18
Discrete Choices and Event Counts

18.1 Introduction

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

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

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

Event Counts: The observed outcome is a count of the number of occurrences. In many cases, this is similar to the preceding settings in that the "dependent variable" measures an individual choice, such as the number of visits to the physician or the hospital, the number of derogatory reports in one's credit history, or the number of visits to a particular recreation site. In other cases, the event count might be the outcome of some less focused natural process, such as incidence of a disease in a population or the number of defects per unit of time in a production process, the number of traffic accidents that occur at a particular location per month, or the number of messages that arrive at a 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 models with discrete outcomes, so the analysis is framed in terms of models of the probabilities attached to those outcomes.

18.2 MODELS FOR UNORDERED MULTIPLE CHOICES

Some studies of multiple-choice settings include the following:

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

In each of these cases, there is a single decision among two or more alternatives. In this and the next section, we will encounter two broad types of multinomial choice sets, **unordered choices models** 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 examined models for ordered choices in Section 18.3. This section will examine models for unordered choice sets. General references on the topics discussed here include Hensher, Louviere, and Swait (2000); Train (2009); and Hensher, Rose, and Greene (2006).

18.2.1 Random Utility Basis of the Multinomial Logit Model

These are all distinct from the multivariate probit model we examined earlier. In that setting, there were several decisions, each between two alternatives. Here there is a single decision among two or more alternatives. We will encounter two broad types of choice sets, **ordered choice models** and **unordered choice models.** The choice among means of getting to work—by car, bus, train, or bicycle—is clearly unordered. A bond rating is, by design, a ranking; that is its purpose. Quite different techniques are used for the two types of models. We examined models for ordered choices in Section 23.10. This section will examine models for unordered choice sets.[53] General references on the topics discussed here include Hensher, Louviere, and Swait (2000); Train (2003); and Hensher, Rose, and Greene (2006).

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

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

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

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

The model is made operational by a particular choice of distribution for the disturbances. As in the binary choice case, two models are usually considered, logit and probit. Because of the need to evaluate multiple integrals of the normal distribution, the probit model has found rather limited use in this setting. The logit model, in contrast, has been widely used in many fields, including economics, market research, politics, finance, and transportation engineering. Let Y_i be a random variable that indicates the choice made. McFadden (1974a) has shown that if (and only if) the J disturbances are independent and identically distributed with Gumbel (type 1 extreme value) distribution,

$$F(\varepsilon_{ij}) = \exp(-\exp(-\varepsilon_{ij})), \qquad (23\text{-}48) \quad 18\text{-}1$$

then

$$\text{Prob}(Y_i = j) = \frac{\exp(\mathbf{z}_{ij}'\theta)}{\sum_{j=1}^{J}\exp(\mathbf{z}_{ij}'\theta)}, \qquad (23\text{-}49) \quad 18\text{-}2$$

which leads to what is called the **conditional logit model.** (It is often labeled the **multinomial logit model,** but this wording conflicts with the usual name for the model discussed in the next section, which differs slightly. Although the distinction turns out to be purely artificial, we will maintain it for the present.)

Utility depends on $\mathbf{z}_{ij}$, which includes aspects specific to the individual as well as to the choices. It is useful to distinguish them. Let $\mathbf{z}_{ij} = [\mathbf{x}_{ij}, \mathbf{w}_i]$ and partition θ conformably into $[\beta', \alpha']'$. Then $\mathbf{x}_{ij}$ varies across the choices and possibly across the individuals as well. The components of $\mathbf{x}_{ij}$ are typically called the **attributes** of the choices. But $\mathbf{w}_i$ contains the **characteristics** of the individual and is, therefore, the same

[53] A hybrid case occurs in which consumers reveal their own specific ordering for the choices in an unordered choice set. Beggs, Cardell, and Hausman (1981) studied consumers' rankings of different automobile types, for example.

Ay! These two KTs are not in chap list

for all choices. If we incorporate this fact in the model, then (18-2) becomes

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

Terms that do not vary across alternatives—that is, those specific to the individual—fall out of the probability. This is as expected in a model that compares the utilities of the alternatives. For example, in a model of a shopping center choice by individuals in various cities that depends on the number of stores at the mall, S_{ij}, the distance from the central business district, D_{ij} and the shoppers' incomes, I_i, the utilities for three choices would be

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

The choice of alternative 1, for example, reveals that

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

The constant term and *Income* have fallen out of the comparison. The result follows from the fact that random utility model is ultimately based on comparisons of pairs of alternatives, not the alternatives themselves. Evidently, if the model is to allow individual specific effects, then it must be modified. One method is to create a set of dummy variables (alternative specific constants), A_j , for the choices and multiply each of them by the common $\mathbf{w}$. 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

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

18-2

for all choices. If we incorporate this fact in the model, then (23-49) becomes

$$\text{Prob}(Y_i = j) = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\alpha})}{\sum_{j=1}^{J}\exp(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\alpha})} = \frac{[\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})]\exp(\mathbf{w}_i'\boldsymbol{\alpha})}{\left[\sum_{j=1}^{J}\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})\right]\exp(\mathbf{w}_i'\boldsymbol{\alpha})}. \quad (23\text{-}50)$$

18-3

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

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

The probabilities for this model would be

$$\text{Prob}(Y_i = j) = \frac{\exp(\beta_1 S_{ij} + \beta_2 D_{ij} + \alpha_1 A_1\, Income_i + \alpha_2 A_2\, Income_i + \alpha_3 A_3\, Income_i)}{\sum_{j=1}^{3}\exp(\beta_1 S_{ij} + \beta_2 D_{ij} + \alpha_1 A_1\, Income_i + \alpha_2 A_2\, Income_i + \alpha_3 A_3\, Income_i)}, \quad \alpha_3 = 0.$$

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

18.2.2 ~~23.11.1~~ THE MULTINOMIAL LOGIT MODEL

To set up the model that applies when data are individual specific, it will help to consider an example. Schmidt and Strauss (1975a, b) estimated a model of occupational choice based on a sample of 1,000 observations drawn from the Public Use Sample for three years: 1960, 1967, and 1970. For each sample, the data for each individual in the sample consist of the following:

1. *Occupation:* 0 = menial, 1 = blue collar, 2 = craft, 3 = white collar, 4 = professional. (Note the slightly different numbering convention, starting at zero, which is standard.)
2. *Characteristics:* constant, education, experience, race, sex.

The model for occupational choice is

$$\text{Prob}(Y_i = j \mid \mathbf{w}_i) = \frac{\exp(\mathbf{w}_i'\boldsymbol{\alpha}_j)}{\sum_{j=0}^{4}\exp(\mathbf{w}_i'\boldsymbol{\alpha}_j)}, \quad j = 0, 1, \ldots, 4. \quad (23\text{-}51)$$

18-4

(The binomial logit model in Section 23.3 is conveniently produced as the special case of $J = 1$.)

