# 7 Variations on a Theme

## 7.1 Introduction

Simulation gives the researcher the freedom to specify models that appropriately represent the choice situations under consideration, without being unduly hampered by purely mathematical concerns. This perspective has been the overarching theme of our book. The discrete choice models that we have discussed – namely, logit, nested logit, probit, and mixed logit – are used in the vast majority of applied work. However, readers should not feel themselves constrained to use these models. In the current chapter, we describe several models that are derived under somewhat different behavioral concepts. These models are variations on the ones already discussed, directed toward specific issues and data. The point is not simply to describe additional models. Rather, the discussion illustrates how the researcher might examine a choice situation and develop a model and estimation procedure that seem appropriate for that particular situation, drawing from, and yet adapting, the standard set of models and tools.

Each section of this chapter is motivated by a type of data, representing the outcome of a particular choice process. The arena in which such data might arise is described, and the limitations of the primary models for these data are identified. In each case, a new model is described that better represents the choice situation. Often this new model is only a slight change from one of the primary models. However, the slight change will often make the standard software unusable, so that the researcher will need to develop her own software, perhaps by modifying the codes that are available for standard models. The ability to revise code to represent new specifications enables the researcher to utilize the freedom that the field offers.

## 7.2 Stated-Preference and Revealed-Preference Data

*Revealed-preference* data relate to people's actual choices in real-world situations. These data are so called because people reveal their tastes, or preferences, through the choices they make in the world. *Stated-preference* data are data collected in experimental or survey situations where respondents are presented with hypothetical choice situations. The term refers to the fact that the respondents state what their choices would be in the hypothetical situations. For example, in a survey, a person might be presented with three cars with different prices and other attributes. The person is asked which of the three cars he would buy if offered only these three cars in the real world. The answer the person gives is the person's stated choice. A revealed-preference datum for the respondent is obtained by asking which car he bought when he last bought a car.

There are advantages and limitations to each type of data. Revealed-preference data have the advantage that they reflect actual choices. This, of course, is a very big advantage. However, such data are limited to the choice situations and attributes of alternatives that currently exist or have existed historically. Often a researcher will want to examine people's responses in situations that do not currently exist, such as the demand for a new product. Revealed-preference data are simply not available for these new situations. Even for choice situations that currently exist, there may be insufficient variation in relevant factors to allow estimation with revealed-preference data. For example, suppose the researcher wants to examine the factors that affect California households' choice of energy supplier. While residential customers have been able to choose among suppliers for many years, there has been practically no difference in price among the suppliers' offers. Customers' response to price cannot be estimated on data that contain little or no price variation. An interesting paradox arises in this regard. If customers were highly price-responsive, then suppliers, knowing this, would offer prices that met their competitors' prices; the well-known equilibrium in this situation is that all firms offer (essentially) the same price. If the data from this market were used in a choice model, the price coefficient would be found to be insignificant, since there is little price variation in the data. The researcher could erroneously conclude from this insignificance that price is unimportant to consumers. This paradox is inherent in revealed-preference data. Factors that are the most important to consumers will often exhibit the least variation due to the natural forces of market equilibrium. Their importance might therefore be difficult to detect with revealed-preference data.

Stated-preference data complement revealed-preference data. A questionnaire is designed in which the respondent is presented with one or more choice experiments. In each experiment, two or more options are described, and the respondent is asked which option he would choose if facing the choice in the real world. For example, in the data that we examine in Chapter 11, each surveyed respondent is presented with 12 experiments. In each experiment, four hypothetical energy suppliers were described, with the price, contract terms, and other attributes given for each supplier. The respondent is asked to state which of the four suppliers he would choose.

The advantage of stated-preference data is that the experiments can be designed to contain as much variation in each attribute as the researcher thinks is appropriate. While there may be little price variation over suppliers in the real world, the suppliers that are described in the experiments can be given sufficiently different prices to allow precise estimation. Attributes can be varied over respondents and over experiments for each respondent. This degree of variation contrasts with market data, where often the same products are available to all customers, such that there is no variation over customers in the attributes of products. Importantly, for products that have never been offered before, or for new attributes of old products, stated-preference data allow estimation of choice models when revealed-preference data do not exist. Louviere *et al.* (2000) describe the appropriate collection and analysis of stated-preference data.

The limitations of stated-preference data are obvious: what people say they will do is often not the same as what they actually do. People may not know what they would do if a hypothetical situation were real. Or they may not be willing to say what they would do. In fact, respondents' idea of what they would do might be influenced by factors that wouldn't arise in the real choice situations, such as their perception of what the interviewer expects or wants as answers.

By combining stated- and revealed-preference data, the advantages of each can be obtained while mitigating the limitations. The stated-preference data provide the needed variation in attributes, while the revealed-preference data ground the predicted shares in reality. To utilize these relative strengths, an estimation procedure is needed that (1) allows the ratios of coefficients (which represent the relative importance of the various attributes) to be estimated primarily from the stated-preference data (or more generally, from whatever variation in the attributes exists, which is usually from the stated-preference data), while (2) allowing the alternative-specific constants and overall scale of the parameters to be determined by the revealed preference data (since the constants and scale determine average shares in base conditions).

Procedures for estimating discrete choice models on a combination of stated- and revealed-preference data are described by Ben-Akiva and Morikawa (1990), Hensher and Bradley (1993), and Hensher *et al.* (1999) in the context of logit models, and by Bhat and Castelar (2002) and Brownstone *et al.* (2000) with mixed logit. These procedures constitute variations on the methods we have already examined. The most prevalent issue when combining stated- and revealed-preference data is that the unobserved factors are generally different for the two types of data. We describe in the following paragraphs how this issue can readily be addressed.

Let the utility that person $n$ obtains from alternative $j$ in situation $t$ be specified as $U_{njt} = \beta' x_{njt} + e_{njt}$, where $x_{njt}$ does not include alternative-specific constants and $e_{njt}$ represents the effect of factors that are not observed by the researcher. These factors have a mean for each alternative (representing the average effect of all excluded factors on the utility of that alternative) and a distribution around this mean. The mean is captured by an alternative-specific constant, labeled $c_j$, and for a standard logit model the distribution around this mean is extreme value with variance $\lambda^2\pi^2/6$. As described in Chapters 2 and 3, the scale of utility is set by normalizing the variance of the unobserved portion of utility. The utility function becomes $U_{njt} = (\beta/\lambda)' x_{njt} + c_j/\lambda + \varepsilon_{njt}$, where the normalized error $\varepsilon_{njt} = (e_{njt} - c_j)/\lambda$ is now iid extreme value with variance $\pi^2/6$. The choice probability is given by the logit formula based on $(\beta/\lambda)' x_{njt} + c_j/\lambda$. The parameters that are estimated are the original parameters divided by the scale factor $\lambda$.

This specification is reasonable for many kinds of data and choice situations. However, there is no reason to expect the alternative-specific constants and the scale factor to be the same for stated-preference data as for revealed-preference data. These parameters reflect the effects of unobserved factors, which are necessarily different in real choice situations than hypothetical survey situations. In real choices, a multitude of issues that affect a person but are not observed by the researcher come into play. In a stated-preference experiment, the respondent is (usually) asked to assume all alternatives to be the same on factors that are not explicitly mentioned in the experiment. If the respondent followed this instruction exactly, there would, by definition, be no unobserved factors in the stated-preference choices. Of course, respondents inevitably bring some outside concepts into the experiments, such that unobserved factors do enter. However, there is no reason to expect that these factors are the same, in mean or variance, as in real-world choices.

To account for these differences, separate constants and scale parameters are specified for stated-preference choice situations and for revealed-preference situations. Let $c_j^s$ and $c_j^r$ represent the mean effect of unobserved factors for alternative $j$ in stated-preference experiments and revealed-preference choices, respectively. Similarly, let $\lambda^s$ and $\lambda^r$ represent the scales (proportional to the standard deviations) of the distributions of unobserved factors around these means in stated- and revealed-preference situations, respectively. To set the overall scale of utility, we normalize either of the scale parameters to 1, which makes the other scale parameter equal the ratio of the two original scale parameters. Let's normalize $\lambda^r$, so that $\lambda^s$ reflects the variance of unobserved factors in stated-preference situations relative to that in revealed-preference situations. Utility then becomes

$$U_{njt} = (\beta/\lambda^s)' x_{njt} + c_j^s/\lambda^s + \varepsilon_{njt}$$

for each $t$ that is a stated-preference situation, and

$$U_{njt} = \beta' x_{njt} + c_j^r + \varepsilon_{njt}$$

for each $t$ that is a revealed-preference situation.

The model is estimated on the data from both the revealed- and stated-preference choices. Both groups of observations are "stacked" together as input to a logit estimation routine. A separate set of alternative-specific constants is estimated for the stated-preference and revealed-preference data. Importantly, the coefficients in the model are divided by a parameter $1/\lambda^s$ for the stated-preference observations. This separate scaling is not feasible in most standard logit estimation packages. However, the researcher can easily modify available codes (or her own code) to allow for this extra parameter. Hensher and Bradley (1993) show how to estimate this model on software for nested logits.

