11 Individual-Level Parameters

11.1 Introduction

Mixed logit and probit models allow random coefficients whose distribution in the population is estimated. Consider, for example, the model in Chapter 6, of anglers' choice among fishing sites. The sites are differentiated on the basis of whether campgrounds are available at the site. Some anglers like having campgrounds at the fishing sites, since they can use the grounds for overnight stays. Other anglers dislike the crowds and noise that are associated with campgrounds and prefer fishing at more isolated spots. To capture these differences in tastes, a mixed logit model was specified that included random coefficients for the campground variable and other site attributes. The distribution of coefficients in the population was estimated. Figure 11.1 gives the estimated distribution of the campground coefficient. The distribution was specified to be normal. The mean was estimated as 0.116, and the standard deviation was estimated as 1.655. This distribution provides useful information about the population. For example, the estimates imply that 47 percent of the population dislike having campgrounds at their fishing sites, while the other 53 percent like having them.

The question arises: where in the distribution of tastes does a particular angler lie? Is there a way to determine whether a given person tends to like or dislike having campgrounds at fishing sites?

A person's choices reveal something about his tastes, which the researcher can, in principle, discover. If the researcher observes that a particular angler consistently chooses sites without campgrounds, even when the cost of driving to these sites is higher, then the researcher can reasonably infer that this angler dislikes campgrounds. There is a precise way for performing this type of inference, given by Revelt and Train (2000).

We explain the procedure in the context of a mixed logit model; however, any behavioral model that incorporates random coefficients can

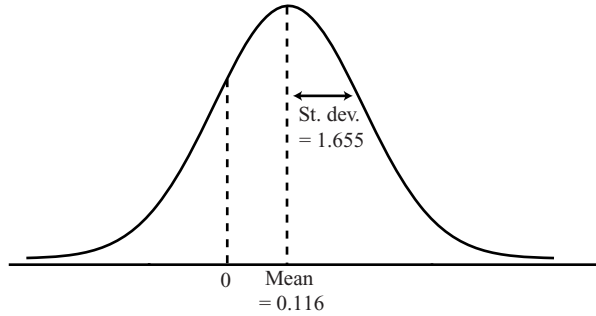

Figure 11.1. Distribution of coefficient of campgrounds in population of all anglers.

be used, including probit. The central concept is a distinction between two distributions: the distribution of tastes in the population, and the distribution of tastes in the subpopulation of people who make particular choices. Denote the random coefficients as vector β. The distribution of β in the population of all people is denoted $g(\beta \mid \theta)$, where θ are the parameters of this distribution, such as the mean and variance.

A choice situation consists of several alternatives described collectively by variables x. Consider the following thought experiment. Suppose everyone in the population faces the same choice situation described by the same variables x. Some portion of the population will choose each alternative. Consider the people who choose alternative i. The tastes of these people are not all the same: there is a distribution of coefficients among these people. Let $h(\beta \mid i, x, \theta)$ denote the distribution of β in the subpopulation of people who, when faced with the choice situation described by variables x, would choose alternative i. Now $g(\beta \mid \theta)$ is the distribution of β in the entire population. $h(\beta \mid i, x, \theta)$ is the distribution of β in the subpopulation of people who would choose alternative i when facing a choice situation described by x.

We can generalize the notation to allow for repeated choices. Let y denote a sequence of choices in a series of situations described collectively by variables x. The distribution of coefficients in the subpopulation of people who would make the sequences of choices y when facing situations described by x is denoted $h(\beta \mid y, x, \theta)$.

Note that $h(\cdot)$ conditions on y, while $g(\cdot)$ does not. It is sometimes useful to call h the conditional distribution and g the unconditional distribution. Two such distributions are depicted in Figure 11.2. If we knew nothing about a person's past choices, then the best we can do in describing his tastes is to say that his coefficients lie somewhere in $g(\beta \mid \theta)$. However, if we have observed that the person made choices y

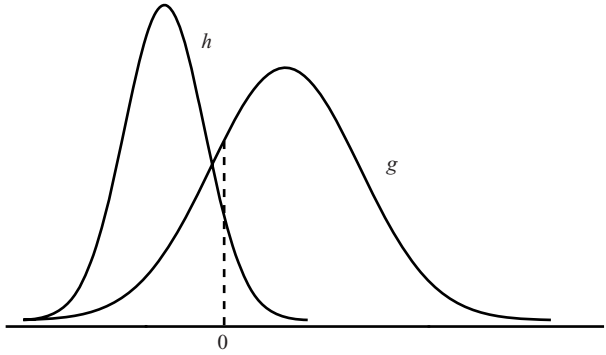

Figure 11.2. Unconditional (population) distribution g and conditional (subpopulation) distribution h for subpopulation of anglers who chose sites without campgrounds.

when facing situations described by x, then we know that that person's coefficients are in the distribution $h(\beta \mid y, x, \theta)$. Since h is tighter than g, we have better information about the person's tastes by conditioning on his past choices.

Inference of this form has long been conducted with linear regression models, where the dependent variable and the distribution of coefficients are both continuous (Griffiths, 1972; Judge *et al.*, 1988). Regime-switching models, particularly in macroeconomics, have used an analogous procedure to assess the probability that an observation is within a given regime (Hamilton and Susmel, 1994; Hamilton, 1996). In these models, the dependent variable is continuous and the distribution of coefficients is discrete (representing one set of coefficients for each regime.) In contrast to both of these traditions, our models have discrete dependent variables. Kamakura and Russell (1989) and DeSarbo *et al.* (1995) developed an approach in the context of a discrete choice model with a discrete distribution of coefficients (that is, a latent class model). They used maximum likelihood procedures to estimate the coefficients for each segment, and then calculated the probability that an observation is within each segment based on the observed choices of the observation. The approach that we describe here applies to discrete choice models with continuous or discrete distributions of coefficients and uses maximum likelihood (or other classical methods) for estimation. The models of Kamakura and Russell (1989) and DeSarbo *et al.* (1995) are a special case of this more general method. Bayesian procedures have been also developed to perform this inference within discrete choice models (Rossi *et al.* 1996; Allenby and Rossi 1999). We describe the Bayesian methods in Chapter 12.