The model in (23-51) is a **multinomial logit model**.[34] The estimated equations provide a set of probabilities for the $J + 1$ choices for a decision maker with characteristics $\mathbf{w}_i$. Before proceeding, we must remove an indeterminacy in the model. If we define $\boldsymbol{\alpha}_j^* = \boldsymbol{\alpha}_j + \mathbf{q}$ for any vector $\mathbf{q}$, then recomputing the probabilities defined later using $\boldsymbol{\alpha}_j^*$ instead of $\boldsymbol{\alpha}_j$ produces the identical set of probabilities because all the terms involving $\mathbf{q}$ drop out. A convenient normalization that solves the problem is $\boldsymbol{\alpha}_0 = \mathbf{0}$. (This arises because the probabilities sum to one, so only J parameter vectors are needed to determine the $J + 1$ probabilities.) Therefore, the probabilities are

$$\text{Prob}(Y_i = j \mid \mathbf{w}_i) = P_{ij} = \frac{\exp(\mathbf{w}_i'\boldsymbol{\alpha}_j)}{1 + \sum_{k=1}^{J} \exp(\mathbf{w}_i'\boldsymbol{\alpha}_k)}, \quad j = 0, 1, \ldots, J, \quad \boldsymbol{\alpha}_0 = \mathbf{0}. \tag{23-52}$$

The form of the binomial model examined in Section 23.4 results if $J = 1$. The model implies that we can compute J **log-odds ratios**

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

From the point of view of estimation, it is useful that the odds ratio, P_{ij}/P_{ik}, does not depend on the other choices, which follows from the independence of the disturbances in the original model. From a behavioral viewpoint, this fact is not very attractive. We shall return to this problem in Section 23.11.3.

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

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

The derivatives have the characteristically simple form

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

The exact second derivatives matrix has $J^2K \times K$ blocks,[35]

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

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

[34] Nerlove and Press (1973).

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

with the number of choices, which is inconvenient because the typical cross section sometimes involves a fairly large number of regressors.

The coefficients in this model are difficult to interpret. It is tempting to associate α_j with the jth outcome, but that would be misleading. By differentiating (23-52), we find that the marginal effects of the characteristics on the probabilities are

$$\delta_{ij} = \frac{\partial P_{ij}}{\partial \mathbf{w}_i} = P_{ij}\left[\alpha_j - \sum_{k=0}^{J} P_{ik}\alpha_k\right] = P_{ij}[\alpha_j - \bar{\alpha}]. \quad (23\text{-}53)$$

Therefore, every subvector of α enters every marginal effect, both through the probabilities and through the weighted average that appears in δ_{ij}. These values can be computed from the parameter estimates. Although the usual focus is on the coefficient estimates, equation (23-53) suggests that there is at least some potential for confusion. Note, for example, that for any particular w_{ik}, $\partial P_{ij}/\partial w_{ik}$ need not have the same sign as α_{jk}. Standard errors can be estimated using the delta method. (See Section 4.9.4.) For purposes of the computation, let $\alpha = [\mathbf{0}, \alpha_1', \alpha_2', \ldots, \alpha_J']'$. We include the fixed $\mathbf{0}$ vector for outcome 0 because although $\alpha_0 = \mathbf{0}$, $\delta_{i0} = -P_{i0}\bar{\alpha}$, which is not $\mathbf{0}$. Note as well that Asy. $\text{Cov}[\hat{\alpha}_0, \hat{\alpha}_j] = \mathbf{0}$ for $j = 0, \ldots, J$. Then

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

$$\frac{\partial \delta_{ij}}{\partial \alpha_l'} = [\mathbf{1}(j=l) - P_{il}][P_{ij}\mathbf{1} + \delta_{ij}\mathbf{w}_i'] + P_{ij}[\delta_{il}\mathbf{w}_i'].$$

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

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

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

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

where p_j is the sample proportion of observations that make choice j. A useful table will give a listing of hits and misses of the prediction rule "predict $Y_i = j$ if $\hat{P}_{ij}$ is the maximum of the predicted probabilities."[36]

[36] It is common for this rule to predict all observation with the same value in an unbalanced sample or a model with little explanatory power. This is not a contradiction of an estimated model with many "significant" coefficients, because the coefficients are not estimated so as to maximize the number of correct predictions.

Example 18.1 Hollingshead Scale of Occupations

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

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

Among the other variables in the data set are *Age*, *Sex* and *Education*. The data are given in Appendix Table F18.1. Table 18.1 lists estimates of a multinomial logit model. (We emphasize that the data are a self selected sample of *Psychology Today* readers in 1976, so it is unclear what contemporary population would be represented. The following serves as an uncluttered numerical example that readers could reproduce. Note, as well, that at least by some viewpoint, the outcome for this experiment is ordered.) The log likelihood for the model is -770.28141 while that for the model with only the constant terms is -982.20533. The likelihood ratio statistic for the hypothesis that all 18 coefficients of the model are zero is 423.85, which is far larger than the critical value of 28.87. In the estimated parameters, it appears that only gender is consistently statistically significant. However, it is unclear how 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 α_6 is. This is an aspect of modeling with multinomial choice models that calls for careful interpretation by the model builder.

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

18.2.3 THE CONDITIONAL LOGIT MODEL

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

$$\text{Prob}(Y_i = j \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iJ}) = \text{Prob}(Y_i = j \mid \mathbf{X}_i) = P_{ij} = \frac{\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}{\sum_{j=1}^{J}\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})}. \quad (23\text{-}54)\ 18\text{-}7$$

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

In this model, the coefficients are not directly tied to the marginal effects. The marginal effects for continuous variables can be obtained by differentiating (23-54) with respect to a particular $\mathbf{x}_m$ to obtain 18-7

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

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

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

Because there is no ambiguity about the scale of the probability itself, whether one should report the derivatives or the elasticities is largely a matter of taste. Some of Hensher's elasticity estimates are given in Table 23.20 later on in this chapter.

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

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

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

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

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

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n}\sum_{j=1}^{J} P_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)', \quad (23\text{-}55)\ 18\text{-}8$$

The usual problems of fit measures appear here. The log-likelihood ratio and tabulation of actual versus predicted choices will be useful. There are two possible constrained log-likelihoods. The model cannot contain a constant term, so the constraint $\boldsymbol{\beta} = \mathbf{0}$ renders all probabilities equal to $1/J$. The constrained log-likelihood for this constraint

is then $L_c = -n \ln J$. Of course, it is unlikely that this hypothesis would fail to be rejected. Alternatively, we could fit the model with only the $J-1$ choice-specific constants, which makes the constrained log-likelihood the same as in the multinomial logit model, $\ln L_0^* = \sum_j n_j \ln p_j$ where, as before, n_j is the number of individuals who choose alternative j.

18.2.4

~~23.11.3~~ THE INDEPENDENCE FROM IRRELEVANT ALTERNATIVES ASSUMPTION

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

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

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

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.[57]

18.2.5

~~23.11.4~~ NESTED LOGIT MODELS

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

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

We had considered this model earlier but found that as a general model of consumer choice, its failings were the practical difficulty of computing the multinormal integral and estimation of an unrestricted correlation matrix. Hausman and Wise (1978) point out that for a model of consumer choice, the probit model may not be as impractical as it might seem. First, for J choices, the comparisons implicit in $U_{ij} > U_{im}$ for $m \neq j$ involve the $J-1$ differences, $\varepsilon_j - \varepsilon_m$. Thus, starting with a J-dimensional problem, we need only consider derivatives of $(J-1)$-order probabilities. Therefore, to come to a concrete example, a model with four choices requires only the evaluation of bivariate

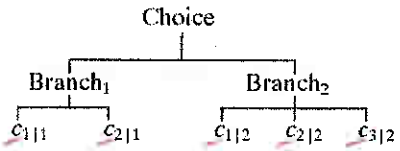
[57] McFadden (1987) shows how this hypothesis can also be tested using a Lagrange multiplier test.

normal integrals, which, albeit still complicated to estimate, is well within the received technology. ~~We will examine the multivariate probit model in Section 23.11.3.~~ For larger models, however, other specifications have proved more useful.

