	*Estimated Elasticity*
Austria	$P(Y > 0)$	−0.051 (−1.467)	−0.012	−0.109 (−0.872)	−0.005
	$E(Y\|Y > 0)$	0.012(0.693)	0.009	0.039(2.167)	0.035
Belgium	$P(Y > 0)$	0.035(1.002)	0.008	0.292(4.004)	0.010
	$E(Y\|Y > 0)$	−0.052(−3.125)	−0.037	−0.055(−4.030)	−0.050
Denmark	$P(Y > 0)$	0.083(1.746)	0.033	0.261 (2.302)	0.023
	$E(Y\|Y > 0)$	0.042 (0.992)	0.021	−0.030 (−1.009)	−0.024
Finland	$P(Y > 0)$	0.054(1.358)	0.024	−0.030 (−0.263)	−0.003
	$E(Y\|Y > 0)$	0.007(0.237)	0.004	−0.048 (−1.706)	−0.037

[69]See, for example, Jones (1989), who applied the model to cigarette consumption.

[70]From Jones and Schurer (2009).

Table 18.30 reports results for a two-class latent class model based on this specification using the 3,377 observations in the 1994 wave of the panel. The estimated prior class probabilities are 0.23298 and 0.76702. For each observation in the sample, the posterior probabilities are computed using

$$\hat{\pi}_{i1} = \frac{\hat{\pi}_1 \hat{L}_{i1}}{\hat{\pi}_1 \hat{L}_{i1} + \hat{\pi}_2 \hat{L}_{i2}}, \hat{L}_{ic} = \frac{\exp(-\hat{\lambda}_{ic})(\hat{\lambda}_{ic})^{DocVis_i}}{DocVis_i!}, \hat{\lambda}_{ic} = \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_c), c = 1, 2,$$

then $\hat{\pi}_{i2} = 1 - \hat{\pi}_{i1}$. The mean values of these posterior probabilities are 0.228309 and 0.771691, which, save for some minor error, match the prior probabilities. (In theory, they match perfectly.) We then define the class assignment to be class 1 if $\hat{\pi}_{i1} \geq 0.5$ and class 2 if $\hat{\pi}_{i1} < 0.5$. By this calculation, there are 771 and 2,606 observations in the two classes, respectively. The sample averages of *DocVis* for the two groups are 11.380 and 1.535, which confirms the idea of a group of high users and low users. Figure 18.8 displays histograms for the two groups. (The sample has been trimmed by dropping a handful of observations larger than 30 in group 1.)

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), were interested in the role of the *AddOn* 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 18.21.

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

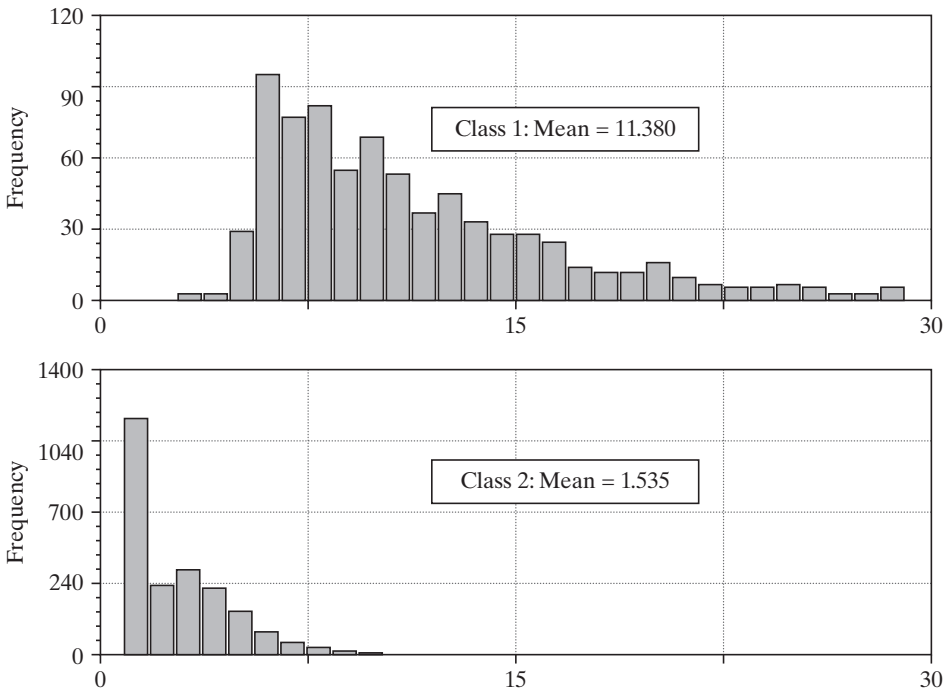
$$Income = \mathbf{z}_i'\boldsymbol{\gamma} + u_i,$$
$$\text{Prob}(DocVis_i = j | \mathbf{x}_i, Income_i) = \exp(-\lambda_i), \lambda_i^j/j!, \lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \delta\, Income_i).$$

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

TABLE 18.30 Estimated Latent Class Model for Doctor Visits

	Latent Class Model				*Poisson Regression*	
	Class 1		*Class 2*			
Variable	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*
Constant	2.67381	0.11876	0.66690	0.17591	1.23358	0.06706
Age	0.01394	0.00149	0.01867	0.00213	0.01866	0.00082
Income	−0.39859	0.08096	−0.51861	0.12012	−0.40231	0.04632
Education	−0.05760	0.00699	−0.06516	0.01140	−0.04457	0.00435
Kids	−0.13259	0.03539	−0.32098	0.05270	−0.14477	0.02065
AddOn	0.00786	0.08795	0.06883	0.15084	0.12270	0.06129
Class Prob.	0.23298	0.00959	0.76702	0.00959	1.00000	0.00000
ln *L*	−9263.76				−13653.41	

FIGURE 18.8 Distributions of Doctor Visits by Class.



