6 Simulation with Qualitative Choice Models

Qualitative choice models operate at the level of individual decisionmakers. However, economists and policy analysts are usually interested in aggregate variables, such as national demand or demand within a state or metropolitan area. Issues concerning the estimation of aggregate variables from qualitative choice models are described in this chapter.

6.1 Aggregation

Introduction

In standard regression models, estimates of aggregate values of the dependent variable are obtained by inserting aggregate values of the explanatory variables. For example, suppose h_n is housing expenditures of person n, y_n is income of person n, and the model relating them is $h_n = \alpha + \beta y_n$. Since this model is linear, the average expenditure on housing is simply calculated as $\alpha + \beta\bar{y}$, where $\bar{y}$ is average income. Similarly, total expenditures on housing within an area (e.g., state) is $\alpha + \beta Y$, where Y is the total income in the area.

Qualitative choice models are not linear in explanatory variables, and, consequently, inserting aggregate values of the explanatory variables into the models will not provide an unbiased estimate of the aggregate value of the dependent variable. Consider a simple binary choice situation in which each household either rents or owns its dwelling. The probability of owning depends only on the household income; assume for convenience that the probability is logit:

$$P_{in} = \exp(\beta y_n)/(1 + \exp(\beta y_n)), \qquad i = \text{owning}.$$

Given this nonlinear model, the average probability of owning, $\bar{P}_i$, is not equal to the logit formula evaluated at average income, $\exp(\beta\bar{y})/(1 + \exp(\beta\bar{y}))$. This inequality is shown graphically in figure 6.1. Households one and two have incomes y_1 and y_2 and ownership probabilities of P_{i1} and P_{i2}, respectively. Their average income is $\bar{y}$. At this average income, the ownership probability given by the logit curve (that is, $\exp(\beta\bar{y})/(1 + \exp(\beta\bar{y}))$) is the value $P(\bar{y})$. This probability is higher (in this example) than the average probability, $\bar{P}_i$, which is the midpoint between P_{i1} and P_{i2}.

Aggregate estimates can be obtained from qualitative choice models in any of several ways. Three of these methods are now described.

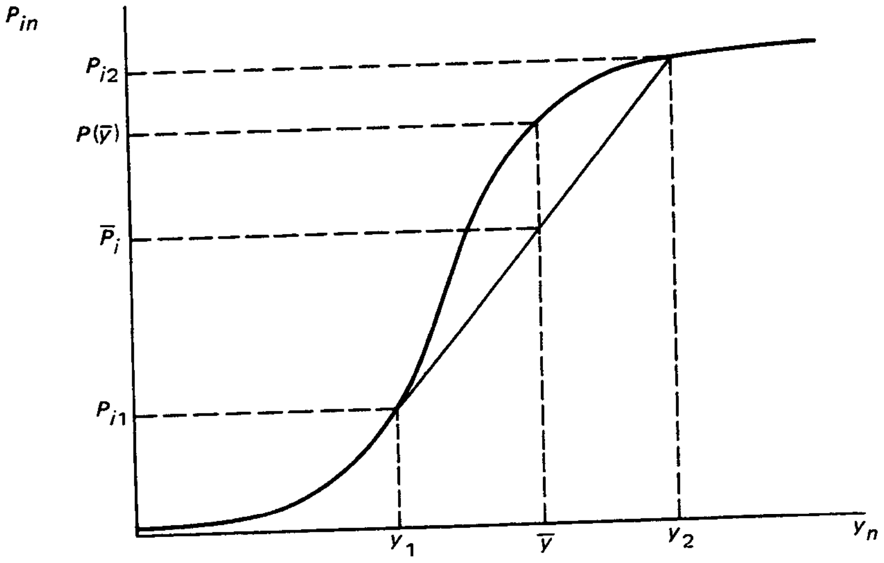

Figure 6.1
Demonstration that $\bar{P}_i \neq \exp(\beta\bar{y})/(1 + \exp(\beta\bar{y}))$, where $\bar{P}_i = (P_{i1} + P_{i2})/2$ and $\bar{y} = (y_1 + y_2)/2$.