One way to relax the homoscedasticity assumption in the conditional logit model that also provides an intuitively appealing structure is to group the alternatives into subgroups that allow the variance to differ across the groups while maintaining the IIA assumption within the groups. This specification defines a **nested logit model.** To fix ideas, it is useful to think of this specification as a two- (or more) level choice problem (although, once again, the model arises as a modification of the stochastic specification in the original conditional logit model, not necessarily as a model of behavior). Suppose, then, that the J alternatives can be divided into B subgroups (branches) such that the choice set can be written

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

Logically, we may think of the choice process as that of choosing among the B choice sets and then making the specific choice within the chosen set. This method produces a tree structure, which for two branches and, say, five choices (twigs) might look as follows:



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

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

where the new parameters τ_l must equal 1 to produce the original model. Therefore, we use the restriction $\tau_l = 1$ to recover the conditional logit model, and the preceding equation just writes this model in another form. The nested logit model arises if this restriction is relaxed. The inclusive value coefficients, unrestricted in this fashion, allow the model to incorporate some degree of heteroscedasticity. Within each branch, the IIA restriction continues to hold. The equal variance of the disturbances within the jth branch are now[5]

$$\sigma_b^2 = \frac{\pi^2}{6\tau_b}. \qquad (18\text{-}10)$$

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

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

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

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 β 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 γ and the τ parameters by treating the choice among branches as a conditional logit model with attributes z_{ib} and I_{ib}.

Because this approach is a two-step estimator, the estimate of the asymptotic covariance matrix of the estimates at the second step must be corrected. [See Section 16.7 and McFadden (1984).] For **full information maximum likelihood** (FIML) estimation of the model, the log-likelihood is

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

[See Hensher (1986, 1991) and Greene (2007a).] The information matrix is not block diagonal in β and (γ, τ), so FIML estimation will be more efficient than two-step estimation. The FIML estimator is now available in several commercial computer packages. The two-step estimator is rarely used in current research.

To specify the nested logit model, it is necessary to partition the choice set into branches. Sometimes there will be a natural partition, such as in the example given by Maddala (1983) when the choice of residence is made first by community, then by

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

dwelling type within the community. In other instances, however, the partitioning of the choice set is ad hoc and leads to the troubling possibility that the results might be dependent on the branches so defined. (Many studies in this literature present several sets of results based on different specifications of the tree structure.) There is no well-defined testing procedure for discriminating among tree structures, which is a problematic aspect of the model.

18.2.6 ~~23.11.5~~ THE MULTINOMIAL PROBIT MODEL

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

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

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

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

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 23.2.) For more than two choices, this specification is far more general than the MNL model, which assumes that $\boldsymbol{\Sigma} = \mathbf{I}$. (The scaling is absorbed in the coefficient vector in the MNL model.) It adds the unrestricted correlations to the heteroscedastic model of the previous section.

17.2

The main obstacle to implementation of the MNP model has been the difficulty in computing the multivariate normal probabilities for any dimensionality higher than 2. Recent results on accurate simulation of multinormal integrals, however, have made estimation of the MNP model feasible. (See Section 17.3.3 and a symposium in the November 1994 issue of the *Review of Economics and Statistics*.) Yet some practical problems remain. Computation is exceedingly time consuming. It is also necessary to ensure that $\boldsymbol{\Sigma}$ remain a positive definite matrix. One way often suggested is to construct the Cholesky decomposition of $\boldsymbol{\Sigma}$, $\mathbf{LL}'$, where $\mathbf{L}$ is a lower triangular matrix, and estimate the elements of $\mathbf{L}$. 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 $\mathbf{L}_{11} = 1$. This is straightforward to implement for an otherwise unrestricted $\boldsymbol{\Sigma}$. A remaining problem, however, is that it is now difficult by this method to impose any other restrictions, such as a zero in a specific location in $\boldsymbol{\Sigma}$, which is common. An alternative approach is estimate the correlations, $\mathbf{R}$, and a diagonal matrix of standard deviations, $\mathbf{S} = \text{diag}(\sigma_1, \ldots, \sigma_{J-2}, 1, 1)$, separately. The normalizations, $\mathbf{R}_{jj} = 1$ and

15.6.2.b

exclusions, $\mathbf{R}_{jl} = 0$, are then simple to impose, and $\boldsymbol{\Sigma}$ is just SRS. The resulting matrix must still be symmetric and positive definite. The restriction $-1 < \mathbf{R}_{jl} < +1$ is necessary but still not sufficient to ensure definiteness. The full set of restrictions is difficult to enumerate explicitly and involves sets of inequalities that would be difficult to impose in estimation. (Typically when this method is employed, the matrix is estimated without the explicit restrictions.) Identification appears to be a serious problem with the MNP model. Although the unrestricted MNP model is fully identified in principle, convergence to satisfactory results in applications with more than three choices appears to require many additional restrictions on the standard deviations and correlations, such as zero restrictions or equality restrictions in the case of the standard deviations.

AG: Run-in or new ¶? Please indicate

18.2.7 ~~23.11.8~~ THE MIXED LOGIT MODEL

Another variant of the multinomial logit model is the **random parameters logit model** (RPL) (also called the **mixed logit model**). [See Revelt and Train (1996); Bhat (1996); Berry, Levinsohn, and Pakes (1995); Jain, Vilcassim, and Chintagunta (1994); and Hensher and Greene (2004).] Train's (2003) formulation of the RPL model (which encompasses the others) is a modification of the MNL model. The model is a **random coefficients** formulation. The change to the basic MNL model is the parameter specification in the distribution of the parameters across individuals, i:

$$\beta_{ik} = \beta_k + \mathbf{z}_i'\boldsymbol{\theta}_k + \sigma_k u_{ik}, \qquad \text{18-9 // (23-56)}$$

where $u_{ik}, k = 1, \ldots, K$, is multivariate normally distributed with correlation matrix $\mathbf{R}$, σ_k is the standard deviation of the kth distribution, $\beta_k + \mathbf{z}_i'\boldsymbol{\theta}_k$ is the mean of the distribution, and $\mathbf{z}_i$ is a vector of person specific characteristics (such as age and income) that do not vary across choices. This formulation contains all the earlier models. For example, if $\boldsymbol{\theta}_k = \mathbf{0}$ for all the coefficients and $\sigma_k = 0$ for all the coefficients except for choice-specific constants, then the original MNL model with a normal-logistic mixture for the random part of the MNL model arises (hence the name).

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

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

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

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

which can be estimated by simulation, using

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

where $\mathbf{u}_{ir}$ is the rth of R draws for observation i. (There are nkR draws in total. The draws for observation i must be the same from one computation to the next, which can be accomplished by assigning to each individual their own seed for the random number generator and restarting it each time the probability is to be computed.) By this

method, the log-likelihood and its derivatives with respect to $(\beta_k, \theta_k, \sigma_k)$, $k = 1, \ldots, K$ and $\mathbf{R}$ are simulated to find the values that maximize the simulated log-likelihood. ~~(See Section 17.5 and Example 23.8.)~~

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

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

The time variation in the coefficients is provided by the choice-invariant variables, which may change through time. Habit persistence is carried by the time-invariant random effect, u_{ik}. If only the constant terms vary and they are assumed to be uncorrelated, then this is logically equivalent to the familiar random effects model. But, much greater generality can be achieved by allowing the other coefficients to vary randomly across individuals and by allowing correlation of these effects.[59] A second degree of flexibility is in (~~23-56~~). 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.

~~23.11.7~~ APPLICATION: CONDITIONAL LOGIT MODEL FOR TRAVEL MODE CHOICE

Hensher and Greene [Greene (2007a)] report estimates of a model of travel mode choice for travel between Sydney and Melbourne, Australia. The data set contains 210 observations on choice among four travel modes, *air*, *train*, *bus*, and *car*. (See Appendix Table ~~F23.2.~~) The attributes used for their example were: choice-specific constants; two choice-specific continuous measures; *GC*, a measure of the generalized cost of the travel that is equal to the sum of in-vehicle cost, *INVC*, and a wagelike measure times *INVT*, the amount of time spent traveling; and *TTME*, the terminal time (zero for car); and for the choice between air and the other modes, *HINC*, the household income. A summary of the sample data is given in Table ~~23.23~~. 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 ~~23.23~~, is dominated by drivers.