11.2 Derivation of Conditional Distribution

The relation between h and g can be established precisely. Consider a choice among alternatives $j = 1, \ldots, J$ in choice situations $t = 1, \ldots, T$. The utility that person n obtains from alternative j in situation t is

$$U_{njt} = \beta_n' x_{njt} + \varepsilon_{njt},$$

where $\varepsilon_{njt} \sim$ iid extreme value, and $\beta_n \sim g(\beta \mid \theta)$ in the population. The variables x_{njt} can be denoted collectively for all alternatives and choice situations as x_n. Let $y_n = \langle y_{n1}, \ldots, y_{nT} \rangle$ denote the person's sequence of chosen alternatives. If we knew β_n, then the probability of the person's sequence of choices would be a product of logits:

$$P(y_n \mid x_n, \beta) = \prod_{t=1}^{T} L_{nt}(y_{nt} \mid \beta),$$

where

$$L_{nt}(y_{nt} \mid \beta) = \frac{e^{\beta' x_{ny_{nt}t}}}{\sum_j e^{\beta' x_{njt}}}.$$

Since we do not know β_n, the probability of the person's sequence of choices is the integral of $P(y_n \mid x_n, \beta)$ over the distribution of β:

$$P(y_n \mid x_n, \theta) = \int P(y_n \mid x_n, \beta) g(\beta \mid \theta)\, d\beta. \tag{11.1}$$

This is the mixed logit probability that we discussed in Chapter 6.

We can now derive $h(\beta \mid y_n, x_n, \theta)$. By Bayes' rule,

$$h(\beta \mid y_n, x_n, \theta) \times P(y_n \mid x_n, \theta) = P(y_n \mid x_n, \beta) \times g(\beta \mid \theta).$$

This equation simply states that the joint density of β and y_n can be expressed as the probability of y_n times the probability of β conditional on y_n (which is the left-hand side), or with the other direction of conditioning, as the probability of β times the probability of y_n conditional on β (which is the right-hand side.) Rearranging,

$$h(\beta \mid y_n, x_n, \theta) = \frac{P(y_n \mid x_n, \beta) g(\beta \mid \theta)}{P(y_n \mid x_n, \theta)}. \tag{11.2}$$

We know all the quantities on the right-hand side. From these, we can calculate h.

Equation (11.2) also provides a way to interpret h intuitively. Note that the denominator $P(y_n \mid x_n, \theta)$ is the integral of the numerator, as given by

the definition in (11.1). As such, the denominator is a constant that makes h integrate to 1, as required for any density. Since the denominator is a constant, h is proportional to the numerator, $P(y_n \mid x_n, \beta)g(\beta \mid \theta)$. This relation makes interpretation of h relatively easy. Stated in words, the density of β in the subpopulation of people who would choose sequence y_n when facing x_n is proportional to the density of β in the entire population *times* the probability that y_n would be chosen if the person's coefficients were β.

Using (11.2), various statistics can be derived conditional on y_n. The mean β in the subpopulation of people who would choose y_n when facing x_n is

$$\bar{\beta}_n = \int \beta \cdot h(\beta \mid y_n, x_n, \theta)\, d\beta.$$

This mean generally differs from the mean β in the entire population. Substituting the formula for h,

$$\begin{aligned} \bar{\beta}_n &= \frac{\int \beta \cdot P(y_n \mid x_n, \beta) g(\beta \mid \theta)\, d\beta}{P(y_n \mid x_n, \theta)} \\ &= \frac{\int \beta \cdot P(y_n \mid x_n, \beta) g(\beta \mid \theta)\, d\beta}{\int P(y_n \mid x_n, \beta) g(\beta \mid \theta)\, d\beta}. \end{aligned} \tag{11.3}$$

The integrals in this equation do not have a closed form; however, they can be readily simulated. Take draws of β from the population density $g(\beta \mid \theta)$. Calculate the weighted average of these draws, with the weight for draw β^r being proportional to $P(y_n \mid x_n, \beta^r)$. The simulated subpopulation mean is

$$\check{\beta}_n = \sum_r w^r \beta^r,$$

where the weights are

$$w^r = \frac{P(y_n \mid x_n, \beta^r)}{\sum_r P(y_n \mid x_n, \beta^r)}. \tag{11.4}$$

Other statistics can also be calculated. Suppose the person faces a new choice situation described by variables $x_{n\,j\,T+1}\ \forall j$. If we had no information on the person's past choices, then we would assign the following probability to his choosing alternative i:

$$P(i \mid x_{n\,T+1}, \theta) = \int L_{n\,T+1}(i \mid \beta) g(\beta \mid \theta)\, d\beta \tag{11.5}$$

where

$$L_{n\,T+1}(i \mid \beta) = \frac{e^{\beta' x_{ni\,T+1}}}{\sum_j e^{\beta' x_{nj\,T+1}}}.$$

This is just the mixed logit probability using the population distribution of β. If we observed the past choices of the person, then the probability can be conditioned on these choices. The probability becomes

$$P(i \mid x_{n\,T+1}, y_n, x_n, \theta) = \int L_{n\,T+1}(i \mid \beta) h(\beta \mid y_n, x_n, \theta)\, d\beta. \tag{11.6}$$

This is also a mixed logit probability, but using the conditional distribution h instead of the unconditional distribution g. When we do not know the person's previous choices, we mix the logit formula over density of β in the entire population. However, when we know the person's previous choices, we can improve our prediction by mixing over the density of β in the subpopulation who would have made the same choices as this person.