Note that, with this setup, the elements of $\beta$ are estimated on both types of data. The estimates will necessarily reflect the amount of variation that each type of data contains for the attributes (that is, the elements of $x$ ). If there is little variance in the revealed-preference data, reflecting conditions in real-world markets, then the $\beta$'s will be determined primarily by the stated-preference data, which contain whatever variation was built into the experiments. Insofar as the revealed-preference data contain usable variation, this information will be incorporated into the estimates.

The alternative-specific constants are estimated separately for the two types of data. This distinction allows the researcher to avoid many of the biases that stated-preference data might exhibit. For example,

respondents often say that they will buy a product far more than they actually end up doing. The average probability of buying the product is captured in the alternative-specific constant for the product. If this bias is occurring, then the estimated constant for the stated-preference data will be greater than that for the revealed-preference data. When forecasting, the researcher can use the constant from the revealed-preference data, thereby grounding the forecast in a market-based reality. Similarly, the scale for the revealed-preference data (which is normalized to 1) can be used in forecasting instead of the scale from the stated-preference data, thereby incorporating correctly the real-world variance in unobserved factors.

## 7.3 Ranked Data

In stated-preference experiments, respondents may be asked to rank the alternatives instead of just identifying the one alternative that they would choose. This ranking can be requested in a variety of ways. The respondents can be asked to state which alternative they would choose, and then, after they have made this choice, can be asked which of the remaining alternatives they would choose, continuing through all the alternatives. Instead, respondents can simply be asked to rank the alternatives from best to worst. In any case, the data that the researcher obtains constitute a ranking of the alternatives that presumably reflects the utility that the respondent obtains from each alternative.

Ranked data can be handled in a standard logit or mixed logit model using currently available software without modification. All that is required is that the input data be constructed in a particular way, which we describe in the following text. For a probit model, the available software would need to be modified slightly to handle ranked data. However, the modification is straightforward. We consider standard and mixed logit first.

### *7.3.1. Standard and Mixed Logit*

Under the assumptions for standard logit, the probability of any ranking of the alternatives from best to worst can be expressed as the product of logit formulas. Consider, for example, a respondent who was presented with four alternatives labeled $A$, $B$, $C$, and $D$. Suppose the person ranked the alternatives as follows: $C$, $B$, $D$, $A$, where $C$ is the first choice. If the utility of each alternative is distributed iid extreme value (as for a logit model), then the probability of this ranking can be expressed as the logit probability of choosing alternative $C$ from the set $A, B, C, D$,

*times* the logit probability of choosing alternative $B$ from the remaining alternatives $A$, $B$, $D$, *times* the probability of choosing alternative $D$ from the remaining alternatives $A$ and $D$.

Stated more explicity, let $U_{nj} = \beta' x_{nj} + \varepsilon_{nj}$ for $j = A, \ldots, D$ with $\varepsilon_{nj}$ iid extreme value. Then

$$\begin{aligned}&\text{Prob(ranking } C, B, D, A\text{)}\\ &= \frac{e^{\beta' x_{nC}}}{\sum_{j=A,B,C,D} e^{\beta' x_{nj}}} \frac{e^{\beta' x_{nB}}}{\sum_{j=A,B,D} e^{\beta' x_{nj}}} \frac{e^{\beta' x_{nD}}}{\sum_{j=A,D} e^{\beta' x_{nj}}}.\end{aligned} \tag{7.1}$$

This simple expression for the ranking probability is an outcome of the particular form of the extreme value distribution, first shown by Luce and Suppes (1965). It does not apply in general; for example, it does not apply with probit models.

Equation (7.1) implies that the ranking of the four alternatives can be represented as being the same as three independent choices by the respondent. These three choices are called *pseudo-observations*, because each respondent's complete ranking, which constitutes an observation, is written as if it were multiple observations. In general, a ranking of $J$ alternatives provides $J - 1$ pseudo-observations in a standard logit model. For the first pseudo-observation, all alternatives are considered available, and the dependent variable identifies the first-ranked alternative. For the second pseudo-observation, the first-ranked alternative is discarded. The remaining alternatives constitute the choice set, and the dependent variable identifies the second-ranked alternative, and so on. In creating the input file for logit estimation, the explanatory variables for each alternative are repeated $J - 1$ times, making that many pseudo-observations. The dependent variable for these pseudo-observations identifies, respectively, the first-ranked, second-ranked, and so on, alternatives. For each pseudo-observation, the alternatives that are ranked above the dependent variable for that pseudo-observation are omitted (i.e., censored out). Once the data are constructed in this way, the logit estimation proceeds as usual.

A logit model on ranked alternatives is often called an *exploded logit*, since each observation is exploded into several pseudo-observations for the purposes of estimation. Prominent applications include Beggs *et al.* (1981), Chapman and Staelin (1982), and Hausman and Ruud (1987).

A mixed logit model can be estimated on ranked data with the same explosion. Assume now that $\beta$ is random with density $g(\beta \mid \theta)$, where $\theta$ are parameters of this distribution. Conditional on $\beta$, the probability of the person's ranking is a product of logits, as given in equation (7.1). The unconditional probability is then the integral of this product over

the density of $\beta$:

$$\begin{aligned}&\text{Prob(ranking } C, B, A, D)\\ &\quad = \int \left( \frac{e^{\beta' x_{nC}}}{\sum_{j=A,B,C,D} e^{\beta' x_{nj}}} \frac{e^{\beta' x_{nB}}}{\sum_{j=A,B,D} e^{\beta' x_{nj}}} \frac{e^{\beta' x_{nD}}}{\sum_{j=A,D} e^{\beta' x_{nj}}} \right)\\ &\qquad \times g(\beta \mid \theta)\, d\beta.\end{aligned}$$

The mixed logit model on ranked alternatives is estimated with regular mixed logit routines for panel data, using the input data setup as described previously for logit, where the $J-1$ pseudo-observations for each ranking are treated as $J-1$ choices in a panel. The mixed logit incorporates the fact that each respondent has his own coefficients and, importantly, that the respondent's coefficients affect his entire ranking, so that the pseudo-observations are correlated. A logit model on ranked data does not allow for this correlation.

### *7.3.2. Probit*

Ranked data can also be utilized effectively in a probit model. Let the utility of the four alternatives be as just stated for a logit except that the error terms are jointly normal: $U_{nj} = \beta' x_{nj} + \varepsilon_{nj}$ for $j = A, B, C, D$, where $\varepsilon_n = \langle \varepsilon_{nA}, \ldots, \varepsilon_{nD} \rangle'$ is distributed $N(0, \Omega)$. As before, the probability of the person's ranking is Prob(ranking $C, B, D, A$) = Prob($U_{nC} > U_{nB} > U_{nD} > U_{nA}$). Decomposing this joint probability into conditionals and a marginal does not help with a probit in the way that it does with logit, since the conditional probabilities do not collapse to unconditional probabilities as they do under independent errors. Another tack is taken instead. Recall that for probit models, we found that it is very convenient to work in utility differences rather than the utilities themselves. Denote $\tilde{U}_{njk} = U_{nj} - U_{nk}$, $\tilde{x}_{njk} = x_{nj} - x_{nk}$, and $\tilde{\varepsilon}_{njk} = \varepsilon_{nj} - \varepsilon_{nk}$. The probability of the ranking can then be expressed as Prob(ranking $C, B, D, A$) = Prob($U_{nC} > U_{nB} > U_{nD} > U_{nA}$) = Prob($\tilde{U}_{nBC} < 0, \tilde{U}_{nDB} < 0, \tilde{U}_{nAD} < 0$).

To express this probability, we define a transformation matrix $M$ that takes appropriate differences. The reader might want to review Section 5.6.3 on simulation of probit probabilities for one chosen alternative, which uses a similar transformation matrix. The same procedure is used for ranked data, but with a different transformation matrix.

Stack the alternatives $A$ to $D$, so that utility is expressed in vector form as $U_n = V_n + \varepsilon_n$, where $\varepsilon_n \sim N(0, \Omega)$. Define the $3 \times 4$ matrix

$$M = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & -1 \end{pmatrix}.$$

This matrix has a row for each inequality in the argument of the probability $\text{Prob}(\tilde{U}_{nBC} < 0, \tilde{U}_{nDB} < 0, \tilde{U}_{nAD} < 0)$. Each row contains a 1 and a $-1$, along with zeros, where the 1 and $-1$ identify the alternatives that are being differenced for the inequality. With this matrix, the probability of the ranked alternatives becomes

$$\begin{aligned}\text{Prob}(\text{ranking } C, B, D, A) &= \text{Prob}(\tilde{U}_{nBC} < 0, \tilde{U}_{nDB} < 0, \tilde{U}_{nAD} < 0)\\ &= \text{Prob}(MU_n < 0)\\ &= \text{Prob}(MV_n + M\varepsilon_n < 0)\\ &= \text{Prob}(M\varepsilon_n < -MV_n).\end{aligned}$$

The error differences defined by $M\varepsilon_n$ are distributed jointly normal with zero mean and covariance $M\Omega M'$. The probability that these correlated error differences fall below $-MV_n$ is simulated by GHK in the manner given in Section 5.6.3. The procedure has been implemented by Hajivassiliou and Ruud (1994) and Schechter (2001).

## 7.4 Ordered Responses

In surveys, respondents are often asked to provide ratings of various kinds. Examples include:

How good a job do you think the president is doing? Check one:

1. very good job
2. good job
3. neither good nor bad
4. poor job
5. very poor job

How well do you like this book? Rate the book from 1 to 7, where 1 is the worst you have ever read (aside from *The Bridges of Madison County*, of course) and 7 is the best

1 2 3 4 5 6 7

How likely are you to buy a new computer this year?