Sample Enumeration of Choice Probabilities

The most straightforward, and by far the most popular, approach is sample enumeration, by which the choice probabilities of each decisionmaker in a sample are summed, or averaged, over decisionmakers. Consider a qualitative choice model that gives the probability, P_{in}, that decisionmaker n will choose alternative i from a set of alternatives. Suppose a sample of N decisionmakers, labeled $n = 1, \ldots, N$, was drawn from the population for which aggregate estimates are required. Each sampled decisionmaker has some weight associated with it, representing the number of decisionmakers similar to it in the population (this weight, for samples based on exogenous factors, is the inverse of the probability that the decisionmaker was selected for the sample). Label the weight for decisionmaker n as w_n. Note that if the sample is purely random, then w_n is the same for all n; if the sample is stratified random, then w_n is the same for all n within a stratum.

A consistent estimate of the total number of decisionmakers in the population that choose alternative i (labeled $\hat{N}_i$) is simply the weighted sum of the individual probabilities:

$$\hat{N}_i = \sum_n w_n P_{in}.$$

The average probability for alternative i is

$$\bar{P}_i = \hat{N}_i/N = (1/N)\sum_n w_n P_{in}.$$

Obviously, the average probability is the estimated share of decisionmakers that choose alternative i.

Sample Enumeration of Randomly Generated Choices

Recall from chapter 1 that for qualitative choice models the decisionmaker is assumed to choose the alternative that provides the greatest utility. Utility is composed of an observed and an unobserved part, and assumptions about the distribution of the unobserved component give rise to the choice probabilities. These facts can be utilized in an alternative method of estimating aggregate variables.

Assume each decisionmaker faces a choice among a set J of alternatives, with utility $V_{in} + e_{in}$ associated with each i in J. Representative utility is calculated from observed data and estimated parameters. In the straightforward procedure just described, choice probabilities are calculated from the values of representative utility. Alternatively, the choice of the decisionmaker can be "mimicked" by selecting a value of e_{in} for each i in J from its assumed distribution and observing for which alternative the quantity $V_{in} + e_{in}$ is greatest. That is, for each decisionmaker, a random number generator assigns a value to each e_{in}, for all i in J. The assumed distribution of e is used as the basis of the random number generator (e.g., if the model is logit, then random numbers are generated from the extreme value distribution; if probit, then from the normal). The value of each e_{in} generated in this way is then added to the representative utility for the alternative, and the decisionmaker is considered to choose the alternative with the highest utility.

The total number of decisionmakers in the population who choose a particular alternative is estimated as the weighted sum of sampled decisionmakers who "chose" that alternative:

$$\hat{N}_i = \sum_i w_n D_{in},$$

where D_{in} = one if $V_{in} + e_{in}$ is greater than $V_{jn} + e_{jn}$ for all $j \neq i$, given the calculated values of V_{in} and generated values of e_{in}, and is zero otherwise.

If each sampled decisionmaker's choice is mimicked numerous times, with the random number generator assigning new values of e_{in} each time, then the proportion of times the decisionmaker "chooses" each alternative will approach the choice probability for that alternative. Alternatively, if the sample size is expanded, then the proportion of the decisionmakers with the same values of observed variables who choose a particular alternative will approach the choice probability for that alternative. Therefore, $\hat{N}_i$ is a consistent estimate of the actual number of decisionmakers in the population that choose alternative i.

For a given sample size, sample enumeration on choice probabilities produces more accurate estimates of aggregate variables than sample enumeration with randomly generated choices. However, if the choice probability is complex (e.g., with probit or a complicated GEV structure), then the computer time required to generate random numbers for each alternative might be considerably less than that required to calculate choice probabilities. For given computer costs, therefore, a larger sample is possible if randomly generated choices are used rather than choice probabilities.