To calculate this probability, we substitute the formula for h from (11.2):

$$P(i \mid x_{n\,T+1}, y_n, x_n, \theta) = \frac{\int L_{n\,T+1}(i \mid \beta) P(y_n \mid x_n, \beta) g(\beta \mid \theta)\, d\beta}{\int P(y_n \mid x_n, \beta) g(\beta \mid \theta)\, d\beta}.$$

The probability is simulated by taking draws of β from the population distribution g, calculating the logit formula for each draw, and taking a weighted average of the results:

$$\check{P}_{n\,i\,T+1}(y_n, x_n, \theta) = \sum_r w^r L_{n\,T+1}(i \mid \beta^r),$$

where the weights are given by (11.4).

11.3 Implications of Estimation of θ

The population parameters θ are estimated in any of the ways described in Chapter 10. The most common approach is maximum simulated likelihood, with the simulated value of $P(y_n \mid x_n, \theta)$ entering the log-likelihood function. An estimate of θ, labeled $\hat{\theta}$, is obtained. We know that there is sampling variance in the estimator. The asymptotic covariance of the estimator is also estimated, which we label $\hat{W}$. The asymptotic distribution is therefore estimated to be $N(\hat{\theta}, \hat{W})$.

The parameter θ describes the distribution of β in the population, giving, for example, the mean and variance of β over all decision makers. For any value of θ, equation (11.2) gives the conditional distribution of β

in the subpopulation of people who would make choices y_n when faced with situations described by x_n. This relation is exact in the sense that there is no sampling or other variance associated with it. Similarly, any statistic based on h is exact given a value of θ. For example, the mean of the conditional distribution, $\bar{\beta}_n$, is exactly equation (11.3) for a given value of θ.

Given this correspondence between θ and h, the fact that θ is estimated can be handled in two different ways. The first approach is to use the point estimate of θ to calculate statistics associated with the conditional distribution h. Under this approach, the mean of the condition distribution, $\bar{\beta}_n$, is calculated by inserting $\hat{\theta}$ into (11.3). The probability in a new choice situation is calculated by inserting $\hat{\theta}$ into (11.6). If the estimator of θ is consistent, then this approach is consistent for statistics based on θ.

The second approach is to take the sampling distribution of $\hat{\theta}$ into consideration. Each possible value of θ implies a value of h, and hence a value of any statistic associated with h, such as $\bar{\beta}_n$. The sampling variance in the estimator of θ induces sampling variance in the statistics that are calculated on the basis of θ. This sampling variance can be calculated through simulation, by taking draws of θ from its estimated sampling distribution and calculating the corresponding statistic for each of these draws.

For example, to represent the sampling distribution of $\hat{\theta}$ in the calculation of $\bar{\beta}_n$, the following steps are taken:

1. Take a draw from $N(\hat{\theta}, \hat{W})$, which is the estimated sampling distribution of $\hat{\theta}$. This step is accomplished as follows. Take K draws from a standard normal density, and label the vector of these draws η^r, where K is the length of θ. Then create $\theta^r = \hat{\theta} + L\eta^r$, where L is the Choleski factor of $\hat{W}$.
2. Calculate $\bar{\beta}_n^r$ based on this θ^r. Since the formula for $\bar{\beta}_n$ involves integration, we simulate it using formula (11.3).
3. Repeat steps 1 and 2 many times, with the number of times labeled R.

The resulting values are draws from the sampling distribution of $\bar{\beta}_n$ induced by the sampling distribution of $\hat{\theta}$. The average of $\bar{\beta}_n^r$ over the R draws of θ^r is the mean of the sampling distribution of $\bar{\beta}_n$. The standard deviation of the draws gives the asymptotic standard error of $\bar{\beta}_n$ that is induced by the sampling variance of $\hat{\theta}$.

Note that this process involves simulation within simulation. For each draw of θ^r, the statistic $\bar{\beta}_n^r$ is simulated with multiple draws of β from the population density $g(\beta \mid \theta^r)$.

Suppose either of these approaches is used to estimate $\bar{\beta}_n$. The question arises: can the estimate of $\bar{\beta}_n$ be considered an estimate of β_n? That is: is the estimated mean of the conditional distribution $h(\beta \mid y_n, x_n, \theta)$, which is conditioned on person n's past choices, an estimate of person n's coefficients?

There are two possible answers, depending on how the researcher views the data-generation process. If the number of choice situations that the researcher can observe for each decision maker is fixed, then the estimate of $\bar{\beta}_n$ is not a consistent estimate of β_n. When T is fixed, consistency requires that the estimate converges to the true value when sample size rises without bound. If sample size rises, but the choice situations faced by person n are fixed, then the conditional distribution and its mean do not change. Insofar as person n's coefficients do not happen to coincide with the mean of the conditional distribution (an essentially impossible event), the mean of the conditional distribution will never equal the person's coefficients no matter how large the sample is. Raising the sample size improves the estimate of θ and hence provides a better estimate of the mean of the conditional distribution, since this mean depends only on θ. However, raising the sample size does not make the conditional mean equal to the person's coefficients.

When the number of choice situations is fixed, then the conditional mean has the same interpretation as the population mean, but for a different, and less diverse, group of people. When predicting the future behavior of the person, one can expect to obtain better predictions using the conditional distribution, as in (11.6), than the population distribution. In the case study presented in the next section, we show that the improvement can be large.

If the number of choice situations that a person faces can be considered to rise, then the estimate of $\bar{\beta}_n$ can be considered to be an estimate of β_n. Let T be the number of choice situations that person n faces. If we observe more choices by the person (i.e., T rises), then we are better able to identify the person's coefficients. Figure 11.3 gives the conditional distribution $h(\beta \mid y_n, x_n, \theta)$ for three different values of T. The conditional distribution tends to move toward the person's own β_n as T rises, and to become more concentrated. As T rises without bound, the conditional distribution collapses onto β_n. The mean of the conditional distribution converges to the true value of β_n as the number of choice situations rises without bound. The estimate of $\bar{\beta}_n$ is therefore consistent for β_n.