1. Not likely at all
2. Somewhat likely
3. Very likely

The main characteristic of these questions, from a modeling perspective, is that the potential responses are ordered. A book rating of 6 is higher than 5, which is higher than 4; and a presidential rating of "very poor" is worse than "poor," which is worse than "neither good nor bad." A standard logit model could be specified with each potential response as an alternative. However, the logit model's assumption of independent errors for each alternative is inconsistent with the fact that the alternatives

are ordered: with ordered alternatives, one alternative is similar to those close to it and less similar to those further away. The ordered nature could be handled by specifying a nested logit, mixed logit, or probit model that accounts for the pattern of similarity and dissimilarity among the alternatives. For example, a probit model could be estimated with correlation among the alternatives, with the correlation between 2 and 3 being greater than that between 1 and 3, and the correlation between 1 and 2 also being greater than that between 1 and 3. However, such a specification, while it might provide fine results, does not actually fit the structure of the data. Recall that the traditional derivation for these models starts with a specification of the utility associated with each alternative. For the ratings question about the president's job, the derivation would assume that there are five utilities, one for each potential response, and that the person chooses the number 1 to 5 that has the greatest utility. While it is perhaps possible to think of the decision process in this way (and the resulting model will probably provide useful results), it is not a very natural way to think about the respondent's decision.

A more natural representation of the decision process is to think of the respondent as having some level of utility or opinion associated with the object of the question and answering the question on the basis of how great this utility is. For example, on the presidential question, the following derivation seems to better represent the decision process. Assume that the respondent has an opinion on how well the president is doing. This opinion is represented in a (unobservable) variable that we label $U$, where higher levels of $U$ mean that the person thinks the president is doing a better job and lower levels mean he thinks the president is doing a poorer job. In answering the question, the person is asked to express this opinion in one of five categories: "very good job," "good job," and so on. That is, even though the person's opinion, $U$, can take many different levels representing various levels of liking or disliking the job the president is doing, the question allows only five possible responses. The person chooses a response on the basis of the level of his $U$. If $U$ is above some cutoff, which we label $k_1$, the respondent chooses the answer "very good job." If $U$ is below $k_1$ but above another cutoff, $k_2$, then he answers "good job." And so on. The decision is represented as

- "very good job" if $U > k_1$
- "good job" if $k_1 > U > k_2$
- "neither good or bad" if $k_2 > U > k_3$
- "poor job" if $k_3 > U > k_4$
- "very poor job" if $k_4 > U$.

The researcher observes some factors that relate to the respondent's opinion, such as the person's political affiliation, income, and so on.

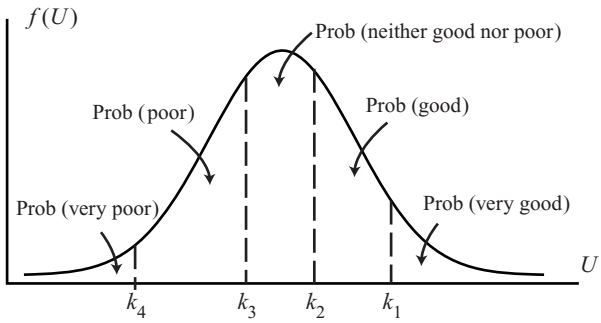


Figure 7.1. Distribution of opinion about president's job.

However, other factors that affect the person's opinion cannot be observed. Decompose $U$ into observed and unobserved components: $U = \beta' x + \varepsilon$. As usual, the unobserved factors $\varepsilon$ are considered random. Their distribution determines the probability for the five possible responses.

Figure 7.1 illustrates the situation. $U$ is distributed around $\beta' x$ with the shape of the distribution following the distribution of $\varepsilon$. There are cutoff points for the possible responses: $k_1, \ldots, k_4$. The probability that the person answers with "very poor job" is the probability that $U$ is less than $k_4$, which is the area in the left tail of the distribution. The probability that the person says "poor job" is the probability that $U$ is above $k_4$, indicating that he doesn't think that the job is very poor, but is below $k_3$. This probability is the area between $k_4$ and $k_3$.

Once a distribution for $\varepsilon$ is specified, the probabilities can be calculated exactly. For simplicity, assume that $\varepsilon$ is distributed logistic, which means that the cumulative distribution of $\varepsilon$ is $F(\varepsilon) = \exp(\varepsilon)/(1 + \exp(\varepsilon))$. The probability of the answer "very poor job" is then

$$\begin{aligned}
\text{Prob}(\text{“very poor job”}) &= \text{Prob}(U < k_4) \\
&= \text{Prob}(\beta' x + \varepsilon < k_4) \\
&= \text{Prob}(\varepsilon < k_4 - \beta' x) \\
&= \frac{e^{k_4 - \beta' x}}{1 + e^{k_4 - \beta' x}}.
\end{aligned}$$

The probability of "poor job" is

$$\begin{aligned}
\text{Prob}(\text{“poor job”}) &= \text{Prob}(k_4 < U < k_3) \\
&= \text{Prob}(k_4 < \beta' x + \varepsilon < k_3) \\
&= \text{Prob}(k_4 - \beta' x < \varepsilon < k_3 - \beta' x) \\
&= \text{Prob}(\varepsilon < k_3 - \beta' x) - \text{Prob}(\varepsilon < k_4 - \beta' x) \\
&= \frac{e^{k_3 - \beta' x}}{1 + e^{k_3 - \beta' x}} - \frac{e^{k_4 - \beta' x}}{1 + e^{k_4 - \beta' x}}.
\end{aligned}$$

Probabilities for the other answers are obtained analogously. The probabilities enter the log-likelihood function as usual, and maximization of the likelihood function provides estimates of the parameters. Note that the parameters consist of $\beta$, which gives the impact of the explanatory variables on people's opinion of the president, as well as the cutoff points $k_1, \ldots, k_4$.

The model is called *ordered logit*, since it uses the logistic distribution on ordered alternatives. Unfortunately, nested logit models have occasionally been called ordered logits; this nomenclature causes confusion and will hopefully be avoided in the future.

Note that the probabilities in the ordered logit model incorporate the binary logit formula. This similarity to binary logit is only incidental: the traditional derivation of a binary logit specifies two alternatives with utility for each, while the ordered logit model has one utility with multiple alternatives to represent the level of that utility. The similarity in formula arises from the fact that, if two random variables are iid extreme value, then their difference follows a logistic distribution. Therefore, assuming that both utilities in a binary logit are iid extreme value is equivalent to assuming that the difference in the utilities is distributed logistic, the same as the utility in the ordered logit model.

A similar model is obtained under the assumption that $\varepsilon$ is distributed standard normal instead of logistic (Zavoina and McKelvey, 1975). The only difference arises in that the binary logit formula is replaced with the cumulative standard normal distribution. That is,

$$\begin{aligned}\text{Prob(“very poor job”)} &= \text{Prob}(\varepsilon < k_4 - \beta' x) \\ &= \Phi(k_4 - \beta' x)\end{aligned}$$

and

$$\begin{aligned}\text{Prob(“poor job”)} &= \text{Prob}(\varepsilon < k_3 - \beta' x) - \text{Prob}(\varepsilon < k_4 - \beta' x) \\ &= \Phi(k_3 - \beta' x) - \Phi(k_4 - \beta' x),\end{aligned}$$

where $\Phi$ is the standard cumulative normal function. This model is called *ordered probit*. Software for ordered logit and probit is available in many commercial packages.

The researcher might believe that the parameters vary randomly in the population. In that case, a mixed version of the model can be specified, as in Bhat (1999). Let the density of $\beta$ be $g(\beta \mid \theta)$. Then the mixed ordered logit probabilities are simply the ordered logit probabilities integrated over the density $g(\cdot)$. For example,

$$\text{Prob(“very poor job”)} = \int \left( \frac{e^{k_4 - \beta' x}}{1 + e^{k_4 - \beta' x}} \right) g(\beta \mid \theta)\, d\beta$$

and

$$\text{Prob}(\text{“poor job”}) = \int \left( \frac{e^{k_3 - \beta' x}}{1 + e^{k_3 - \beta' x}} - \frac{e^{k_4 - \beta' x}}{1 + e^{k_4 - \beta' x}} \right) g(\beta \mid \theta)\, d\beta,$$

and so on. These probabilities are simulated in the same way as mixed logits, by drawing values of $\beta$ from $g(\cdot)$, calculating the ordered logit probability for each draw, and averaging the results. Mixed ordered probit is derived similarly.

### *7.4.1. Multiple Ordered Responses*

Respondents' answers to different questions are often related. For example, a person's rating of how well the president is doing is probably related to the person's rating of how well the economy is doing. The researcher might want to incorporate into the analysis the fact that the answers are related. To be concrete, suppose that respondents are asked to rate both the president and the economy on a five-point scale, like the rating given for the president. Let $U$ be the respondent's opinion of the job the president is doing, and let $W$ be the respondent's assessment of the economy. Each of these assessments can be decomposed into observed and unobserved factors: $U = \beta' x + \varepsilon$ and $W = \alpha' z + \mu$. Insofar as the assessments are related due to observed factors, the same variables can be included in $x$ and $z$. To allow for the possibility that the assessments are related due to unobserved factors, we specify $\varepsilon$ and $\mu$ to be jointly normal with correlation $\rho$ (and unit variances by normalization). Let the cutoffs for $U$ be denoted $k_1, \ldots, k_4$ as before, and the cutoffs for $W$ be denoted $c_1, \ldots, c_4$. We want to derive the probability of each possible combination of responses to the two questions.