Segmentation

When the number of explanatory variables in a qualitative choice model is low, and those variables take only a few values, it is possible to estimate aggregate variables without utilizing a sample of decisionmakers. Consider, for example, a model with only two variables entering the representative utility of each alternative: education level and sex of the decisionmaker. Suppose the education variable consists of four categories: did not complete high school (A), completed high school, but had no college (B), had some college, but did not receive a degree (C), and received a college degree (D). Then the total number of different types of decisionmakers is eight; these eight segments are depicted in figure 6.2 and are labeled $s = 1, \ldots, 8$.

If the researcher has data on the number of people in each segment of the population (i.e., the number of decisionmakers in each cell in figure 6.2), then aggregate variables can be estimated by calculating choice probabilities for each of the eight types of decisionmakers and taking the weighted sum of these choice probabilities. That is, an estimate of the number of decisionmakers in the population who choose alternative i is

$$N_i = \sum_{s=1}^{8} w_s P_{is},$$

	Male	Female
(A) Did not complete high school	1	2
(B) Completed high school, but had no college	3	4
(C) Had some college, but did not receive a degree	5	6
(D) Received a college degree	7	8

Figure 6.2
Segmentation of population.

where P_{is} is the probability that a decisionmaker in segment s (i.e., with a given education level and sex) chooses alternative i, and w_s is the number of decisionmakers in the population who are in segment s.

Note that this procedure is entirely dependent on the researcher knowing the number of decisionmakers in the population who are in each segment. Sometimes this information can be obtained from published population statistics, such as census summaries. Often, however, the information can only be estimated from a sample drawn from the population. In these cases, the procedure does not allow the researcher to avoid taking a sample. However, if the number of segments is smaller than the sample size, then the procedure can reduce computer costs; choice probabilities are calculated for each segment rather than each sampled decisionmaker, and the sample is used simply to estimate the number of decisionmakers in each segment.

6.2 Forecasting

For forecasting into some future year, the same basic procedures described are applied. However, the exogenous variables and/or the weights are adjusted to reflect changes that are anticipated over time. For the sample enumeration procedures, the sample is adjusted in either of these two ways so that it **looks like** a sample that would be drawn in the future year. For example, to forecast the number of people who will choose a given alternative five years in the future, a sample drawn in the current year is adjusted to reflect changes in socioeconomic and other factors that are expected to occur over the next five years. The sample is adjusted in either or both of two ways, (1) by changing the values of the variables relating to each sampled decisionmaker (e.g., increasing each decisionmaker's income to represent real income growth over time) and/or (2) by changing the weight, w_n, attached to each decisionmaker to reflect changes over time in the number of decisionmakers in the population that are similar to the sampled decisionmaker (e.g., increasing the weights for one-person households and decreasing the weights for six-person households to represent expected decreases in household size over time).

For the segmentation approach, changes in explanatory variables over time are represented by changes in the number of decisionmakers in each segment. The explanatory variables themselves cannot logically be adjusted since the distinct values of the explanatory variables define the segments. Changing the variables associated with a decisionmaker in one segment simply shifts the decisionmaker to another segment.

Changing the weights associated with each sampled decisionmaker in the sample enumeration procedure, and adjusting the number of decisionmakers in each segment for the segmentation approach, are essentially the same process. Consider a choice model with one explanatory variable, a dummy indicating whether the decisionmaker is over 30 years old or under 30. Label the number of decisionmakers over and under 30 in the base year as $\mathrm{O30_b}$ and $\mathrm{U30_b}$, respectively, where b denotes the base year (i.e., the year in which the sample used for forecasting was drawn). Suppose that the researcher predicts (or assumes) that in the forecast year the number of decisionmakers over and under 30 will be $\mathrm{O30_f}$ and $\mathrm{U30_f}$, where f denotes the forecast year. For sample enumeration, the appropriate adjustment in weights in this case is the following. For each sampled decisionmaker under 30 in the base year, the weight for the forecast year is calculated as $(\mathrm{U30_f}/\mathrm{U30_b})$ times the decisionmaker's original weight in the base year. Similarly, the forecast year weight for a decisionmaker over 30 is $(\mathrm{O30_f}/\mathrm{O30_b})$ times the base year weight. For the segmentation approach, the number of people in the under 30 segment is considered to be $\mathrm{U30_f}$ instead of $\mathrm{U30_b}$; and similarly for the over 30 segment.