In Chapter 12, we describe the Bernstein–von Mises theorem. This theorem states that, under fairly mild conditions, the mean of a posterior distribution for a parameter is asymptotically equivalent to the maximum

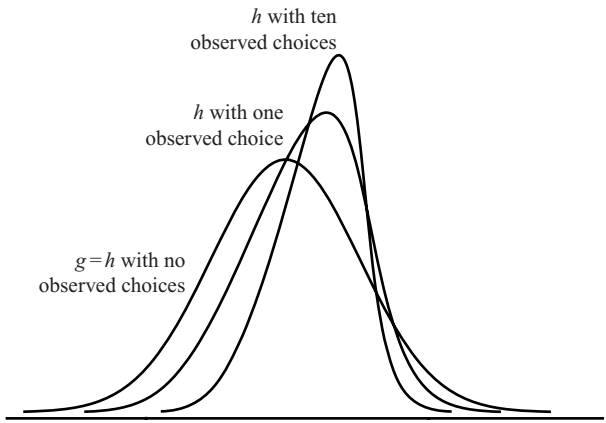

Figure 11.3. Conditional distribution with $T = 0$, 1, and 10.

of the likelihood function. The conditional distribution h is a posterior distribution: by (11.2) h is proportional to a density g, which can be interpreted as a prior distribution on β_n, *times* the likelihood of person n's T choices given β_n, which is $P(y_n \mid x_n, \beta_n)$. By the Bernstein–von Mises theorem, the mean of h is therefore an estimator of β_n that is asymptotically equivalent to the maximum likelihood estimator of β_n, where the asymptotics are defined as T rising. These concepts are described more fully in Chapter 12; we mention them now simply to provide another interpretation of the mean of the conditional distribution.

11.4 Monte Carlo Illustration

To illustrate the concepts, I constructed a hypothetical data set where the true population parameters θ are known as well as the true β_n for each decision maker. These data allow us to compare the mean of the conditional distribution for each decision maker's choices, $\bar{\beta}_n$, with the β_n for that decision maker. It also allows us to investigate the impact of increasing the number of choice situations on the conditional distribution. For this experiment, I constructed data sets consisting of 300 "customers" each facing $T = 1$, 10, 20, and 50 choice situations. There are three alternatives and four variables in each data set. The coefficients for the first two variables are held fixed for the entire population at 1.0, and the coefficients for the last two variables are distributed normal with a mean and variance of 1.0. Utility is specified to include these variables plus a final iid term that is distributed extreme value, so that the model is a mixed logit. The dependent variable for each customer was created by taking a draw from the density of the random terms, calculating the

Table 11.1. *Monte Carlo illustration*

	1st Coef.	2nd Coef.
1 choice situation:		
Standard deviation of $\bar{\beta}_n$	0.413	0.416
Absolute difference between $\bar{\beta}_n$ and β_n	0.726	0.718
10 choice situations:		
Standard deviation of $\bar{\beta}_n$	0.826	0.826
Absolute difference between $\bar{\beta}_n$ and β_n	0.422	0.448
20 choice situations:		
Standard deviation of $\bar{\beta}_n$	0.894	0.886
Absolute difference between $\bar{\beta}_n$ and β_n	0.354	0.350
50 choice situations:		
Standard deviation of $\bar{\beta}_n$	0.951	0.953
Absolute difference between $\bar{\beta}_n$ and β_n	0.243	0.243

utility of each alternative with this draw, and determining which alternative had the highest utility. To minimize the effect of simulation noise in the creation of the data, I constructed 50 datasets for each level of T. The results that are reported are the average over these 50 datasets.

The mean of the conditional distribution for each customer, $\bar{\beta}_n$, was calculated. The standard deviation of $\bar{\beta}_n$ over the 300 customers was calculated, as well as the average absolute deviation of $\bar{\beta}_n$ from the customer's β_n (i.e., the average over n of $|\,\bar{\beta}_n - \beta_n\,|$). Table 11.1 presents these statistics. Consider first the standard deviation. If there were no observed choice situations on which to condition ($T = 0$), then the conditional distribution for each customer would be the unconditional (population) distribution. Each customer would have the same $\bar{\beta}_n$ equal to the population mean of β. In this case, the standard deviation of $\bar{\beta}_n$ would be zero, since all customers have the same $\bar{\beta}_n$. At the other extreme, if we observed an unboundedly large number of choice situations ($T \to \infty$), then the conditional distribution for each customer would collapse to their own β_n. In this case, the standard deviation of $\bar{\beta}_n$ would equal the standard deviation of the population distribution of β_n, which is 1 in this experiment. For T between 0 and ∞, the standard deviation of $\bar{\beta}_n$ is between 0 and the standard deviation of β_n in the population.

In Table 11.1, we see that conditioning on only a few choice situations captures a large share of the variation in β's over customers. With only one choice situation, the standard deviation of $\bar{\beta}_n$ is over 0.4. Since the standard deviation of β_n in the population is 1 in this experiment, which means that conditioning on one choice situation captures over 40 percent of the variation in β_n. With 10 choice situations, over 80 percent

of the variation is captured. There are strongly decreasing returns to observing more choice situations. Doubling from $T = 10$ to $T = 20$ only increases the proportion of variation captured from about .83 to about .89. Increasing T to 50 increases it to about .95.

Consider now the absolute difference between the mean of the customer's conditional distribution, $\bar{\beta}_n$, and the customer's actual β_n. With no conditioning ($T = 0$), the average absolute difference would be 0.8, which is the expected absolute difference for deviates that follow a standard normal as we have in our experiment. With perfect conditioning ($T \to \infty$), $\bar{\beta}_n = \beta_n$ for each customer, and so the absolute difference is 0. With only one choice situation, the average absolute deviation drops from 0.8 (without conditioning) to about 0.72, for a 10 percent improvement. The absolute deviation drops further as the number of choice situations rises.