The probability that the person says the president is doing a "very poor job" and also that the economy is doing "very poorly" is derived as follows:

$$\begin{aligned}
&\text{Prob}(\text{President “very poor” and economy “very poor”}) \\
&\quad = \text{Prob}(U < k_4 \text{ and } W < c_4) \\
&\quad = \text{Prob}(\varepsilon < k_4 - \beta' x \text{ and } \mu < c_4 - \alpha' z) \\
&\quad = \text{Prob}(\varepsilon < k_4 - \beta' x) \\
&\qquad \times \text{Prob}(\mu < c_4 - \alpha' z \mid \varepsilon < k_4 - \beta' x).
\end{aligned}$$

Similarly, the probability of a rating of "very poor" for the president and

"good" for the economy is

$$\begin{aligned}
&\text{Prob(President "very poor" and economy "good")}\\
&\quad = \text{Prob}(U < k_4 \text{ and } c_2 < W < c_1)\\
&\quad = \text{Prob}(\varepsilon < k_4 - \beta' x \text{ and } c_2 - \alpha' z < \mu < c_1 - \alpha' z)\\
&\quad = \text{Prob}((\varepsilon < k_4 - \beta' x)\\
&\qquad \times \text{Prob}(c_2 - \alpha' z < \mu < c_1 - \alpha' z \mid \varepsilon < k_4 - \beta' x).
\end{aligned}$$

The probabilities for other combinations are derived similarly, and generalization to more than two related questions is straightforward. The model is called multivariate (or multiresponse) ordered probit. The probabilities can be simulated by GHK in a manner similar to that described in Chapter 5. The explanation in Chapter 5 assumes that truncation of the joint normal is only on one side (since for a standard probit the probability that is being calculated is the probability that all utility differences are below zero, which is truncation from above), while the probabilities for multivariate ordered probit are truncated on two sides (as for the second probability listed earlier). However, the logic is the same, and interested readers can refer to Hajivassiliou and Ruud (1994) for an explicit treatment of GHK with two-sided truncation.

## 7.5 Contingent Valuation

In some surveys, respondents are asked to express their opinions or actions relative to a specific number that the interviewer states. For example, the interviewer might ask: "Consider a project that protected the fish in specific rivers in Montana. Would you be willing to spend $50 to know that the fish in these rivers are safe?" This question is sometimes followed by another question that depends on the respondent's answer to the first question. For example, if the person said "yes" to the above question, the interviewer might follow up by asking, "How about $75? Would you be willing to pay $75?" If the person answered "no" to the first question, indicating that he was not willing to pay $50, the interviewer would follow up with "Would you be willing to pay $25?"

These kinds of questions are used in environmental studies where the lack of markets for environmental quality prevent valuation of resources by revelation procedures; the papers edited by Hausman (1993) provide a review and critique of the procedure, which is often called "contingent valuation." When only one question is asked, such as whether the person is willing to pay $50, the method is called *single-bounded*, since the person's answer gives one bound on his true willingness to pay. If the person answers "yes," the researcher knows that his true willingness to

pay is at least \$50, but she does not know how *much* more. If the person answers "no," the researcher knows that the person's willingness to pay is less than \$50. Examples of studies using single-bounded methods are Cameron and James (1987) and Cameron (1988).

When a follow-up question is asked, the method is called *double-bounded.* If the person says that he is willing to pay \$50 but not \$75, the researcher knows his true willingness to pay is between \$50 and \$75, that is, is bounded on both sides. If the person says he is not willing to pay \$50 but is willing to pay \$25, his willingness to pay is known to be between \$25 and \$50. Of course, even with a double-bounded method, some respondents' willingness to pay is only singly bounded, such as that of a person who says he is willing to pay \$50 and also willing to pay \$75. Examples of this approach include Hanemann *et al.* (1991), Cameron and Quiggin (1994), and Cai *et al.* (1998).

The figure that is used as the prompt (i.e., the \$50 in our example) is varied over respondents. The answers from a sample of people are then used to estimate the distribution of willingness to pay. The estimation procedure is closely related to that just described for ordered logits and probits, except that the cutoff points are given by the questionnaire design rather than estimated as parameters. We describe the procedure as follows.

Let $W_n$ represent the true willingness to pay of person $n$. $W_n$ varies over people with distribution $f(W \mid \theta)$, where $\theta$ are the parameters of the distribution, such as the mean and variance. The researcher's goal is to estimate these population parameters. Suppose the researcher designs a questionnaire with a single-bounded approach, giving a different prompt (or reference value) for different respondents. Denote the prompt that is given to person $n$ as $k_n$. The person answers the question with a "yes" if $W_n > k_n$ and "no" otherwise. The researcher assumes that $W_n$ is distributed normally in the population with mean $\bar{W}$ and variance $\sigma^2$.

The probability of "yes" is $\text{Prob}(W_n > k_n) = 1 - \text{Prob}(W_n < k_n) = 1 - \Phi((k_n - \bar{W})/\sigma)$, and the probability of "no" is $\Phi((k_n - \bar{W})/\sigma)$, where $\Phi(\cdot)$ is the standard cumulative normal function. The log-likelihood function is then $\sum_n y_n \ln(1 - \Phi((k_n - \bar{W})/\sigma)) + (1 - y_n) \ln(\Phi((k_n - \bar{W})/\sigma))$, where $y_n = 1$ if person $n$ said "yes" and 0 otherwise. Maximizing this function provides estimates of $\bar{W}$ and $\sigma$.

A similar procedure is used if the researcher designs a double-bounded questionnaire. Let the prompt for the second question be $k_{nu}$ if the person answered "yes" to the first question, where $k_{nu} > k_n$, and let $k_{nl}$ be the second prompt if the person initially answered "no," where $k_{nl} < k_n$. There are four possible sequences of answers to the two questions. The

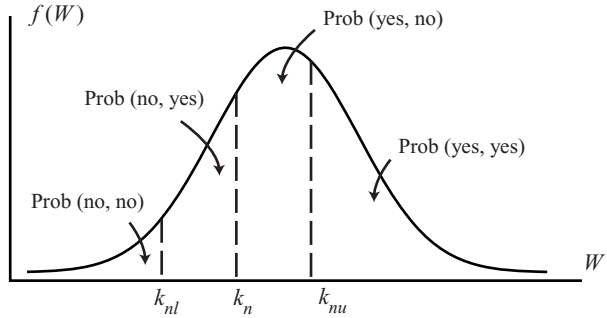


Figure 7.2. Distribution of willingness to pay.

probabilities for these sequences are illustrated in Figure 7.2 and given below:

- "no" then "no": $P = \text{Prob}(W_n < k_{nl}) = \Phi((k_{nl} - \bar{W})/\sigma)$
- "no" then "yes": $P = \text{Prob}(k_{nl} < W_n < k_n) = \Phi((k_n - \bar{W})/\sigma) - \Phi((k_{nl} - \bar{W})/\sigma)$
- "yes" then "no": $P = \text{Prob}(k_n < W_n < k_{nu}) = \Phi((k_{nu} - \bar{W})/\sigma) - \Phi((k_n - \bar{W})/\sigma)$
- "yes" then "yes": $P = \text{Prob}(W_n > k_{nu}) = 1 - \Phi((k_{nu} - \bar{W})/\sigma)$.

These probabilities enter the log-likelihood function, which is maximized to obtain estimates of $\bar{W}$ and $\sigma$. Other distributions can of course be used instead of normal. Lognormal is attractive if the researcher assumes that all people have a positive willingness to pay. Or the researcher might specify a distribution that has a mass at zero to represent the share of people who are not willing to pay anything, and a lognormal for the remaining share. Generalization to multiple dimensions is straightforward, to reflect, for example, that people's willingness to pay for one environmental package might also be related to their willingness to pay for another. As with multiresponse ordered probit, the GHK simulator comes in handy when the multiple values are assumed to be distributed jointly normal.

## 7.6 Mixed Models

We have discussed mixed logit and mixed ordered logit. Of course, mixed models of all kinds can be developed using the same logic. Any model whose probabilities can be written as a function of parameters can also be mixed by allowing the parameters to be random and integrating the function over the distribution of parameters (Greene, 2001). The

probability is simulated by drawing from the distribution, calculating the function for each draw, and averaging the results. We give two examples in the following section, but researchers will inevitably develop others that meet the needs of their particular projects, such as Bhat's (1999) use of mixed ordered logit.

### *7.6.1. Mixed Nested Logit*

The mixed logit model does not exhibit the independence from irrelevant alteratives property as logit does, and can approximate any substitution pattern by appropriate specification of variables and mixing distribution. This fact has led some people to feel that there is no further need for nested logit models. A mixed logit can be estimated that provides correlation–substitution patterns analogous to those of a nested logit. For example, consider a nested logit with two nests of alternatives labeled $A$ and $B$. Provided the log-sum coefficients are between 0 and 1, substitution within each nest is greater than substitution across nests. This substitution pattern can be represented in a mixed logit model by specifying a dummy variable for each nest and allowing the coefficients on the dummies to be random (constraining, for identification purposes, the means to be zero if a full set of alternative-specific constants are included, and the two variances to be the same).