The model specified is

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

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

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

which has standard deviation $\pi^2/6$. The mean is absorbed in the constants. Estimates of the conditional logit model are shown in Table ~~23.24.~~ The model was fit with and

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

18.2.8 A Generalized Mixed Logit Model

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

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

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

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

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

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

18.2

TABLE ~~23.23~~ Summary Statistics for Travel Mode Choice Data

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

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

18.3

TABLE ~~23.24~~ Parameter Estimates

	Unweighted Sample		*Choice-Based Weighting*	
	Estimate	*t Ratio*	*Estimate*	*t Ratio*
β_G	−0.15501	−3.517	−0.01333	~~−2.724~~
β_T	−0.09612	−9.207	−0.13405	~~−7.164~~
γ_H	0.01329	1.295	−0.00108	−0.087
α_{air}	5.2074	6.684	6.5940	~~5.906~~
α_{train}	3.8690	8.731	3.6190	~~7.447~~
α_{bus}	3.1632	7.025	3.3218	~~5.698~~
Log-likelihood at $\beta = 0$	−291.1218		−291.1218	
Log-likelihood (sample shares)	−283.7588		~~−223.0578~~	
Log-likelihood at convergence	−199.1284		−147.5896	

without the corrections for choice-based sampling. Because the sample shares do not differ radically from the population proportions, the effect on the estimated parameters is fairly modest. Nonetheless, it is apparent that the choice-based sampling is not completely innocent. A cross tabulation of the predicted versus actual outcomes is given in Table ~~23.25~~. The predictions are generated by tabulating the integer parts of $m_{jk} = \sum_{i=1}^{210} \hat{p}_{ij} d_{ik}$, j, k = *air, train, bus, car,* where $\hat{p}_{ij}$ is the predicted probability of outcome j for observation i and d_{ik} is the binary variable which indicates if individual i made choice k.

Are the odds ratios *train/bus* and *car/bus* really independent from the presence of the *air* alternative? To use the Hausman test, we would eliminate choice *air*, from the

18.4

TABLE ~~23.25~~ Predicted Choices Based on Model Probabilities (predictions based on choice-based sampling in parentheses)

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

18.5

TABLE ~~23.26~~ Results for IIA Test

	Full-Choice Set				*Restricted-Choice Set*			
	β_G	β_T	α_{train}	α_{bus}	β_G	β_T	α_{train}	α_{bus}
Estimate	−0.0155	−0.0961	3.869	3.163	−0.0639	−0.0699	4.464	3.105
	Estimated Asymptotic Covariance Matrix				*Estimated Asymptotic Covariance Matrix*			
β_G	0.194e-04				0.000101			
β_T	−0.46e-06	0.000109			−0.0000013	0.000221		
α_{train}	−0.00060	−0.0038	0.196		−0.000244	−0.00759	0.410	
α_{bus}	−0.00026	−0.0037	0.161	0.203	−0.000113	−0.00753	0.336	0.371

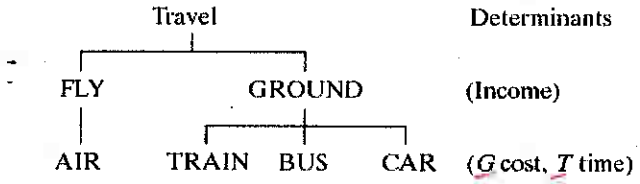
Note: 0.nnne-*p* indicates times 10 to the negative *p* power.
$H = 33.3363$. Critical chi-squared[4] = 9.488.

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 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 ~~23.26~~ 18.5

TB 18.5

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

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



TB 18.6

Note that one of the branches has only a single choice, so the conditional probability, $P_{j|fly} = P_{air|fly} = 1$. The estimates marked "unconditional" in Table ~~23.27~~ 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,

18.6

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

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

18.6

TABLE 23.27 Estimates of a Mode Choice Model (standard errors in parentheses)

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

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 the heteroscedastic extreme value (HEV) model. [See Greene (2007b).] This model allows a separate variance, $\sigma_j^2 = \pi^2/(6\theta_j^2)$ for each ε_{ij} in (23.48). The results are shown in Table 23.28. This model is

18.7

TABLE 23.28 Estimates of a Heteroscedastic Extreme Value Model (standard errors in parentheses)

Parameter	*HEV Model*		*Heteroscedastic HEV Model*		*Restricted HEV Model*		*Nested Logit Model*	
α_{air}	7.8326	(10.951)	5.1815	(6.042)	2.973	(0.995)	6.062	(1.199)
α_{bus}	7.1718	(9.135)	5.1302	(5.132)	4.050	(0.494)	4.096	(0.615)
α_{train}	6.8655	(8.829)	4.8654	(5.071)	3.042	(0.429)	5.065	(0.662)
β_{GC}	−0.05156	(0.0694)	−0.03326	(0.0378)	−0.0289	(0.00580)	−0.03159	(0.00816)
β_{TTME}	−0.1968	(0.288)	−0.1372	(0.164)	−0.0828	(0.00576)	−0.1126	(0.0141)
γ	0.04024	(0.0607)	0.03557	(0.0451)	0.0238	(0.0186)	0.01533	(0.00938)
τ_{fly}							0.5860	(0.141)
τ_{ground}							0.3890	(0.124)
θ_{air}	0.2485	(0.369)	0.2890	(0.321)	0.4959	(0.124)		
θ_{train}	0.2595	(0.418)	0.3629	(0.482)	1.0000	(0.000)		
θ_{bus}	0.6065	(1.040)	0.6895	(0.945)	1.0000	(0.000)		
θ_{car}	1.0000	(0.000)	1.0000	(0.000)	1.0000	(0.000)		
ϕ	0.0000	(0.000)	0.00552	(0.00573)	0.0000	(0.000)		
	Implied Standard Deviations							
σ_{air}	5.161	(7.667)						
σ_{train}	4.942	(7.978)						
σ_{bus}	2.115	(3.623)						
σ_{car}	1.283	(0.000)						
ln L	−195.6605		−194.5107		−200.3791		−193.6561	

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

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

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 Tble 18.7. This model is less restrictive than the nested logit model. To make them comparable, we note that we found that $\sigma_{air} = \pi/(\tau_{fly}\sqrt{6}) = 2.1886$ and $\sigma_{train} = \sigma_{bus} = \sigma_{car} = \pi/(\tau_{ground}\sqrt{6}) = 3.2974$. The HEV model thus relaxes an additional restriction because it has three free variances whereas the nested logit model has two. On the other hand, the important degree of freedom is that the HEV model does not impose the IIA assumptions anywhere in the choice se, whereas the nexted logit does, within each branch. Table 18.7 contains two additional results for HEV specifications. In the one denoted "Heteroscedastic HEV Model," we have allowed heteroscedasticity across individuals as well as across choices by specifying

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

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

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

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 heteroscedastic extreme value (HEV) model thus relaxes one variance restriction, because it has three free variance parameters instead of two. On the other hand, the important degree of freedom here is that the HEV model does not impose the IIA assumption anywhere in the choice set, whereas the nested logit does, within each branch.

A primary virtue of the HEV model, the nested logit model, and other alternative models is that they relax the IIA assumption. This assumption has implications for the cross elasticities between attributes in the different probabilities. Table ~~23.29~~ 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.