Notice that the drop in the absolute deviation is smaller than the increase in the standard deviation. For example, with one choice situation the absolute deviation moves 10 percent of the way from no conditioning to perfect knowledge (from .80 with $T = 0$ to .72 with $T = 1$, which is 10 percent of the way to 0 with $T \to \infty$). Yet the standard deviation moves about 40 percent of the way from no conditioning to perfect knowledge (.4 with $T = 1$ is 40 percent of the distance from 0 with $T = 0$ to 1 with $T \to \infty$). This difference is due to the fact that the standard deviation incorporates movement of $\bar{\beta}_n$ away from β_n as well as movement toward β_n. This fact is important to recognize when evaluating the standard deviation of $\bar{\beta}_n$ in empirical applications, where the absolute difference cannot be calculated since β_n is not known. That is, the standard deviation of $\bar{\beta}_n$ expressed as a percentage of the estimated standard deviation in the population is an overestimate of the amount of information that is contained in the $\bar{\beta}_n$'s. With ten choice situations, the average standard deviation in $\bar{\beta}_n$ is over 80 percent of the value that it would have with perfect knowledge, and yet the absolute deviation is less than half as high as would be attained without conditioning.

11.5 Average Conditional Distribution

For a correctly specified model at the true population parameters, the conditional distribution of tastes, aggregated over all customers, equals the population distribution of tastes. Given a series of choice situations described by x_n, there is a set of possible sequences of choices. Label these possible sequences as y_s for $s = 1, \ldots, S$. Denote the true frequency of y_s as $m(y_s \mid x_n, \theta^*)$, expressing its dependence on the true parameters θ^*. If the model is correctly specified and consistently

estimated, then $P(y_s \mid x_n, \hat{\theta})$ approaches $m(y_s \mid x_n, \theta^*)$ asymptotically. Conditional on the explanatory variables, the expected value of $h(\beta \mid y_s, x_n, \hat{\theta})$ is then

$$\begin{aligned} E_y h(\beta \mid y, x_n, \hat{\theta}) &= \sum_s \frac{P(y_s \mid x_n, \beta) g(\beta \mid x_n, \hat{\theta})}{P(y_s \mid x_n, \hat{\theta})} m(y_s \mid x_n, \theta^*) \\ &\to \sum_s P(y_s \mid x_n, \beta) g(\beta \mid x_n, \hat{\theta}) \\ &= g(\beta \mid x_n, \hat{\theta}). \end{aligned}$$

This relation provides a diagnostic tool (Allenby and Rossi 1999). If the average of the sampled customers' conditional taste distributions is similar to the estimated population distribution, the model is correctly specified and accurately estimated. If they are not similar, the difference could be due to (1) specification error, (2) an insufficient number of draws in simulation, (3) an inadequate sample size, and/or (4) the maximum likelihood routine converging at a local rather than global maximum.

11.6 Case Study: Choice of Energy Supplier

11.6.1. Population Distribution

We obtained stated-preference data on residential customers' choice of electricity supplier. Surveyed customers were presented with 8–12 hypothetical choice situations called *experiments*. In each experiment, the customer was presented with four alternative suppliers with different prices and other characteristics. The suppliers differed in price (fixed price given in cents per kilowatthour (c/kWh), TOD prices with stated prices in each time period, or seasonal prices with stated prices in each time period), the length of the contract (during which the supplier is required to provide service at the stated price and the customer would need to pay a penalty for leaving the supplier), and whether the supplier was their local utility, a well-known company other than their local utility, or an unfamiliar company. The data were collected by Research Triangle Institute (1997) for the Electric Power Research Institute and have been used by Goett (1998) to estimate mixed logits. We utilize a specification similar to Goett's, but we eliminate or combine variables that he found to be insignificant.

Two mixed logit models were estimated on these data, based on different specifications for the distribution of the random coefficients. All choices except the last situation for each customer are used to estimate

Table 11.2. *Mixed logit model of energy supplier choice*

	Model 1	Model 2
Price, kWh	−0.8574	−0.8827
	(0.0488)	(0.0497)
Contract length, years		
m	−0.1833	−0.2125
	(0.0289)	(0.0261)
s	0.3786	0.3865
	(0.0291)	(0.0278)
Local utility		
m	2.0977	2.2297
	(0.1370)	(0.1266)
s	1.5585	1.7514
	(0.1264)	(0.1371)
Known company		
m	1.5247	1.5906
	(0.1018)	(0.0999)
s	0.9520	0.9621
	(0.0998)	(0.0977)
TOD rate[a]		
m	−8.2857	2.1328
	(0.4577)	(0.0543)
s	2.5742	0.4113
	(0.1676)	(0.0397)
Seasonal rate[b]		
m	−8.5303	2.1577
	(0.4468)	(0.0509)
s	2.1259	0.2812
	(0.1604)	(0.0217)
Log likelihood at convergence	−3646.51	−3618.92

Standard errors in parentheses.
[a] TOD rates: 11c/kWh, 8 A.M.–8 P.M., 5c/kWh, 8 P.M.–8 A.M.
[b] Seasonal rates: 10c/kWh, summer; 8c/kWh, winter, 6c/kWh, spring and fall.

the parameters of the population distribution, and the customer's last choice situation was retained for use in comparing the predictive ability of different models and methods.

Table 11.2 gives the estimated population parameters. The price coefficient in both models is fixed across the population in such a way that the distribution of willingness to pay for each nonprice attribute (which is the ratio of the attribute's coefficient to the price coefficient) has the same distribution as the attribute's coefficient. For model 1, all of the nonprice coefficients are specified to be normally distributed in the population. The mean *m* and standard deviation *s* of each coefficient are

estimated. For model 2, the first three nonprice coefficients are specified to be normal, and the fourth and fifth are log-normal. The fourth and fifth variables are indicators of TOD and seasonal rates, and their coefficients must logically be negative for all customers. The lognormal distribution (with the signs of the variables reversed) provides for this necessity. The log of these coefficients is distributed normal with mean m and standard deviation s, which are the parameters that are estimated. The coefficients themselves have mean $\exp(m + (s^2/2))$ and standard deviation equal to the mean times $\sqrt{\exp(s^2) - 1}$.