While a mixed logit can be specified in this way, doing so misses the point of simulation. As discussed in Chapter 1, simulation is used as a way to approximate integrals when a closed form does not exist. Analytic integration is always more accurate than simulation and should be used whenever feasible, unless there is a compelling reason to the contrary. Using a mixed logit to represent the substitution patterns of a nested logit, while feasible, replaces the closed-form integral of the nested logit with an integral that needs to be simulated. From a numerical perspective, this replacement can only reduce accuracy. The only possible advantages of mixed logit in this context are that (1) it might be easier for the researcher to test numerous nesting structures, including overlapping nests, within a mixed logit than a nested logit, and (2) the researcher might have specified other coefficients to be random, so that a mixed logit is already being used.

The second reason suggests a mixed nested logit. Suppose the researcher believes that some of the coefficients in the model are random and also that, conditional on these coefficients, the unobserved factors are correlated over alternatives in a way that can be represented by a nested logit. A mixed nested logit model can be specified to represent this situation. Conditional on the coefficients that enter utility, the choice

probabilities are nested logit, which is a closed form and can be calculated exactly. The unconditional probability is the nested logit formula integrated over the distribution of the random coefficients. Software for mixed logit can be modified by simply locating the logit formula within the code and changing it to the appropriate nested logit formula. Experience indicates that maximizing the likelihood function for unmixed nested logits is often difficult numerically, and mixing the model will compound this difficulty. Hierarchical Bayes estimation (Chapter 12) could prove particularly useful in this situation, since it does not involve maximizing the likelihood function.

### *7.6.2. Mixed Probit*

A constraint of probit models, and in fact their defining characteristic, is that all random terms enter utility linearly and are randomly distributed in such a way that utility itself is normally distributed. This constraint can be removed by specifying a *mixed probit*. Suppose that some random terms enter nonlinearly or are not randomly distributed, but that *conditional* on these, utility is normally distributed. For example, a price coefficient might be lognormal to assure that it is negative for all people, and yet all other coefficients be either fixed or normal, and the final error terms jointly normal. A mixed probit model is appropriate for this specification. Conditional on the price coefficient, the choice probabilities follow the standard probit formula. The unconditional probabilities are the integral of this probit formula over the distribution of the price coefficient. Two layers of simulation are used to approximate the probabilities: (1) a draw of the price coefficient is taken, and (2) for this draw, the GHK or other probit simulator is used to approximate the conditional choice probability. This process is repeated many times, and the results are averaged.

Long run times can be expected for the mixed probit model, since the GHK simulator is calculated for each draw of the price coefficient. However, the number of draws in the GHK simulator can be reduced, since the averaging over draws of the price coefficient reduces the variance generated by the GHK simulator. In principle, the GHK simulator can be based on only one draw for each draw of the price coefficient. In practice, it may be advisable to use more than one draw, but far fewer than would be used in an unmixed probit.

The mixed probit model provides a way for the researcher to avoid some of the practical difficulties that can arise with a mixed logit model. For example, to represent pure heteroskedasticity (i.e., a different variance for each alternative's utility) or a fixed correlation pattern among

alternatives (i.e., a covariance matrix that does not depend on the variables), it can often be easier to estimate a probit instead of specifying numerous error components within a mixed logit. As emphasized by Ben-Akiva *et al.* (2001), specification of covariance and heteroskedasticity can be more complex in a mixed logit model than in a probit, because iid extreme value terms are necessarily added to whatever other random elements the researcher specifies. Probit is a more natural specification in these situations. However, if the researcher wants to include some nonnormal random terms, an unmixed probit cannot be used. Mixing the probit allows the researcher to include nonnormal terms while still maintaining the simplicity of probit's representation of fixed covariance for additive errors. Conceptually, the specification and estimation procedure are straightforward. The cost comes only in extra computation time, which becomes less relevant as computers get faster.

## 7.7 Dynamic Optimization

In previous chapters we examined certain types of dynamics, by which choices in one period affect choices in another period. For example, we described how a lagged dependent variable can be included to capture inertia or variety-seeking behavior. These discussions suggest a much wider realm of dynamics than we had actually considered. In particular: if past choices affect current choices, then current choices affect future choices, and a decision maker who is aware of this fact will take these future effects into consideration. A link from the past to the present necessarily implies a link from the present to the future.

In many situations, the choices that a person makes at one point in his life have a profound influence on the options that are available to him in the future. Going to college, while expensive and sometimes irritating, enhances future job possibilities. Saving money now allows a person to buy things later that he otherwise would not be able to afford. Going to the gym today means that we can skip going tomorrow. Most of us take future effects like these into consideration when choosing among current alternatives.

The question is: how can behavior such as this be represented in discrete choice models? In general the situation can be described as follows. A person makes a series of choices over time. The alternative that is chosen in one period affects the attributes and availability of alternatives in the future. Sometimes the future effects are not fully known, or depend on factors that have not yet transpired (such as the future state of the economy). However, the person knows that he will, in the future, maximize utility among the alternatives that are available at that time under

the conditions that prevail at that time. This knowledge enables him to choose the alternative in the current period that maximizes his expected utility over the current and future periods. The researcher recognizes that the decision maker acts in this way, but does not observe everything that the decision maker considers in the current and future periods. As usual, the choice probability is an integral of the decision maker's behavior over all possible values of the factors that the researcher does not observe.

In this section we specify models in which the future consequences of current decisions are incorporated. For these models, we will assume that the decision maker is fully rational in the sense that he optimizes perfectly in each time period given the information that is available to him at that point in time and given that he knows he will act optimally in the future when future information is revealed. The procedures for modeling these decisions were first developed for various applications by, for example, Wolpin (1984) on women's fertility, Pakes (1986) on patent options, Wolpin (1987) on job search, Rust (1987) on engine replacement, Berkovec and Stern (1991) on retirement, and others. Eckstein and Wolpin (1989) provide an excellent survey of these early contributions. The thrust of more recent work has primarily been toward solving some of the computational difficulties that can arise in these models, as discussed below.

Before embarking on this endeavor, it is important to keep the concept of rationality in perspective. A model of rational decision making over time does not necessarily represent behavior more accurately than a model of myopic behavior, where the decision maker ignores future consequences. In fact, the truth in a given situation might lie between these two extremes: decision makers might be acting in ways that are neither completely myopic nor completely rational. As we will see, the truly optimizing behavior is very complex. People might engage in behavior that is only approximately optimal simply because they (we) can't figure out the truly optimal way to proceed. Viewed in another light, one could argue that people always optimize when the realm of optimization is broadened sufficiently. For example, rules of thumb or other behavior that seem only to approximate optimality may actually turn out to be optimal when the costs of optimization are considered.

The concepts and procedures that are developed to examine optimizing behavior carry over, in modified form, to other types of behavior that recognize future effects of current choices. Furthermore, the researcher can often test alternative behavioral representations. Myopic behavior nearly always appears as a testable restriction on a fully rational model, namely, a zero coefficient for the variable that captures future effects.

Sometimes, the standard rational model is a restriction on a supposedly nonrational one. For example, O'Donoghue and Rabin (1999), among others, argue that people are time-inconsistent: when it is Monday, we weigh the benefits and costs that will come on, say, Wednesday only marginally more than those that will arrive on Thursday, and yet when Wednesday actually arrives, we weigh Wednesday's (today's) benefits and costs far more than Thursday's. Essentially, we have a bias for the present. The standard rational model, where the same discount rate is used between any two periods independent of whether the person is in one of the periods, constitutes a restriction on the time-inconsistent model.

The concepts in this area of analysis are more straightforward than the notation. To develop the concepts with a minimum of notation, we will start with a two-period model in which the decision maker knows the exact effect of first-period choices on the second-period alternatives and utilities. We will then expand the model to more periods and to situations where the decision maker faces uncertainty about future effects.

### *7.7.1. Two Periods, No Uncertainty about Future Effects*

To make the explication concrete, consider a high school student's choice of whether or not to go to college. The choice can be examined in the context of two periods: the college years and the post-college years. In the first period, the student either goes to college or not. Even though these are called the college years, the student need not go to college but can take a job instead. In the second period the student chooses among the jobs that are available to him at that time. Going to college during the college years means less income during that period but better job options in the post-college years. $U_{1C}$ is the utility that the student obtains in period 1 from going to college, and $U_{1W}$ is the utility he obtains in the first period if he works in the first period instead of going to college. If the student were myopic, he would choose college only if $U_{1C} > U_{1W}$. However, we assume that he is not myopic. For the second period, let $J$ denote the set of all possible jobs. The utility of job $j$ in period 2 is $U_{2j}^{C}$ if the student went to college and $U_{2j}^{W}$ if he worked in the first period. The utility from a job depends on the wage that the person is paid as well as other factors. For many jobs, people with a college degree are paid higher wages and granted greater autonomy and responsibility. For these jobs, $U_{2j}^{C} > U_{2j}^{W}$. However, working in the first period provides on-the-job experience that commands higher wages and responsibility than a college degree for some jobs; for these, $U_{2j}^{W} > U_{2j}^{C}$.