This concept can be generalized to any number of segments. Suppose the explanatory variables in a particular model can take K distinct combinations of values, labeled $k = 1, \ldots, K$, and called segments. Assume the number of decisionmakers in segment k in the base and forecast years is M_b^k and M_f^k, respectively. With sample enumeration, the weight for any sampled decisionmaker who is in segment k in the base year is adjusted by M_f^k/M_b^k for the forecast year. For estimation of aggregates by the segmentation procedure, the choice probabilities for segment k are weighted by M_{f}^{k} in the forecast year rather than M_{b}^{k}.[1]

6.3 Recalibration of Alternative-Specific Constants

Often the representative utility for each alternative in a qualitative choice model includes a constant term, for example,

$$V_{in} = \beta z_{in} + \alpha_i,$$

where z_{in} is a vector of variables relating to decisionmaker n's utility for alternative i, β is a vector of parameters, and α_i is a scalar parameter. The true value of α_i is the mean of all factors that affect the utility of alternative i but are not included in the vector z_i (see section 2.3).

The value of α_i for each i is estimated along with β on the sample used for estimation. However, if, in simulation, the model is run on a sample from a different area or different time than the sample used for estimation (e.g., if the forecasting sample is drawn from one state while the estimation sample was nationwide, or the forecast sample is drawn in 1984 while the estimation sample was drawn in 1980), then the value of α_i for each i will need to be reestimated to reflect the fact that the mean of unincluded variables in the area for which forecasts are made is not the same as those in the area from which the estimation sample was drawn.[2]

The α_i for all i are recalibrated with an iterative procedure that utilizes information on the number of decisionmakers that actually chose alternative i in the forecast area in some base year. The procedure can be described as follows. Let S_i denote the number of decisionmakers that chose alternative i in the forecast area in the base year. Run the model with its original values of α_i for all i on the sample of decisionmakers for the forecast area and estimate the number of decisionmakers to choose alternative i; label the predicted number for alternative i as N_i^{o}, where the superscript o denotes that these predictions are based on the original values of the α_i.

The next step is to compare the proportion of decisionmakers predicted to choose each alternative with the proportion who actually did. That is, let the predicted and actual proportions be denoted

$$n_i^{\mathrm{o}} = N_i^{\mathrm{o}} / \sum_j N_j^{\mathrm{o}}; \qquad s_i = S_i / \sum_j S_j.$$

The model with its original values of the α_i is overpredicting alternative i if n_i^o is larger than s_i, and underpredicting if s_i is larger than n_i^o. This misprediction can be attributed to the fact that the original α_i for all i represent the mean of unincluded variables in the estimation area rather than in the forecast area. Consequently, each α_i should be corrected. In particular, each α_i is adjusted to new values using the formula

$$\alpha_i^1 = \alpha_i^o + \ln(s_i/n_i^o),$$

where α_i^o is the original value of α_i and α_i^1 is the first adjusted value. Note that if s_i is larger than n_i^o, and the model is underpredicting alternative i, then the adjustment increases the value of α_i, thereby increasing its desirability as measured by V_i. Conversely, if n_i^o is greater than s_i, the model is overpredicting alternative i, and the adjustment decreases α_i and hence the representative utility of alternative i.