The estimates provide the following qualitative results:

- The average customer is willing to pay about $\frac{1}{5}$ to $\frac{1}{4}$ c/kWh in higher price, depending on the model, in order to have a contract that is shorter by one year. Stated conversely, a supplier that requires customers to sign a four- to five-year contract must discount its price by 1 c/kWh to attract the average customer.
- There is considerable variation in customers' attitudes toward contract length, with a sizable share of customers preferring a longer to a shorter contract. A long-term contract constitutes insurance for the customer against price increases, the supplier being locked into the stated price for the length of the contract. Such contracts, however, prevent the customer from taking advantage of lower prices that might arise during the term of the contract. Apparently, many customers value the insurance against higher prices more than they mind losing the option to take advantage of lower prices. The degree of customer heterogeneity implies that the market can sustain contracts of different lengths with suppliers making profits by writing contracts that appeal to different segments of the population.
- The average customer is willing to pay a whopping 2.5 c/kWh more for its local supplier than for an unknown supplier. Only a small share of customers prefer an unknown supplier to their local utility. This finding has important implications for competition. It implies that entry in the residential market by previously unknown suppliers will be very difficult, particularly since the price discounts that entrants can offer in most markets are fairly small. The experience in California, where only 1 percent of residential customers have switched away from their local utility after several years of open access, is consistent with this finding.
- The average customer is willing to pay 1.8 c/kWh more for a known supplier than for an unknown one. The estimated values of s imply that a sizable share of customers would be willing

to pay more for a known supplier than for their local utility, presumably because of a bad experience or a negative attitude toward the local utility. These results imply that companies that are known to customers, such as their long-distance carriers, local telecommunications carriers, local cable companies, and even retailers like Sears and Home Depot, may be more successful in attracting customers for electricity supply than companies that were unknown prior to their entry as an energy supplier.

- The average customer evaluates the TOD rates in a way that is fairly consistent with TOD usage patterns. In model 1, the mean coefficient of the dummy variable for the TOD rates implies that the average customer considers these rates to be equivalent to a fixed price of 9.7 c/kWh. In model 2, the estimated mean and standard deviation of the log of the coefficient imply a median willingness to pay of 8.4 and a mean of 10.4 c/kWh, which span the mean from model 1. Here 9.5 c/kWh is the average price that a customer would pay under the TOD rates if 75 percent of its consumption occurred during the day (between 8 A.M. and 8 P.M.) and the other 25 percent occurred at night. These shares, while perhaps slightly high for the day, are not unreasonable. The estimated values of s are highly significant, reflecting heterogeneity in usage patterns and perhaps in customers' ability to shift consumption in response to TOD prices. These values are larger than reasonable, implying that a nonnegligible share of customers treat the TOD prices as being equivalent to a fixed price that is higher than the highest TOD price or lower than the lowest TOD price.
- The average customer seems to avoid seasonal rates for reasons beyond the prices themselves. The average customer treats the seasonal rates as being equivalent to a fixed 10 c/kWh, which is the highest seasonal price. A possible explanation for this result relates to the seasonal variation in customers' bills. In many areas, electricity consumption is highest in the summer, when air conditioners are being run, and energy bills are therefore higher in the summer than in other seasons, even under fixed rates. The variation in bills over months without commensurate variation in income makes it more difficult for customers to pay their summer bills. In fact, nonpayment for most energy utilities is most frequent in the summer. Seasonal rates, which apply the highest price in the summer, increase the seasonal variation in bills. Customers would rationally avoid a rate plan that exacerbates

an already existing difficulty. If this interpretation is correct, then seasonal rates combined with bill smoothing (by which the supplier carries a portion of the summer bills over to the winter) could provide an attractive arrangement for customers and suppliers alike.

Model 2 attains a higher log-likelihood value than model 1, presumably because the lognormal distribution assures negative coefficients for the TOD and seasonal variables.

11.6.2. Conditional Distributions

We now use the estimated models to calculate customers' conditional distributions and the means of these distributions. We calculate $\bar{\beta}_n$ for each customer in two ways. First, we calculate $\bar{\beta}_n$ using equation (11.3) with the point estimates of the population parameters, $\hat{\theta}$. Second, we use the procedure in Section 11.3 to integrate over the sampling distribution of the estimated population parameters.

The means and standard deviations of $\bar{\beta}_n$ over the sampled customers calculated by these two methods are given in Tables 11.3 and 11.4, respectively. The price coefficient is not listed in Table 11.3, since it is fixed across the population. Table 11.4 incorporates the sampling distribution of the population parameters, which includes variance in the price coefficient.

Consider the results in Table 11.3 first. The mean of $\bar{\beta}_n$ is very close to the estimated population mean given in Table 11.2. This similarity is expected for a correctly specified and consistently estimated model. The standard deviation of $\bar{\beta}_n$ would be zero if there were no conditioning and would equal the population standard deviation if each customer's coefficient were known exactly. The standard deviations in Table 11.3 are considerably above zero and are fairly close to the estimated population standard deviations in Table 11.2. For example, in model 1, the conditional mean of the coefficient of contract length has a standard deviation of 0.318 over customers, and the point estimate of the standard deviation in the population is 0.379. Thus, variation in $\bar{\beta}_n$ captures more than 70 percent of the total estimated variation in this coefficient. Similar results are obtained for other coefficients. This result implies that the mean of a customer's conditional distribution captures a fairly large share of the variation in coefficients across customers and has the potential to be useful in distinguishing customers.