TB 18.8

18.9

Table ~~23.29~~ 18.9 lists the estimates of the parameters of the multinomial probit and random parameters logit models. For the multinomial probit model, we fit three specifications: (1) free correlations among the choices, which implies an unrestricted 3×3 correlation matrix and two free standard deviations; (2) uncorrelated disturbances, but free standard deviations, a model that parallels the heteroscedastic extreme value model; and (3) uncorrelated disturbances and equal standard deviations, a model that is the same as the original conditional logit model save for the normal distribution of the disturbances instead of the extreme value assumed in the logit model. In this case,

TB 18.9

18.8

TABLE ~~23.29~~ Estimated Elasticities with Respect to Generalized Cost

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

the scaling of the utility functions is different by a factor of $(\pi^2/6)^{1/2} = 1.283$, as the probit model assumes ε_j has a standard deviation of 1.0.

We also fit three variants of the random parameters logit. In these cases, the choice-specific variance for each utility function is $\sigma_j^2 + \theta_j^2$ where σ_j^2 is the contribution of the logit model, which is $\pi^2/6 = 1.645$, and θ_j^2 is the estimated constant specific variance estimated in the random parameters model. The combined estimated standard deviations are given in the table. The estimates of the specific parameters, θ_j, are given in the footnotes. The estimated models are (1) unrestricted variation and correlation among the three intercept parameters—this parallels the general specification of the multinomial probit model; (2) only the constant terms randomly distributed but uncorrelated, a model that is parallel to the multinomial probit model with no cross-equation correlation and to the heteroscedastic extreme value model shown in Table ~~23.28~~; and (3) random but uncorrelated parameters. This model is more general than the others, but is somewhat restricted as the parameters are assumed to be uncorrelated. Identification of the correlation matrix is weak in this model—after all, we are attempting to estimate a 6 × 6 correlation matrix for all unobserved variables. Only the estimated parameters are shown in Table ~~23.30~~. Estimated standard errors are similar to (although generally somewhat larger than) those for the basic multinomial logit model.

18.7

18.9

18.9

The standard deviations and correlations shown for the multinomial probit model are parameters of the distribution of ε_{ij}, the overall randomness in the model. The counterparts in the random parameters model apply to the distributions of the parameters. Thus, the full disturbance in the model in which only the constants are random is $\varepsilon_{iair} + u_{air}$ for air, and likewise for train and bus. Likewise, the correlations shown

TABLE ~~23.30~~ 18.9 Parameter Estimates for Normal-Based Multinomial Choice Models

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

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

for the first two models are directly comparable, although it should be noted that in the random parameters model, the disturbances have a distribution that is that of a sum of an extreme value and a normal variable, while in the probit model, the disturbances are normally distributed. With these considerations, the "unrestricted" models in each case are comparable and are, in fact, fairly similar.

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

18.2.9 ~~18.11.8.~~ PANEL DATA AND STATED CHOICE EXPERIMENTS

Panel data in the unordered discrete choice setting typically come in the form of sequential choices. Train (~~2003~~ 2009, Chapter 6) reports an analysis of the site choices of 258 anglers who chose among 59 possible fishing sites for total of 962 visits. Allenby and Rossi (1999) modeled brand choice for a sample of shoppers who made multiple store trips. The mixed logit model is a framework that allows the counterpart to a random effects model. The random utility model would appear

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

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

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

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

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

18.2.10 Estimating Willingness to Pay

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

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

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

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

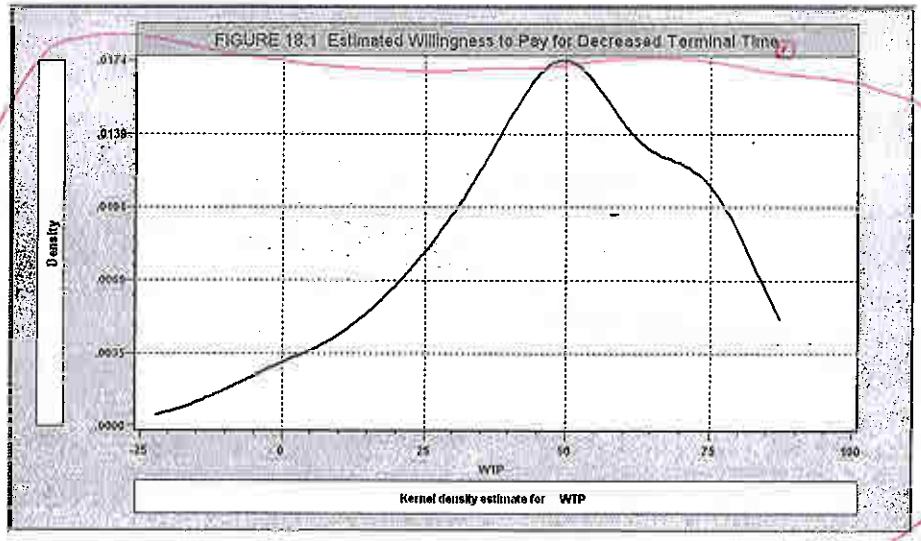
(We use the negative because additional time spent waiting at the terminal provides disutility, as evidenced by its coefficient's negative sign.) In settings in which income is not observed, researchers often use the negative of the coefficient on a cost variable as a proxy for the marginal utility of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

In the basic multinomial logit model, the estimator of WTP is a simple ratio of parameters. In our estimated model in Table 18.3, for example, using the household income coefficient as the numeraire, the estimate of WTP for a shorter wait at the terminal is -0.09612/0.01329 = 7.239. The units of measurement must be resolved in this computation, since terminal time is measured in minutes while the cost is in \$1000/year. Multiplying this result by \$60 minutes/hour and dividing by the equivalent hourly income of income times 8760/1000 gives \$49.54 per hour of waiting time. To compute the estimated asymptotic standard error, for convenience, we first rescaled the terminal time to hours by dividing it by 60 and the income variable to \$/hour by multiplying it by 1000/8760. The resulting estimated asymptotic distribution for the estimators is

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

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

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



Willingness to pay measures computed as suggested above are ultimately based on a ratio of two asymptotically normally distributed parameter estimators. In general, ratios of normally distributed random variables do not have a finite variance. This often becomes apparent when using the delta method, as it seems above. 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, e.g., by Train and Weeks (2005), Sonnier, Ainsle and Otter (2007), and Scarpa, Thiene and Train (2008), is to recast the original model in "**willingness to pay space.**" In the multinomial logit case, this amounts to a trivial reparameterization of the model. Using our application as an example, we would write

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

This obviously returns the original model, though in the process, it transforms a linear estimation problem into a nonlinear one. But, in principle, with the model reparameterized in "WTP space," we have sidestepped the problem noted earlier – λ_{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 $\lambda_{TTME,i} = \lambda_{TTME} + \theta_{TTME} v_{TTME,i}$. Greene and Hensher (2010b) apply this method to the generalized mixed logit model in Section 18.2.8.

18-26

for the first two models are directly comparable, although it should be noted that in the random parameters model, the disturbances have a distribution that is that of a sum of an extreme value and a normal variable, while in the probit model, the disturbances are normally distributed. With these considerations, the "unrestricted" models in each case are comparable and are, in fact, fairly similar.

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

18.2.9 ~~18.11.8~~ PANEL DATA AND STATED CHOICE EXPERIMENTS

Panel data in the unordered discrete choice setting typically come in the form of sequential choices. Train (~~2003~~ 2009, Chapter 6) reports an analysis of the site choices of 258 anglers who chose among 59 possible fishing sites for a total of 962 visits. Allenby and Rossi (1999) modeled brand choice for a sample of shoppers who made multiple store trips. The mixed logit model is a framework that allows the counterpart to a random effects model. The random utility model would appear

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

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

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

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

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

preferences is to build different scaling parameters into the model for the stated and revealed preference components of the model. Greene and Hensher (2007) suggest a nested logit model that groups the hypothetical choices in one branch of a tree and the observed choices in another.