A job not being available is represented as having a utility of negative infinity. For example, if job $j$ is available only to college graduates, then $U_{2j}^{W} = -\infty$.

How will the high school student decide whether to go to college? We assume for now that the student knows $U_{2j}^{C}$ and $U_{2j}^{W}$ for all $j \in J$ when deciding whether to go to college in the first period. That is, the student has perfect knowledge of his future options under whatever choice he makes in the first period. We will later consider how the decision process changes when the student is uncertain about these future utilities. The student knows that when the second period arrives he will choose the job that provides the greatest utility. That is, he knows in the first period that the utility that he will obtain in the second period if he chooses college in the first period is the maximum of $U_{2j}^{C}$ over all possible jobs. We label this utility as $U_{2}^{C} = \max_j(U_{2j}^{C})$. The student therefore realizes that, if he chooses college in the first period, his total utility over both periods will be

$$\begin{aligned} \mathrm{TU}_C &= U_{1C} + \lambda U_{2}^{C} \\ &= U_{1C} + \lambda \max_j\left(U_{2j}^{C}\right), \end{aligned}$$

where $\lambda$ reflects the relative weighting of the two periods' utilities in the student's decision process. Given the way we have defined time periods, $\lambda$ incorporates the relative time spans of each period as well as the traditional discounting of future utility relative to current utility. Thus, $\lambda$ can exceed one, even with discounting, if the second period represents say forty years while the first period is four years. Myopic behavior is represented as $\lambda = 0$.

The same logic is applied to the option of working in the first period instead of going to school. The student knows that he will choose the job that offers the greatest utility, so that $U_{2}^{W} = \max_j(U_{2j}^{W})$ and the total utility over both period from choosing to work in the first period is

$$\begin{aligned} \mathrm{TU}_W &= U_{1W} + \lambda U_{2}^{W} \\ &= U_{1W} + \lambda \max_j\left(U_{2j}^{W}\right). \end{aligned}$$

The student chooses college if $\mathrm{TU}_C > \mathrm{TU}_W$ and otherwise chooses to work in the first period.

This completes the description of the decision maker's behavior. We now turn to the researcher. As always, the researcher observes only some of the factors that affect the student's utility. Each utility in each period is decomposed into an observed and unobserved component:

$$\begin{aligned} U_{1C} &= V_{1C} + \varepsilon_{1C}, \\ U_{1W} &= V_{1W} + \varepsilon_{1W} \end{aligned}$$

and

$$U_{2j}^{C} = V_{2j}^{C} + \varepsilon_{2j}^{C},$$
$$U_{2j}^{W} = V_{2j}^{W} + \varepsilon_{2j}^{W}$$

for all $j \in J$. Collect the unobserved components into vector $\varepsilon = \langle \varepsilon_{1C}, \varepsilon_{1W}, \varepsilon_{2j}^{C}, \varepsilon_{2j}^{W}, \forall j \rangle$, and denote the density of these terms as $f(\varepsilon)$. The probability of the student choosing college is

$$\begin{aligned} P_C &= \text{Prob}(\text{TU}_C > \text{TU}_W) \\ &= \text{Prob}\big[U_{1C} + \lambda \max_j \big(U_{2j}^{C}\big) > U_{1W} + \lambda \max_j \big(U_{2j}^{W}\big)\big] \\ &= \text{Prob}\big[V_{1C} + \varepsilon_{1C} + \lambda \max_j \big(V_{2j}^{C} + \varepsilon_{2j}^{C}\big) \\ &> V_{1W} + \varepsilon_{1W} + \lambda \max_j \big(V_{2j}^{W} + \varepsilon_{2j}^{W}\big)\big] \\ &= \int I\big[V_{1C} + \varepsilon_{1C} + \lambda \max_j \big(V_{2j}^{C} + \varepsilon_{2j}^{C}\big) \\ &> V_{1W} + \varepsilon_{1W} + \lambda \max_j \big(V_{2j}^{W} + \varepsilon_{2j}^{W}\big)\big] f(\varepsilon)\, d\varepsilon, \end{aligned}$$

where $I[\cdot]$ is an indicator of whether the statement in brackets is true.

The integral can be approximated through simulation. For an accept–reject simulator:

1. Take a draw from $f(\varepsilon)$, with its components labeled $\varepsilon_{1C}^{r}$, $\varepsilon_{2j}^{Cr}, \ldots$.
2. Calculate $U_{2j}^{C} = V_{2j}^{C} + \varepsilon_{2j}^{Cr}$ for all $j$, determine the highest one, and label it $U_{2}^{Cr}$. Similarly, calculate $U_{2}^{Wr}$.
3. Calculate the total utilities as $\text{TU}_C^r = V_{1C}^r + \varepsilon_{1C}^r + \lambda U_2^{Cr}$, and similarly for $\text{TU}_W^r$.
4. Determine whether $\text{TU}_C^r > \text{TU}_W^r$. If so, set $I^r = 1$. Otherwise, let $I^r = 0$.
5. Repeat steps 1–4 $R$ times. The simulated probability of choosing college is $\tilde{P}_C = \sum_r I^r / R$.

Convenient error partitioning (as explained in Section 1.2) can be utilized to obtain a smooth and more accurate simulator than accept–reject, provided that the integral over the first-period errors has a closed form conditional on the second-period errors. Suppose for example that $\varepsilon_{1C}$ and $\varepsilon_{1W}$ are iid extreme value. Label the second-period errors collectively as $\varepsilon_2$ with any density $g(\varepsilon_2)$. Conditional on the second-period errors, the probability of the student going to college is given by a standard logit model with an extra explanatory variable that captures the future effect of the current choice. That is,

$$P_C(\varepsilon_2) = \frac{e^{V_{1C} + \lambda U_2^{C}(\varepsilon_2)}}{e^{V_{1C} + \lambda U_2^{C}(\varepsilon_2)} + e^{V_{1W} + \lambda U_2^{W}(\varepsilon_2)}},$$

where $U_2^C(\varepsilon_2)$ is calculated from the second-period errors as $U_2^C(\varepsilon_2) = \max_j(V_{2j}^C + \varepsilon_{2j}^C)$, and similarly for $U_2^W(\varepsilon_2)$. The unconditional probability is then the integral of this logit formula over all possible values of the second-period errors:

$$P_C = \int P_C(\varepsilon_2)g(\varepsilon_2)\,d\varepsilon_2.$$

The probability is simulated as follows: (1) Take a draw from density $g(\cdot)$ and label it $\varepsilon_2^r$ . (2) Using this draw of the second-period errors, calculate the utility that would be obtained from each possible job if the person went to college. That is, calculate $U_{2j}^{Cr} = V_{2j}^C + \varepsilon_{2j}^{Cr}$ for all $j$. (3) Determine the maximum of these utilities, and label it $U_2^{Cr}$. This is the utility that the person would obtain in the second period if he went to college in the first period, based on this draw of the second-period errors. (4)–(5) Similarly, calculate $U_{2j}^{Wr}\ \forall j$, and then determine the maximum $U_2^{Wr}$. (6) Calculate the conditional choice probability for this draw as

$$P_C^r = \frac{e^{V_{1C}+\lambda U_2^{Cr}}}{e^{V_{1C}+\lambda U_2^{Cr}} + e^{V_{1W}+\lambda U_2^{Wr}}}.$$

(7) Repeat steps 1–6 many times, labeled $r = 1, \ldots, R$. (8) The simulated probability is $\tilde{P}_C = \sum_r P_C^r/R$.

If the second-period errors are also iid extreme value, then the probability of taking a particular job in the second period is standard logit. The probability of going to college and taking job $j$ is

$$P_{Cj} = \left(\int \left[\frac{e^{V_{1C}+\lambda U_2^C(\varepsilon_2)}}{e^{V_{1C}+\lambda U_2^C(\varepsilon_2)} + e^{V_{1W}+\lambda U_2^W(\varepsilon_2)}}\right] g(\varepsilon_2)d\varepsilon_2\right)\left(\frac{e^{V_{2j}^C}}{\sum_k e^{V_{2k}^C}}\right).$$

The choice probabilities for the first period are simulated by taking draws of the second-period errors, as just described, with $g(\cdot)$ being the extreme value distribution. However, the probabilities for the second period are calculated exactly. The draws of the second-period errors are used only in calculating the first-period probabilities where they do not integrate out in closed form. The second-period errors integrate out of the second-period probabilities in closed form, which is used to calculate the second-period probabilities exactly. Application to other distributions that allow correlation over alternatives, such as GEV or normal, is straightforward. Allowing the errors to be correlated over time can be accomplished with a joint normal distribution and simulation of both periods' probabilities.

*7.7.2. Multiple Periods*

We first expand to three periods and then generalize to any number of periods. The model of college choice can be extended by considering retirement options. When a person reaches retirement age, there are usually several options available. He can continue working full time, or work part time and spend part of his retirement funds, or retire fully and collect social security and perhaps a pension. The person's income under these alternatives depends largely on the job that the person has held and the retirement plan that the job provided. Three periods are sufficient to capture the decision process. The person goes to college or not in the first period, chooses a job in the second period, and chooses among the available retirement-age options in the third period. The high school student knows, when deciding whether to go to college, that this decision will affect his job opportunities, which in turn will affect his retirement options. (This foreknowledge is starting to seem like a mighty big burden for a high school student.)