As discussed in Section 11.5, a diagnostic check on the specification and estimation of the model is obtained by comparing the sample average

Table 11.3. *Average $\bar{\beta}_n$ using point estimate $\hat{\theta}$*

	Model 1	Model 2
Contract length		
Mean	−0.2028	−0.2149
Std. dev.	0.3175	0.3262
Local utility		
Mean	2.1205	2.2146
Std. dev.	1.2472	1.3836
Known company		
Mean	1.5360	1.5997
Std. dev.	0.6676	0.6818
TOD rate		
Mean	−8.3194	−9.2584
Std. dev.	2.2725	3.1051
Seasonal rate		
Mean	−8.6394	−9.1344
Std. dev.	1.7072	2.0560

Table 11.4. *Average $\bar{\beta}_n$ with sampling distribution of $\hat{\theta}$*

	Model 1	Model 2
Price		
Mean	−0.8753	−0.8836
Std. dev.	0.5461	0.0922
Contract length		
Mean	−0.2004	−0.2111
Std. dev.	0.3655	0.3720
Local utility		
Mean	2.1121	2.1921
Std. dev.	1.5312	1.6815
Known company		
Mean	1.5413	1.5832
Std. dev.	0.9364	0.9527
TOD rate		
Mean	−9.1615	−9.0216
Std. dev.	2.4309	3.8785
Seasonal rate		
Mean	−9.4528	−8.9408
Std. dev.	1.9222	2.5615

of the conditional distributions with the estimated population distribution. The means in Table 11.3 represent the means of the sample average of the conditional distributions. The standard deviation of the sample-average conditional distribution depends on the standard deviation of $\bar{\beta}_n$, which is given in Table 11.3, plus the standard deviation of $\beta_n - \bar{\beta}_n$. When this latter portion is added, the standard deviation of each coefficient matches very closely the estimated population standard deviation. This equivalence suggests that there is no significant specification error and that the estimated population parameters are fairly accurate. This suggestion is somewhat tempered, however, by the results in Table 11.4.

Table 11.4 gives the sample mean and standard deviation of the mean of the sampling distribution of $\bar{\beta}_n$ that is induced by the sampling distribution of $\hat{\theta}$. The means in Table 11.4 are the means of the sample average of $h(\beta \mid y_n, x_n, \hat{\theta})$ integrated over the sampling distribution of $\hat{\theta}$. For model 1, a discrepancy occurs that indicates possible misspecification. In particular, the means of the TOD and seasonal rates coefficients in Table 11.4 exceed their estimated population means in Table 11.2. Interestingly, the means for these coefficients in Table 11.4 for model 1 are closer to the analogous means for model 2 than to the estimated population means for model 1 in Table 11.2. Model 2 has the more reasonably shaped lognormal distribution for these coefficients and obtains a considerably better fit than model 1. The conditioning in model 1 appears to be moving the coefficients closer to the values in the better-specified model 2 and away from its own misspecified population distributions. This is an example of how a comparison of the estimated population distribution with the sample average of the conditional distribution can reveal information about specification and estimation.

The standard deviations in Table 11.4 are larger than those in Table 11.3. This difference is due to the fact that the sampling variance in the estimated population parameters is included in the calculations for Table 11.4 but not for Table 11.3. The larger standard deviations do not mean that the portion of total variance in β_n that is captured by variation in $\bar{\beta}_n$ is larger when the sampling distribution is considered than when not.

Useful marketing information can be obtained by examining the $\bar{\beta}_n$ of each customer. The value of this information for targeted marketing has been emphasized by Rossi *et al.* (1996). Table 11.5 gives the calculated $\bar{\beta}_n$ for the first three customers in the data set, along with the population mean of β_n.

The first customer wants to enter a long-term contract, in contrast with the vast majority of customers who dislike long-term contracts. He is willing to pay a higher energy price if the price is guaranteed through a long term contract. He evaluates TOD and seasonal rates very

Table 11.5. *Condition means for three customers*

	Population	Customer 1	Customer 2	Customer 3
Contract length	−0.213	0.198	−0.208	−0.401
Local utility	2.23	2.91	2.17	0.677
Known company	1.59	1.79	2.15	1.24
TOD rates	−9.19	−5.59	−8.92	−12.8
Seasonal rates	−9.02	−5.86	−11.1	−10.9

generously, as if all of his consumption were in the lowest-priced period (note that the lowest price under TOD rates is 5 c/kWh and the lowest price under seasonal rates is 6 c/kWh). That is, the first customer is willing to pay, to be on TOD or seasonal rates, probably more than the rates are actually worth in terms of reduced energy bills. Finally, this customer is willing to pay more than the average customer to stay with the local utility. From a marketing perspective, the local utility can easily retain and make extra profits from this customer by offering a long-term contract under TOD or seasonal rates.

The third customer dislikes seasonal and TOD rates, evaluating them as if all of his consumption were in the highest-priced periods. He dislikes long-term contracts far more than the average customer, and yet, unlike most customers, prefers to receive service from a known company that is not his local utility. This customer is a prime target for capture by a well-known company if the company offers him a fixed price without requiring a commitment.

The second customer is less clearly a marketing opportunity. A well-known company is on about an equal footing with the local utility in competing for this customer. This in itself might make the customer a target of well-known suppliers, since he is less tied to the local utility than most customers. However, beyond this information, there is little beyond low prices (which all customers value) that would seem to attract the customer. His evaluation of TOD and seasonal rates is sufficiently negative that it is unlikely that a supplier could attract and make a profit from the customer by offering these rates. The customer is willing to pay to avoid a long-term contract, and so a supplier could attract this customer by not requiring a contract if other suppliers were requiring contracts. However, if other suppliers were not requiring contracts either, there seems to be little leverage that any supplier would have over its competitors. This customer will apparently be won by the supplier that offers the lowest fixed price.

The discussion of these three customers illustrates the type of information that can be obtained by conditioning on customer's choices, and

how the information translates readily into characterizing each customer and identifying profitable marketing opportunities.

11.6.3. Conditional Probability for the Last Choice

Recall that the last choice situation faced by each customer was not included in the estimation. It can therefore be considered a new choice situation and used to assess the effect of conditioning on past choices. We identified which alternative each customer chose in the new choice situation and calculated the probability of this alternative. The probability was first calculated without conditioning on previous choices. This calculation uses the mixed logit formula (11.5) with the population distribution of β_n and the point estimates of the population parameters. The average of this unconditional probability over customers is 0.353. The probability was then calculated conditioned on previous choices. Four different ways of calculating this probability were used:

1. Based on formula (11.6) using the point estimates of the population parameters.
2. Based on formula (11.6) along with the procedure in Section 11.3 that takes account of the sampling variance of the estimates of the population parameters.