23.12 SUMMARY AND CONCLUSIONS

This chapter has surveyed techniques for modeling discrete choice. We examined three classes of models: binary choice, ordered choice, and multinomial choice. These are quite far removed from the regression models (linear and nonlinear) that have been the focus of the preceding 22 chapters. The most important difference concerns the modeling approach. Up to this point, we have been primarily interested in modeling the conditional mean function for outcomes that vary continuously. In this chapter, we have shifted our approach to one of modeling the conditional probabilities of events.

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

Key Terms and Concepts

- Attributes
- Binary choice model
- Bivariate ordered probit
- Bivariate probit
- Bootstrapping
- Butler and Moffitt method
- Characteristics
- Choice-based sampling
- Chow test
- Conditional likelihood function
- Conditional logit model
- Control function
- Discriminant analysis
- Fixed effects model
- Full information maximum likelihood (FIML)
- Generalized residual
- Goodness of fit measure
- Gumbel model
- Heterogeneity
- Heteroscedasticity
- Incidental parameters problem
- Inclusive value
- Independence from irrelevant alternatives
- Index function model
- Initial conditions
- Kernel density estimator
- Kernel function
- Lagrange multiplier test
- Latent regression
- Likelihood equations
- Likelihood ratio test
- Limited information ML
- Linear probability model
- Logit
- Log-odds ratios
- Marginal effects
- Maximum likelihood
- Maximum score estimator
- Maximum simulated likelihood
- Mean-squared deviation
- Method of kernels
- Method of scoring
- Minimal sufficient statistic
- Mixed logit model
- Multinomial logit model
- Multinomial probit model
- Multivariate probit
- Negative binomial model

18.2.12 Aggregate Market Share Data – The BLP Random Parameters Model

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

Markets, denoted $t = 1,...,T$;
Consumers in the markets, denoted $i = 1,...,n_t$;
Products, denoted $j = 1,...,J$.

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

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

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

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

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

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

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

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),

$$u_{ijt} = \alpha_i y_i + \delta_{jt}(\mathbf{x}_{jt}, p_{jt}, \Delta_{jt}; \alpha, \beta) + \tau_{ijt}(\mathbf{x}_{jt}, p_{jt}, \mathbf{v}_i, w_i; \pi, \Pi, \gamma, \Gamma) + \varepsilon_{ijt}$$

$$\delta_{jt} = \mathbf{x}_{jt}'\beta - \alpha p_{jt} + \Delta_{jt}$$

$$\tau_{jt} = [-p_{jt}, \mathbf{x}_{jt}']\left[\begin{pmatrix}\pi' \\ \Pi\end{pmatrix} d_i + \begin{pmatrix}\gamma w_i \\ \Gamma \mathbf{v}_i\end{pmatrix}\right].$$

The preceding 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 ($\Gamma = \mathbf{0}$ 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

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

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

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

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

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

These would be simulated using the method of Section 18.2.7.

18.3 Random Utility Models for Ordered Choices

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

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

For another example, *Netflix* (www.netflix.com) is an internet company that rents movies. Subscribers order the film online for download or home delivery of a DVD. The next time the customer logs on to the website, they are invited to rate the movie on a five point scale, where five is the highest, most favorable rating. The ratings of the many thousands of subscribers who rented that movie are averaged to provide a recommendation to prospective viewers. As of April 5, 2009, the average rating of the 2007 movie *National Treasure: Book of Secrets* given by approximately 12,900 visitors to the site was 3.8. Many other internet sellers of product and services, such as Barnes and Noble, Amazon, Hewlett Packard and Best Buy, employ ratings schemes such as this. Many recently developed national survey data sets, such as the British Household Panel Data Set (http://www.iser.essex.ac.uk/survey/bhps) (BHPS) and the German Socioeconomic Panel (http://www.diw.de/en/soep) (GSOEP) contain questions that elicit self assessed ratings of health, health satisfaction, or overall well being. Like the other examples listed, these survey questions are answered on a discrete scale, such as the zero to ten scale of the question about health satisfaction in the GSOEP. Ratings such as these provides applications of the models and methods that interest us in this section.[8]

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

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

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

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

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

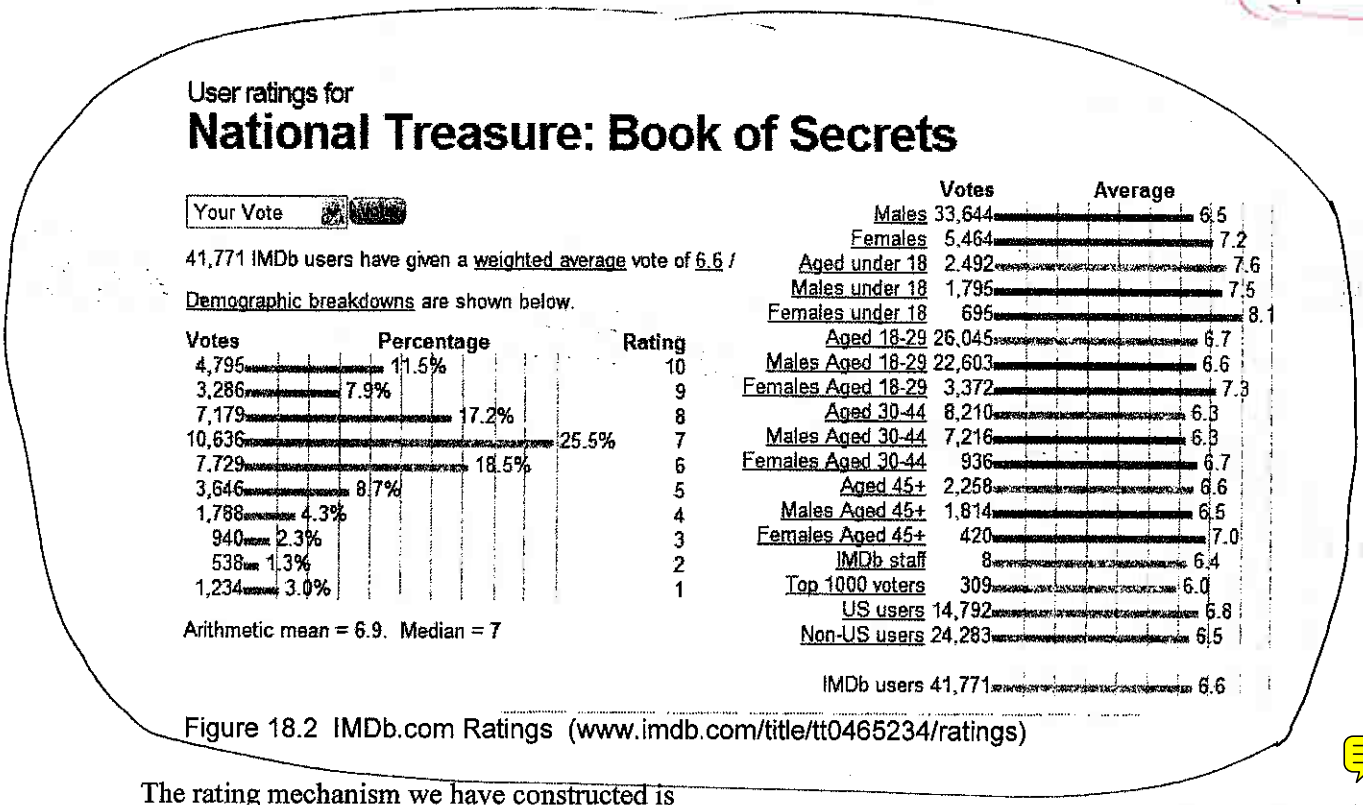
The same mapping would characterize the bond ratings, since the qualities of bonds that produce the ratings will vary continuously and the self assessed health and well being questions in the panel survey data sets based on an underlying utility or preference structure. The crucial feature of the description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

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

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

Example 18.2 Movie Ratings

The website www.imdb.com invites visitors to rate movies that they have seen, in the same fashion as the ne site. This site uses a ten point scale. On December 1, 2008, they reported the following results for the movie noted above for 41,771 users of the site: The panel at the left below shows the overall ratings. The panel at the right shows how the average rating varies across age, gender and whether the rater is a US viewer or not.