The set of retirement-age alternatives is labeled $S$, and its elements indexed by $s$. In the third period, the utility that the person obtains from alternative $s$ if he went to college in the first period and had job $j$ in the second period is $U_{3s}^{Cj}$. Conditional on these previous choices, the person chooses option $s$ if $U_{3s}^{Cj} > U_{3t}^{Cj}$ for all $s \neq t$ and $s, t \in S$. Similar notation and behavior apply conditional on other choices in the first and second periods.

In the second period, the person recognizes that his job choice will affect his retirement-age options. He knows he will maximize among the available options when retirement age arrives. Suppose he chose college in the first period. In the second period, he knows that the utility he will obtain in the third period if he chooses job $j$ is $\max_s U_{3s}^{Cj}$. The total utility of choosing job $j$ in the second period, given that he chose college in the first period, is therefore $\mathrm{TU}_j^C = U_{2j}^C + \theta \max_s U_{3s}^{Cj}$, where $\theta$ weights period three relative to period two. He chooses job $j$ if $\mathrm{TU}_j^C > \mathrm{TU}_k^C$ for all $k \neq j$ and $j, k \in J$. Similar notation and behavior occur if he chose to work in the first period.

Consider now the first period. He knows that, if he chooses college, he will choose the job that maximizes his utility from jobs conditional on going to college, and then will choose the retirement-age option that maximizes his utility conditional on that chosen job. The total utility from college is

$$\begin{aligned}\mathrm{TU}_C &= U_{1c} + \lambda \max_j \mathrm{TU}_j^C \\ &= U_{1c} + \lambda \max_j \left(U_{2j}^C + \theta \max_s U_{3s}^{Cj}\right).\end{aligned}$$

This expression is similar to that in the two-period model except that it includes an additional layer of maximization: the maximization for the third period is contained in each maximization for the second period. A similar expression gives the total utility of working in the first period, $\mathrm{TU}_W$. The person chooses college if $\mathrm{TU}_C > \mathrm{TU}_W$.

This completes the description of the person's behavior. The researcher observes a portion of each utility function: $U_{1C}$, $U_{1W}$, $U_{2j}^C$, and $U_{2j}^W \; \forall j \in J$, and $U_{3s}^{Cj}$ and $U_{3s}^{Wj} \; \forall s \in S, \; j \in J$. The unobserved portions are collectively labeled by the vector $\varepsilon$ with density $f(\varepsilon)$. The probability that the person chooses college is

$$P_C = \int I(\varepsilon) f(\varepsilon) \, d\varepsilon,$$

where

$$I(\varepsilon) = 1$$

if

$$V_{1C} + \varepsilon_{1C} + \lambda \max_j \left( V_{2j}^C + \varepsilon_{2j}^C + \theta \max_s \left( V_{3s}^{Cj} + \varepsilon_{3s}^{Cj} \right) \right)$$
$$> V_{1W} + \varepsilon_{1W} + \lambda \max_j \left( V_{2j}^W + \varepsilon_{2j}^W + \theta \max_s \left( V_{3s}^{Wj} + \varepsilon_{3s}^{Wj} \right) \right).$$

This expression is the same as in the two-period model except that now the term inside the indicator function has an extra level of maximization. An accept–reject simulator is obtained: (1) draw from $f(\varepsilon)$; (2) calculate the third-period utility $U_{3s}^{Cj}$ for each $s$; (3) identify the maximum over $s$; (4) calculate $\mathrm{TU}_{2j}^C$ with this maximum; (5) repeat steps (2)–(5) for each $j$, and identify the maximum of $\mathrm{TU}_{2j}^C$ over $j$; (6) calculate $\mathrm{TU}_C$ using this maximum; (7) repeat steps (2)–(6) for $\mathrm{TU}_W$; (8) determine whether $\mathrm{TU}_C > \mathrm{TU}_W$, and set $I = 1$ if it is; (9) repeat steps (1)–(8) many times, and average the results. Convenient error partitioning can also be used. For example if all errors are iid extreme value, then the first-period choice probabilities, conditional on draws of the second- and third-period errors, are logit; the second-period probabilities, conditional on the third-period errors, are logit; and the third-period probabilities are logit.

We can now generalize these concepts and introduce some widely used terminology. Note that the analysis of the person's behavior and the simulation of the choice probabilities by the researcher start with the last period and work backward in time to the first period. This process is called backwards recursion. Suppose there are $J$ alternatives in each of $T$ equal-length time periods. Let a sequence of choices up to

period $t$ be denoted $\{i_1, i_2, \ldots, i_t\}$. The utility that the person obtains in period $t$ from alternative $j$ is $U_{tj}(i_1, i_2, \ldots, i_{t-1})$, which depends on all previous choices. If the person chooses alternative $j$ in period $t$, he will obtain this utility plus the future utility of choices conditioned on this choice. The total utility (current and future) that the person obtains from choosing alternative $j$ in period $t$ is $\mathrm{TU}_{tj}(i_1, i_2, \ldots, i_{t-1})$. He chooses the alternative in the current period that provides the greatest total utility. Therefore the total utility he receives from his optimal choice in period $t$ is $\mathrm{TU}_t(i_1, i_2, \ldots, i_{t-1}) = \max_j \mathrm{TU}_{tj}(i_1, i_2, \ldots, i_{t-1})$. This total utility from the optimal choice at time $t$, $\mathrm{TU}_t$, is called the valuation function at time $t$.

The person chooses optimally in the current period with knowledge that he will choose optimally in the future. This fact establishes a convenient relation between the valuation function in successive periods. In particular,

$$\mathrm{TU}_t(i_1, \ldots, i_{t-1}) = \max_j [U_{jt}(i_1, \ldots, i_{t-1}) + \delta \mathrm{TU}_{t+1}(i_1, \ldots, i_t = j)],$$

where $\delta$ is a parameter that discounts the future. $\mathrm{TU}_{t+1}$ on the right-hand side is the total utility that the person will obtain from period $t+1$ onward if he chooses alternative $j$ in period $t$ (i.e., if $i_t = j$). The equation states that the total utility that the person obtains from optimizing behavior from period $t$ onward, given previous choices, is the maximum over $j$ of the utility from $j$ in period $t$ plus the discounted total utility from optimizing behavior from period $t+1$ onward conditional on choosing $j$ in period $t$. This relation is Bellman's equation (1957) applied to discrete choice with perfect information.

$\mathrm{TU}_{tj}(i_1, \ldots, i_{t-1})$ is sometimes called the conditional valuation function, conditional on choosing alternative $j$ in period $t$. A Bellman equation also operates for this term:

$$\begin{aligned}\mathrm{TU}_{tj}(i_1, \ldots, i_{t-1}) &= U_{jt}(i_1, \ldots, i_{t-1}) \\ &\quad + \delta \max_k [\mathrm{TU}_{t+1,k}(i_1, \ldots, i_t = j)].\end{aligned}$$

Since by definition $\mathrm{TU}_t(i_1, \ldots, i_{t-1}) = \max_j [\mathrm{TU}_{tj}(i_1, \ldots, i_{t-1})]$, the Bellman equation in terms of the conditional valuation function is equivalent to that in terms of the unconditional valuation function.

If $T$ is finite, the Bellman equation can be applied with backward recursion to calculate $\mathrm{TU}_{tj}$ for each time period. At $t = T$, there is no future time period, and so $\mathrm{TU}_{Tj}(i_1, \ldots, i_{T-1}) = U_{Tj}(i_1, \ldots, i_{T-1})$. Then $\mathrm{TU}_{T-1,j}(i_1, \ldots, i_{T-2})$ is calculated from $\mathrm{TU}_{Tj}(i_1, \ldots, i_{T-1})$ using Bellman's equation, and so on forward to $t = 1$. Note that $U_{tj}(i_1, \ldots, i_{t-1})$ must be calculated for each $t$, each $j$, and, importantly,

each possible sequence of past choices, $i_1, \ldots, i_{t-1}$. With $J$ alternatives in $T$ time periods, the recursion requires calculation of $(J^T)T$ utilities (that is, $J^T$ possible sequences of choices, with each sequence containing $T$ one-period utilities). To simulate the probabilities, the researcher must calculate these utilities for each draw of unobserved factors. And these probabilities must be simulated for each value of the parameters in the numerical search for the estimates. This huge computational burden is called the *curse of dimensionality* and is the main stumbling block to application of the procedures with more than a few time periods and/or alternatives. We discuss in the next subsection procedures that have been suggested to avoid or mitigate this curse, after showing that the curse is even greater when uncertainty is considered.

### *7.7.3. Uncertainty about Future Effects*

In the analysis so far we have assumed that the decision maker knows the utility for each alternative in each future time period and how this utility is affected by prior choices. Usually, the decision maker does not possess such foreknowledge. A degree of uncertainty shrouds the future effects of current choices.