The adjustment just described completes the first iteration of the recalibration procedure. For the second iteration, the model is run with the new values of α_i (that is, the α_i^1) and new predictions are obtained. Label the proportion of decisionmakers that are predicted to choose alternative i with these new α_i^1 as n_i^1. Compare n_i^1 with s_i for all i. If these values are close, then use the α_i^1 as the final recalibrated values. If n_i^1 and s_i are not close for all i, then adjust each α_i by the formula

$$\alpha_i^2 = \alpha_i^1 + \ln(s_i/n_i^1),$$

where α_i^2 is the twice-adjusted value of α_i. Continue this process, obtaining new values of the α_i with each iteration, until the predicted proportion for each alternative is close to the actual proportion.

6.4 Pivot Point Analysis with Logit Models

The standard way to analyze policies and "what if" situations with a qualitative choice model is to simulate demand with the model twice, once with "base case" values for explanatory variables (i.e., observed values for the base year and assumed values for the forecast years) and a second time with one or more of the explanatory variables changed to represent the policy or situation being examined. For example, the effect of a gas tax on automobile demand is usually assessed by simulating aggregate demand for each class of vehicle using expected gas prices, and then estimating demand again with higher gas prices (representing the tax) entering the model. The difference in the two simulation results is the estimated impact of the gasoline tax.

This standard procedure is the most accurate and is applicable for all qualitative choice models. If the model is logit, however, an approach called "pivot point analysis" is sometimes used instead. It is much easier and less expensive than the standard approach. And, for small changes in explanatory variables, it is perhaps not too inaccurate.

The method is based on the derivatives of the logit formula. Suppose a researcher is interested in examining the impact of a change in a particular explanatory variable affecting the utility of alternative i for all decisionmakers. For decisionmaker n, this variable is labeled X_{in}. If the choice probabilities are logit, the change in the probability that decisionmaker n will choose alternative i (see section 2.4) is

$$\partial P_{in}/\partial X_{in} = (\partial V_{in}/\partial X_{in})P_{in}(1 - P_{in}).$$

That is, the researcher can estimate the effect of the change in X_{in} by running the model only once, to obtain the choice probabilities prior to the change. By knowing the derivative of representative utility with respect to X_{in} (e.g., if V_{in} is linear in X_{in}, then $\partial V_{in}/\partial X_{in}$ is simply the estimated coefficient of X_{in}), the researcher calculates the impact on decisionmaker n by the formula just given. The impact at the aggregate level is similarly determined:

$$\partial \hat{N}_i/\partial X_{in\,\text{for all}\,n} = \sum_n w_n(\partial V_{in}/\partial X_{in})P_{in}(1 - P_{in}),$$

where $\hat{N}_i$ is the estimated number of decisionmakers who choose alternative i.

When a sample of decisionmakers is unavailable for the area for which policy analysis is being performed, a common practice has been to estimate the impact of changes in explanatory variables by applying pivot point analysis to the average probabilities. That is, the change in the proportion of decisionmakers choosing alternative i is estimated as

$$((\partial \bar{V}_i/\partial \bar{X}_i)\bar{P}_i(1 - \bar{P}_i)),$$

where $\bar{X}_i$, $\bar{V}_i$, and $\bar{P}_i$ are the averages of X_{in}, V_{in}, and P_{in}, respectively, over all n. $\bar{P}_i$ can be observed from aggregate data. It is simply the proportion of decisionmakers who actually chose alternative i. The quantity $\partial \bar{V}_i/\partial \bar{X}_i$ can also be known in many cases without sample information; for example, if V_{in}

is linear in X_{in}, then $\partial \bar{V}_i/\partial \bar{X}_i$ is simply the coefficient of X_{in}. Thus, the change in average probabilities is estimated without a sample.

This procedure does not produce a consistent estimate of the impact of changes in explanatory variables. It misses the fundamental point that the average probability is not the probability calculated at the average of the explanatory variables and similarly that the derivative of the average probability is not the derivative of the probability calculated at the average explanatory variables.