User ratings for
National Treasure: Book of Secrets

Your Vote [Vote]

41,771 IMDb users have given a weighted average vote of 6.6 /

Demographic breakdowns are shown below.

Votes	Percentage	Rating
4,795	11.5%	10
3,286	7.9%	9
7,179	17.2%	8
10,636	25.5%	7
7,729	18.5%	6
3,646	8.7%	5
1,788	4.3%	4
940	2.3%	3
538	1.3%	2
1,234	3.0%	1

Arithmetic mean = 6.9. Median = 7

	Votes	Average
Males	33,644	6.5
Females	5,464	7.2
Aged under 18	2,492	7.6
Males under 18	1,795	7.5
Females under 18	695	8.1
Aged 18-29	26,045	6.7
Males Aged 18-29	22,603	6.6
Females Aged 18-29	3,372	7.3
Aged 30-44	8,210	6.3
Males Aged 30-44	7,216	6.3
Females Aged 30-44	936	6.7
Aged 45+	2,258	6.6
Males Aged 45+	1,814	6.5
Females Aged 45+	420	7.0
IMDb staff	8	6.4
Top 1000 voters	309	6.0
US users	14,792	6.8
Non-US users	24,283	6.5
IMDb users	41,771	6.6

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

The rating mechanism we have constructed is

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

AU Note: Screenshots are permissions? Can this data be reconfigured into a new table or figure, rather than attempting to obtain permission?

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

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

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

$$
\begin{aligned}
E_{im} &= 1 \text{ if } R_{im} = 5\\
E_{im} &= 0 \text{ if } R_{im} < 5.
\end{aligned}
$$

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 choice of the normal distribution for the random term makes it an *ordered probit model*.) Ordered choice models are appropriate for a wide variety of settings in the social and biological sciences. The essential

ingredient is the *mapping from an underlying, naturally ordered preference scale to a discrete ordered observed outcome*, such as the rating scheme described above. The model of ordered choice pioneered by Aitcheson and Silvey (1957), Snell (1964) and Walker and Duncan (1967) and articulated in its modern form by Zavoina and McElvey (1975) has become a widely used tool in many fields. The number of applications in the current literature is large and increasing rapidly, including:

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

18.3.1 THE ORDERED PROBIT MODEL

The ordered probit model is built around a latent regression in the same manner as the binomial probit model. We begin with

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

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

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

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

As before, we assume that ε is normally distributed across observations.[9] For the same reasons as in the binomial probit model (which is the special case of $J = 1$), we normalize the mean and variance of ε to zero and one. We then have the following probabilities:

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

For all the probabilities to be positive, we must have

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

Figure 23.4 shows the implications of the structure. This is an extension of the univariate probit model we examined earlier. The log-likelihood function and its derivatives can be obtained readily, and optimization can be done by the usual means.

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

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

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

18-35

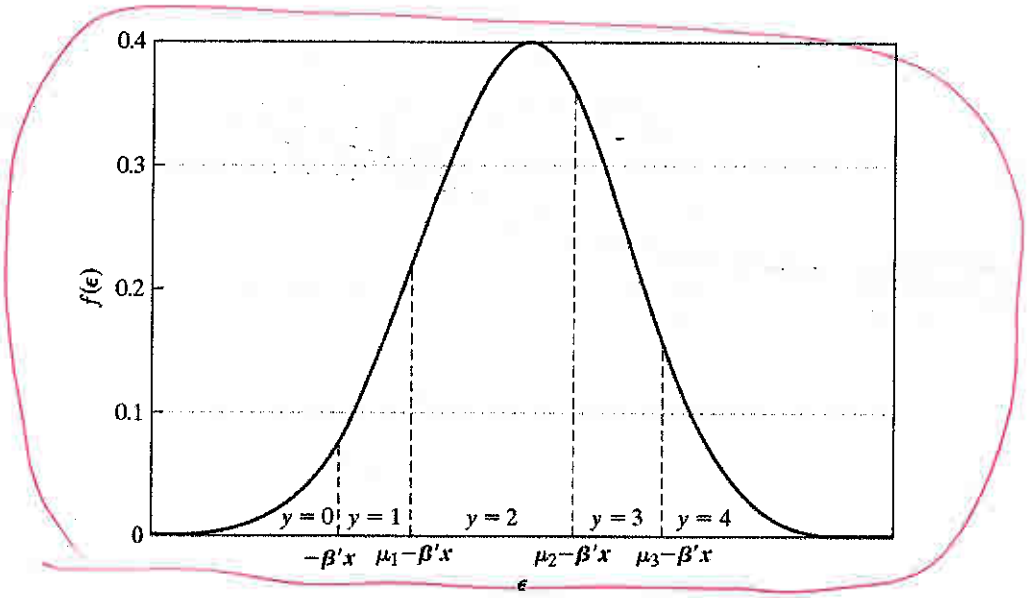


FIGURE 23.4 Probabilities in the Ordered Probit Model.

18.3

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

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

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

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

18.4

FIG 18.4

Figure 23.5 illustrates the effect. The probability distributions of y and y^* are shown in the solid curve. Increasing one of the x's while holding β 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 β 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 β. 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 β. Assuming that the particular β 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.[56]

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

18-36

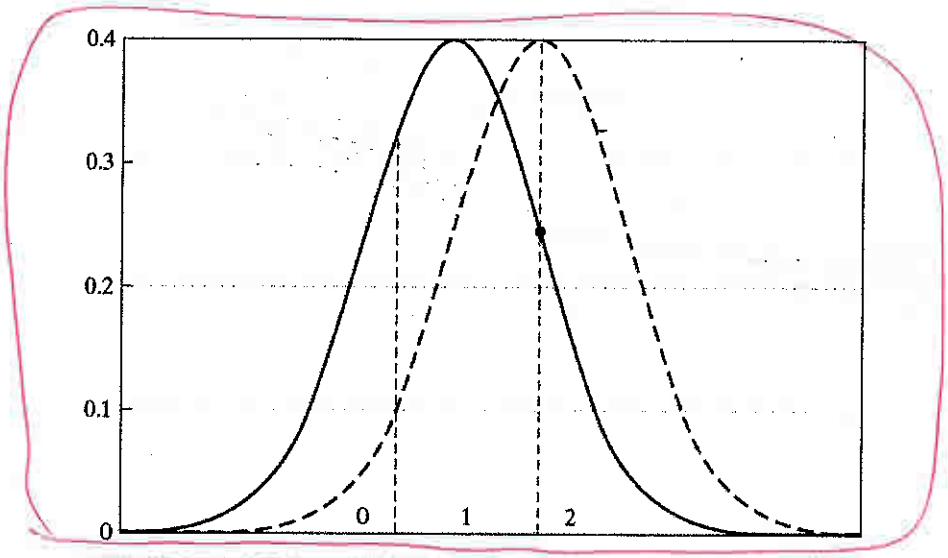


FIGURE ~~23.8~~ 18.5 Effects of Change in x on Predicted Probabilities.