The behavioral model can be adapted to incorporate uncertainty. For simplicity, return to the two-period model for our high school student. In the first period, the student does not know for sure the second-period utilities, $U^C_{2j}$ and $U^W_{2j}$ $\forall j$. For example, the student does not know, before going to college, how strong the economy, and hence his job possibilities, will be when he graduates. These utilities can be expressed as functions of unknown factors $U^C_{2j}(e)$, where $e$ refers collectively to all factors in period two that are unknown in period one. These unknown factors will become known (that is, will be revealed) when the student reaches the second period, but are unknown to the person in the first period. The student has a subjective distribution on $e$ that reflects the likelihood that he ascribes to the unknown factors taking a particular realization in the second period. This density is labeled $g(e)$. He knows that, whatever realization of $e$ actually occurs, he will, in the second period, choose the job that gives him the maximum utility. That is, he will receive utility $\max_j U^C_{2j}(e)$ in the second period if he chooses college in the first period and the unknown factors end up being $e$. In the first period, when evaluating whether to go to college, he takes the expectation of this future utility over all possible realizations of the unknown factors, using his subjective distribution over these realizations. The expected utility that he will obtain in the second period if he chooses college in the first period is therefore $\int [\max_j U^C_{2j}(e)] g(e)\, de$. The total expected

utility from choosing college in the first period is then

$$\text{TEU}_C = U_{1C} + \lambda \int \left[\max_j U_{2j}^C(e)\right] g(e)\, de.$$

$\text{TEU}_W$ is defined similarly. The person chooses college if $\text{TEU}_C > \text{TEU}_W$. In the second period, the unknown factors become known, and the person chooses job $j$ if he had chosen college if $U_{2j}^C(e^*) > U_{2k}^C(e^*)$ for all $k \neq j$, where $e^*$ is the realization that actually occurred.

Turning to the researcher, we have an extra complication introduced by $g(e)$, the decision maker's subjective distribution for unknown factors. In addition to not knowing utilities in their entirety, the researcher has only partial knowledge of the decision maker's subjective probability $g(e)$. This lack of information is usually represented through parameterization. The researcher specifies a density, labeled $h(e \mid \theta)$, that depends on unknown parameters $\theta$. The researcher then assumes that the person's subjective density is the specified density evaluated at the true parameters $\theta^*$. That is, the researcher assumes $h(e \mid \theta^*) = g(e)$. Stated more persuasively and accurately: the true parameters are, by definition, the parameters for which the researcher's specified density $h(e \mid \theta)$ becomes the density $g(e)$ that the person actually used. With a sufficiently flexible $h$, any $g$ can be represented as $h$ evaluated at some parameters, which are called the true parameters. These parameters are estimated along with the parameters that enter utility. (Other ways of representing the researcher's lack of knowledge about $g(e)$ can be specified; however, they are generally more complex.)

Utilities are decomposed into their observed and unobserved portions, with the unobserved portions collectively called $\varepsilon$ with density $f(\varepsilon)$. The probability that the person goes to college is

$$\begin{aligned} P_C &= \text{Prob}(\text{TEU}_C > \text{TEU}_W) \\ &= \int I(\text{TEU}_C > \text{TEU}_W) f(\varepsilon) d\varepsilon. \end{aligned}$$

where $\text{TEU}_C$ and $\text{TEU}_W$, by the definitions above, each include an integral over $e$ with density $h(e \mid \theta)$. The probability can be approximated by simulating the integrals in $\text{TEU}_C$ and $\text{TEU}_W$ within the simulation of the integral over $I(\text{TEU}_C > \text{TEU}_W)$. (1) Take a draw of $\varepsilon$. (2a) Take a draw of $e$ from $h(e \mid \theta)$. (2b) Using this draw, calculate the integrands in $\text{TEU}_C$ and $\text{TEU}_W$. (2c) Repeat steps 2a–b many times and average the results. (3) Using the value from 2c, calculate $I(\text{TEU}_C > \text{TEU}_W)$. (4) Repeat steps 1–3 many times and average the results. As the reader can see, the curse of dimensionality grows worse.

Several authors have suggested ways to reduce the computational burden. Keane and Wolpin (1994) calculate the valuation function at

selected realizations of the unknown factors and past choices; they then approximate the valuation function at other realizations and past choices through interpolating from the calculated valuations. Rust (1997) suggests simulating future paths and using the average over these simulated paths as an approximation in the valuation function. Hotz and Miller (1993) and Hotz *et al.* (1993) show that there is a correspondence between the valuation function in each time period and the choice probabilities in future periods. This correspondence allows the valuation functions to be calculated with these probabilities instead of backward recursion.

Each of these procedures has limitations and is applicable only in certain situations, which the authors themselves describe. As Rust (1994) has observed, it is unlikely that a general-purpose breakthrough will arise that makes estimation simple for all forms of dynamic optimization models. Inevitably the researcher will need to make trade-offs in specifying the model to assure feasibility, and the most appropriate specification and estimation method will depend on the particulars of the choice process and the goals of the research. In this regard, I have found that two simplifications are very powerful in that they often provide a large gain in computational feasibility for a relatively small loss (and sometimes a gain) in content.

The first suggestion is for the researcher to consider ways to capture the nature of the choice situation with as few time periods as possible. Sometimes, in fact usually, time periods will need to be defined not by the standard markers, such as the year or month, but rather in a way that is more structural with respect to the decision process. For example, for the high school student deciding whether to go to college, it might seem natural to say that he makes a choice each year among the jobs and schooling options that are available in that year, given his past choices. Indeed, this statement is true: the student does indeed make annual (or even monthly, weekly, daily) choices. However, such a model would clearly face the curse of dimensionality. In contrast, the specification that we discussed earlier involves only two time periods, or three if retirement is considered. Estimation is quite feasible for this specification. In fact, the two-period model might be more accurate than an annual model: students deciding on college probably think in terms of the college years and their post-college options, rather than trying to anticipate their future choices in each future year. McFadden and Train (1996) provide an example of how a dynamic optimization model with only a few well-considered periods can accurately capture the nature of the choice situation.

A second powerful simplification was first noted by Rust (1987). Suppose that the factors that the decision maker does not observe beforehand

are also the factors that the researcher does not observe (either before or after), and that these factors are thought by the decision maker to be iid extreme value. Under this admittedly restrictive assumption, the choice probabilities take a closed form that is easy to calculate. The result can be readily derived for our model of college choice. Assume that the student, when in the first period, decomposes second-period utility into a known and unknown part, e.g., $U_{2j}^C(e) = V_{2j}^C + e_{2j}^C$, and assumes that $e_{2j}^C$ follows an extreme value distribution independent of all else. This unknown factor becomes known to the student in the second period, so that second-period choice entails maximization over known $U_{2j}^C \; \forall j$. However, in the first period it is unknown. Recall from Section 3.5 that the expected maximum of utilities that are iid extreme value takes the familiar log-sum formula. In our context, this result means that

$$E\left(\max_j \left(V_{2j}^C + \varepsilon_{2j}^C\right)\right) = \alpha \ln\left(\sum_j e^{V_{2j}^C}\right),$$

which we can label $\mathrm{LS}_2^C$. $\mathrm{LS}_2^W$ is derived similarly. The person chooses college if then

$$\begin{aligned} \mathrm{TEU}_C &> \mathrm{TEU}_W, \\ U_{1C} + \lambda\,\mathrm{LS}_2^C &> U_{1W} + \lambda\,\mathrm{LS}_2^W. \end{aligned}$$

Note that this decision rule is in closed form: the integral over unknown future factors becomes the log-sum formula. Consider now the researcher. Each first-period utility is decomposed into an observed and an unobserved part ($U_{1C} = V_{1C} + \varepsilon_{1C}$, $U_{1W} = V_{1W} + \varepsilon_{1W}$), and we assume that the unobserved parts are iid extreme value. For the second-period utilities, we make a fairly restrictive assumption. We assume that the part of utility that the researcher does not observe is the same as the part that the student does not know beforehand. That is, we assume $U_{2j}^C = V_{2j}^C + \varepsilon_{2j}^C \; \forall j$, where the researcher's $\varepsilon_{2j}^C$ is the same as the student's $e_{2j}^C$. Under this assumption, the researcher can calculate the log-sum terms for future utility, $\mathrm{LC}_2^C$ and $\mathrm{LS}_2^W$, exactly, since these terms depend only on the observed portion of utility in the second period, $V_{2j}^C \; \forall j$, which is observed by the researcher and known beforehand by the decision maker. The probability of the student choosing college is then

$$\begin{aligned} P_C &= \mathrm{Prob}(\mathrm{TEU}_C > \mathrm{TEU}_W) \\ &= \mathrm{Prob}\left(U_{1C} + \lambda\,\mathrm{LS}_2^C > U_{1W} + \lambda\,\mathrm{LS}_2^W\right) \\ &= \mathrm{Prob}\left(V_{1C} + \varepsilon_{1C} + \lambda\,\mathrm{LS}_2^C > V_{1W} + \varepsilon_{1W} + \lambda\,\mathrm{LS}_2^W\right) \\ &= \frac{e^{V_{1C} + \mathrm{LS}_2^C}}{e^{V_{1C} + \lambda\,\mathrm{LS}_2^C} + e^{V_{1W} + \lambda\,\mathrm{LS}_2^W}}. \end{aligned}$$

The model takes the same form as the upper part of a nested logit model: the first-period choice probability is the logit formula with a log-sum term included as an extra explanatory variable. Multiple periods are handled the same way as multilevel nested logits.

It is doubtful that the researcher, in reality, observes everything that the decision maker knows beforehand. However, the simplification that arises from this assumption is so great, and the curse of dimensionality that would arise otherwise is so severe, that proceeding as if it were true is perhaps worthwhile in many situations.