3–4. With the logit formula

$$\frac{e^{\beta_n' x_{n\,i\,T+1}}}{\sum_j e^{\beta_n' x_{n\,j\,T+1}}},$$

with the conditional mean $\bar{\beta}_n$ being used for β_n. This method is equivalent to using the customer's $\bar{\beta}_n$ as if it were an estimate of the customer's true coefficients, β_n. The two versions differ in whether $\bar{\beta}_n$ is calculated on the basis of the point estimate of the population parameters (method 3) or takes the sampling distribution into account (method 4).

Results are given in Table 11.6 for model 2. The most prominent result is that conditioning on each customer's previous choices improves the forecasts for the last choice situation considerably. The average probability of the chosen alternative increases from 0.35 without conditioning to over 0.50 with conditioning. For nearly three-quarters of the 361 sampled customers, the prediction of their last choice situation is better with conditioning than without, with the average probability rising by more than 0.25. For the other customers, the conditioning makes the prediction

Table 11.6. *Probability of chosen alternative in last choice situation*

	Method 1	Method 2	Method 3	Method 4
Average probability	0.5213	0.5041	0.5565	0.5487
Number of customers whose probability rises with conditioning	266	260	268	264
Average rise in probability for customers with a rise	0.2725	0.2576	0.3240	0.3204
Number of customers whose probability drops with conditioning	95	101	93	97
Average fall in probability for customers with a drop	0.1235	0.1182	0.1436	0.1391

in the last choice situations less accurate, with the average probability for these customers dropping.

There are several reasons why the predicted probability after conditioning is not always greater. First, the choice experiments were constructed so that each situation would be fairly different from the other situations, so as to obtain as much variation as possible. If the last situation involves new trade-offs, the previous choices will not be useful and may in fact be detrimental to predicting the last choice. A more appropriate test might be to design a series of choice situations that elicited information on the relevant trade-offs and then design an extra "holdout" situation that is within the range of trade-offs of the previous ones.

Second, we did not include in our model all of the attributes of the alternatives that were presented to customers. In particular, we omitted attributes that did not enter significantly in the estimation of the population parameters. Some customers might respond to these omitted attributes, even though they are insignificant for the population as a whole. Insofar as the last choice situation involves trade-offs of these attributes, the conditional distributions of tastes would be misleading, since the relevant tastes are excluded. This explanation suggests that, if a mixed logit is going to be used for obtaining conditional densities for each customer, the researcher might include attributes that could be important for some individuals even though they are insignificant for the population as a whole.

Third, regardless of how the survey and model are designed, some customers might respond to choice situations in a quixotic manner, such

that the tastes that are evidenced in previous choices are not applied by the customer in the last choice situation.

Last, random factors can cause the probability for some customers to drop with conditioning even when the first three reasons do not.

While at least one of these reasons may be contributing to the lower choice probabilities for some of the customers in our sample, the gain in predictive accuracy for the customers with an increase in probability after conditioning is over twice as great as the loss in accuracy for those with a decrease, and the number of customers with a gain is almost three times as great as the number with a loss.

The third and easiest method, which simply calculates the standard logit formula using the customers' $\bar{\beta}_n$ based on the point estimate of the population parameters, gives the highest probability. This procedure does not allow for the distribution of β_n around $\bar{\beta}_n$ or for the sampling distribution of $\hat{\theta}$. Allowing for either variance reduces the average probability: using the conditional distribution of β_n rather than just the mean $\bar{\beta}_n$ (methods 1 and 2 compared with methods 3 and 4, respectively) reduces the average probability, and allowing for the sampling distribution of $\hat{\theta}$ rather than the point estimate (methods 2 and 4 compared with methods 1 and 3, respectively) also reduces the average probability. This result does not mean that method 3, which incorporates the least variance, is superior to the others. Methods 3 and 4 are consistent only if the number of choice situations is able to rise without bound, so that $\bar{\beta}_n$ can be considered to be an estimate of β_n. With fixed T, methods 1 and 2 are more appropriate, since they incorporate the entire conditional density.

11.7 Discussion

This chapter demonstrates how the distribution of coefficients conditioned on the customer's observed choices are obtained from the distribution of coefficients in the population. While these conditional distributions can be useful in several ways, it is important to recognize the limitations of the concept. First, the use of conditional distributions in forecasting is limited to those customers whose previous choices are observed. Second, while the conditional distribution of each customer can be used in cluster analysis and for other identification purposes, the researcher will often want to relate preferences to observable demographics of the customers. Yet, these observable demographics of the customers could be entered directly into the model itself, so that the population parameters vary with the observed characteristics of the customers in the population. In fact, entering demographics into the model is more direct and more accessible to hypothesis testing than estimating

a model without these characteristics, calculating the conditional distribution for each customer, and then doing cluster and other analyses on the moments of the conditional distributions.

Given these issues, there are three main reasons that a researcher might benefit from calculating customers' conditional distributions. First, information on the past choices of customers is becoming more and more widely available. Examples include scanner data for customers with club cards at grocery stores, frequent flier programs for airlines, and purchases from internet retailers. In these situations, conditioning on previous choices allows for effective targeted marketing and the development of new products and services that match the revealed preferences of subgroups of customers.

Second, the demographic characteristics that differentiate customers with different preferences might be more evident through cluster analysis on the conditional distributions than through specification testing in the model itself. Cluster analysis has its own unique way of identifying patterns, which might in some cases be more effective than specification testing within a discrete choice model.

Third, examination of customers' conditional distributions can often identify patterns that cannot be related to observed characteristics of customers but are nevertheless useful to know. For instance, knowing that a product or marketing campaign will appeal to a share of the population because of their particular preferences is often sufficient, without needing to identify the people on the basis of their demographics. The conditional densities can greatly facilitate analyses that have these goals.