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.[71] If the model is reformulated to accommodate heterogeneity, as in

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

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:

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

[71] See Terza (2009, pp. 555–556) for discussion of this issue.

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 $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, δ, σ_ε, σ_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,

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

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 17.6.2 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 the joint density of *DocVis* and *Income* as $P(DocVis \,|\, Income) f(Income)$. Thus,

$$f(DocVis, Income) = \frac{\exp[-\lambda_i(w_i)][\lambda_i(w_i)]^{y_i}}{y_i!} \frac{1}{\sigma_u} \phi\left(\frac{Income - \mathbf{z}_i'\boldsymbol{\gamma}}{\sigma_u} \right),$$
$$\lambda_i(w_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \delta\, Income_i + \tau(Income_i - \mathbf{z}_i'\boldsymbol{\gamma})/\sigma_u + \theta w_i).$$

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 $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \delta, \sigma_u, \sigma_\varepsilon, \rho)$. The invariance principle can be 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 $\boldsymbol{\alpha} = (1/\sigma_u)\boldsymbol{\gamma}$.

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

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

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 17.3),

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

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

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

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

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

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 feasible 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.[72]

Example 18.22 Endogenous Treatment in Health Care Utilization

Table 18.31 reports estimates of the treatment effects model for our health care utilization data. The main result is the causal parameter on *Addon,* which is shown in the boxes in the table. We have fit the model with the full panel (pooled) and with the final (1994) wave of the panel. The results are nearly identical. The large negative value is, of course, inconsistent with any suggestion of moral hazard, and seems extreme enough to cast some suspicion on the model specification. We, like Riphahn et al. (2003) and others they discuss, did not find evidence of moral hazard in the demand for physician visits. (The authors did find more suggestive results for hospital visits.)

[72]See Greene (2005, 2007d).

TABLE 18.31 Estimated Treatment Effects Model (Standard errors in parentheses)

	Full Panel		*1994 Wave*	
	Treatment	***Outcome***	***Treatment***	***Outcome***
Variable	***(Probit: Addon)***	***(Poisson: DocVis)***	***(Probit: Addon)***	***(Poisson: DocVis)***
Health Sat.	0.10824		0.13202	
	(0.00677)		(0.00903)	
Married	0.12325		0.14827	
	(0.03564)		(0.07314)	
Income	0.61812		0.31412	
	(0.05873)		(0.14664)	
Working	−0.05864		0.19407	
	(0.03297)		(0.12375)	
Education	0.05233		0.04755	
	(0.00588)		(0.01020)	
Kids	−0.10872	−0.17063	−0.00065	−0.23349
	(0.03306)	(0.01879)	(0.07519)	(04933)
Constant	−3.56368	−0.74006	−3.70407	−0.20658
	(0.08364)	(0.04094)	(0.16509)	(0.10440)
Age		0.02099		0.01431
		(0.00079)		(0.00214)
Female		0.42599		0.50918
		(0.01619)		(0.04400)
AddOn		−2.73847		−2.86428
		(0.04978)		(0.09289)
Sigma		1.43070		1.42112
		(0.00653)		(0.01866)
Rho		0.93299		0.99644
		(0.00754)		(0.00376)
ln *L*	−62366.61		−8313.88	
N	27,326,		3,377	

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. First, 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

- Attribute nonattendance
- Bivariate ordered probit
- Censoring
- Characteristics
- Choice-based sample
- Conditional logit model
- Count data
- Deviance
- Differential item functioning (DIF)
- Exposure
- Generalized mixed logit model
- Hurdle model
- Identification through functional form
- Inclusive value
- Independence from irrelevant alternatives (IIA)
- Limited information
- Log-odds
- Method of simulated moments
- Mixed logit model
- Multinomial choice
- Multinomial logit model
- Multinomial probit model (MNP)
- Negative binomial distribution
- Negative binomial model
- Negbin 1 (NB1) form
- Negbin 2 (NB2) form
- Negbin P (NBP) model
- Nested logit model
- Ordered choice
- Overdispersion
- Parallel regression assumption
- Random coefficients
- Random parameters logit model (RPL)
- Revealed preference data
- Specification error
- Stated choice data
- Stated choice experiment
- Subjective well-being (SWB)
- Unlabeled choices
- Unordered choice model
- Willingness to pay space

Exercises

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

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

   Using the preceding data, obtain maximum likelihood estimates of the unknown parameters of the model. (*Hint:* Consider the probabilities as the unknown parameters.)
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 F17.2 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 partial effects for the variables you choose. Compare the results you obtain for a probit model to those for a logit model. Are there any substantial differences in the results for the two models?

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 five-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, *NLOGIT*. 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 *DocVis*, 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 *HospVis*, 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 overdispersion 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 F18.3 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 type 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 overdispersion? Show your results.