18.3 18.4

Example ~~23.17~~ Rating Assignments

Marcus and Greene (1985) estimated an ordered probit model for the job assignments of new Navy recruits. The Navy attempts to direct recruits into job classifications in which they will be most productive. The broad classifications the authors analyzed were technical jobs with three clearly ranked skill ratings: "medium skilled," "highly skilled," and "nuclear qualified/highly skilled." Because the assignment is partly based on the Navy's own assessment and needs and partly on factors specific to the individual, an ordered probit model was used with the following determinants: (1) ENSPE = a dummy variable indicating that the individual entered the Navy with an "A school" (technical training) guarantee; (2) EDMA = educational level of the entrant's mother; (3) AFQT = score on the Armed Forces Qualifying Test; (4) EDYRS = years of education completed by the trainee; (5) MARR = a dummy variable indicating that the individual was married at the time of enlistment; and (6) AGEAT = trainee's age at the time of enlistment. (The data used in this study are not available for distribution.) The sample size was 5,641. The results are reported in Table ~~23.19~~. 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.

TB 18.10

18.10

18.10

TABLE ~~23.19~~ 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	—

18.11

TABLE 23.20 Marginal Effect of a Binary Variable

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

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

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

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

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

Note that the marginal effects sum to zero, which follows from the requirement that the probabilities add to one. This approach is not appropriate for evaluating the effect of a dummy variable. We can analyze a dummy variable by comparing the probabilities that result when the variable takes its two different values with those that occur with the other variables held at their sample means. For example, for the MARR variable, we have the results given in Table 23.20.

18.11.

18.3.4 **23.10.2. BIVARIATE ORDERED PROBIT MODELS** 17.5

There are several extensions of the ordered probit model that follow the logic of the bivariate probit model we examined in Section 23.8. A direct analog to the base case two-equation model is used in the study in Example 23.18. 18.3

18.34

Example 23.18 Calculus and Intermediate Economics Courses

Butler et al. (1994) analyzed the relationship between the level of calculus attained and grades in intermediate economics courses for a sample of Vanderbilt students. The two-step estimation approach involved the following strategy. (We are stylizing the precise formulation a bit to compress the description.) Step 1 involved a direct application of the ordered probit model of Section 23.10.1 to the level of calculus achievement, which is coded 0, 1, . . . , 6:

18.3.1

$$\begin{aligned} m_i^* &= \mathbf{x}_i'\beta + \varepsilon_i,\ \varepsilon_i \mid \mathbf{x}_i \sim \text{N}[0, 1], \\ m_i &= 0 \text{ if } -\infty < m_i^* \le 0 \\ &= 1 \text{ if } 0 < m_i^* \le \mu_1 \\ &\cdots \\ &= 6 \text{ if } \mu_6 < m_i^* < +\infty. \end{aligned}$$

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.

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,

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

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

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

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

Rejection of the null hypothesis calls the model specification into question. An alternative model in which there is a different $\boldsymbol{\beta}$ for each value of y has two problems; it does not force the probabilities to be positive and it is internally inconsistent. On the latter point, consider the suggested latent regression, $y^* = \mathbf{x}'\boldsymbol{\beta}_j + \varepsilon$. If the "$\boldsymbol{\beta}$" is different for each j, then it is not possible to construct a data generating mechanism for y^* (or, for example, simulate it); the realized value of y^* cannot be defined without knowing y (i.e., the realized j), since the applicable $\boldsymbol{\beta}$ depends on j, but y is supposed to be determined from y^* through, e.g., (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; (3) misspecification of the distributional form for the latent variable, i.e., "nonlogistic link function."

EXAMPLE 18.8 Brant Test for an Ordered Logit Model of Health Satisfaction

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

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

Variables KIDS, MARRIED and WORKING are binary indicators of whether there are children in the household, marital status and whether the individual was working at the time of the survey. (These data are examined further in Example 18.6 following.) 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.

18.11

TABLE 23.20 Marginal Effect of a Binary Variable

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

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

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

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

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

Note that the marginal effects sum to zero, which follows from the requirement that the probabilities add to one. This approach is not appropriate for evaluating the effect of a dummy variable. We can analyze a dummy variable by comparing the probabilities that result when the variable takes its two different values with those that occur with the other variables held at their sample means. For example, for the MARR variable, we have the results given in Table 23.20.

18.11

18.3.4

23.10.2 BIVARIATE ORDERED PROBIT MODELS

17.5

There are several extensions of the ordered probit model that follow the logic of the bivariate probit model we examined in Section 23.8. A direct analog to the base case two-equation model is used in the study in Example 23.18.

18.3

18.3.5

Example 23.18 Calculus and Intermediate Economics Courses

Butler et al. (1994) analyzed the relationship between the level of calculus attained and grades in intermediate economics courses for a sample of Vanderbilt students. The two-step estimation approach involved the following strategy. (We are stylizing the precise formulation a bit to compress the description.) Step 1 involved a direct application of the ordered probit model of Section 23.10.1 to the level of calculus achievement, which is coded 0, 1, . . . , 6:

18.3.1

$$m_i^* = \mathbf{x}_i'\beta + \varepsilon_i,\ \varepsilon_i \mid \mathbf{x}_i \sim \text{N}[0, 1],$$

$$m_i = 0 \text{ if } -\infty < m_i^* \le 0$$

$$= 1 \text{ if } 0 < m_i^* \le \mu_1$$

$$\cdots$$

$$= 6 \text{ if } \mu_5 < m_i^* < +\infty.$$

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

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

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

Independent variables in this regression include, among others, (1) dummy variables for which outcome in the ordered probit model applies to the student (with the zero reference case omitted), (2) grade in the last calculus course, (3) several other variables related to prior courses, (4) class size, (5) freshman GPA, etc. The unobservables in the *Grade* equation and the math attainment are clearly correlated, a feature captured by the additional assumption that $(\varepsilon_i, u_i \mid \mathbf{x}_i, \mathbf{z}_i) \sim \mathrm{N}_2[(0, 0), (1, \sigma_u^2), \rho\sigma_u]$. A nonzero ρ captures this "selection" effect. With this in place, the dummy variables in (1) have now become endogenous. The solution is a "selection" correction that we will examine in detail in Chapter 24. The modified equation becomes

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

They thus adopt a "control function" approach to accommodate the endogeneity of the math attainment dummy variables. [See Section 23.7 and (23-43) for another application of this method.] The term $\lambda(\mathbf{x}_i'\beta, \mu_1, \ldots, \mu_5)$ is a generalized residual that is constructed using the estimates from the first-stage ordered probit model. [A precise statement of the form of this variable is given in Li and Tobias (2006).] Linear regression of the course grade on $\mathbf{z}_i$ and this constructed regressor is computed at the second step. The standard errors at the second step must be corrected for the use of the estimated regressor using what amounts to a Murphy and Topel (2002) correction. (See Section ~~16.4.~~)

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 *Grade* equation above could also be treated as an ordered probit model. The resulting **bivariate ordered probit** model would be

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

where

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

Li and Tobias extended their analysis to this case simply by "transforming" the dependent variable in Butler et al.'s second equation. Computing the log-likelihood using sets of bivariate normal probabilities is fairly straightforward for the bivariate ordered probit model. [See Greene (2007).] However, the classical study of these data using the bivariate ordered approach remains to be done, so a side-by-side comparison to Li and Tobias's Bayesian alternative estimator is not possible. The endogeneity of the calculus dummy variables in (1) remains a feature of the model, so both the MLE and the Bayesian posterior are less straightforward than they might appears. Whether the results in Section 23.8.4 on the recursive bivariate probit model extend to this case also remains to be determined.

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