change, it drops out of the difference and can therefore be ignored when calculating changes in consumer surplus.

To calculate the change in consumer surplus, the researcher must know or have estimated the marginal utility of income α_n. Usually a price or cost variable enters representative utility, in which case the negative of its coefficient is α_n by definition. (A price or cost coefficient is negative; the negative of a negative coefficient gives a positive α_n). For example, in the choice between car and bus, utility is $U_{nj} = \beta_1 t_{nj} + \beta_2 c_{nj}$ where t is time, c is cost, and both β_1 and β_2 are negative, indicating that utility decreases as the time or cost for a trip increases. The negative of the cost coefficient, $-\beta_2$, is the amount that utility rises due to a one-dollar decrease in costs. A one-dollar reduction in costs is equivalent to a one-dollar increase in income, since the person gets to spend the dollar that he saves in travel costs just the same as if he got the extra dollar in income. The amount $-\beta_2$ is therefore the increase in utility from a one-dollar increase in income: the marginal utility of income. It is the same amount in this case for all n. If c_{nj} entered representative utility interacting with characteristics of the person other than income, such as $c_{nj}H_n$ where H_n is household size, then the marginal utility of income would be $-\beta_2 H_n$, which varies over n.

Throughout this discussion, α_n has been assumed to be fixed for a given person independent of their income. The formula (3.10) for expected consumer surplus depends critically on the assumption that the marginal utility of income is constant with respect to income. If the marginal utility of income changes with income, then a more complicated formula is needed, since α_n itself becomes a function of the change in attributes. McFadden (1999) and Karlstrom (2000) provide procedures for calculating changes in consumer surplus under these conditions.

The conditions for using expression (3.10) are actually less severe than stated. Since only changes in consumer surplus are relevant for policy analysis, formula (3.10) can be used if the marginal utility of income is constant over the range of implicit income changes that are considered by the policy. Thus, for policy changes that change consumer surplus by small amounts per person relative to their income, the formula can be used even though the marginal utility of income in reality varies with income.

The assumption that α_n does not depend on income has implication for the specification of representative utility. As discussed above, α_n is usually taken as the absolute value of the coefficient of price or cost. Therefore, if the researcher plans to use her model to estimate changes in consumer surplus and wants to apply formula (3.10), this coefficient cannot be specified to depend on income. In the mode choice example, cost can be multiplied by household size such that the cost coefficient, and hence the marginal utility of income, varies over households of different size. However, if cost is divided by the household's income, then the coefficient of cost depends on income, violating the assumption needed for expression (3.10). This violation might not be important for small changes in consumer surplus, but certainly becomes important for large changes.

3.6 Derivatives and Elasticities

Since choice probabilities are a function of observed variables, it is often useful to know the extent to which these probabilities change in response to a change in some observed factor. For example, in a household's choice of make and model of car to buy, a natural question is: to what extent will the probability of choosing a given car increase if the vehicle's fuel efficiency is improved? From competing manufacturers points of view, a related question is: to what extent will the probability of households' choosing, say, a Toyota decrease if the fuel efficiency of a Honda improves.

To address these questions, derivatives of the choice probabilities are calculated. The change in the probability that decision-maker n chooses alternative i given a change in an observed factor, z_{ni}, entering the representative utility of that alternative (and holding the representative utility of other alternatives constant) is

$$\begin{aligned}
\frac{\partial P_{ni}}{\partial z_{ni}} &= \frac{\partial[(e^{V_{ni}})/(\sum_j e^{V_{nj}})]}{\partial z_{ni}} \\
&= [e^{V_{ni}}/\sum e^{V_{nj}}](\frac{\partial V_{ni}}{\partial z_{ni}}) - [e^{V_{ni}}/(\sum e^{V_{nj}})^2](e^{V_{ni}})(\frac{\partial V_{ni}}{\partial z_{ni}}) \\
&= \frac{\partial V_{ni}}{\partial z_{ni}}(P_{ni} - P_{ni}^2) \\
&= \frac{\partial V_{ni}}{\partial z_{ni}} P_{ni}(1 - P_{ni}).
\end{aligned}$$

If representative utility is linear in z_{ni} with coefficient β_z, the derivative becomes $\beta_z P_{ni}(1 - P_{ni})$. This derivative is largest when $P_{ni} = 1 - P_{ni}$, which occurs when $P_{ni} = .5$. It becomes smaller as P_{ni} approaches zero or one. The sigmoid probability curve in Figure 3.1 is consistent with these facts. Stated intuitively, the effect of a change in an observed variable is highest when the choice probabilities indicate a high degree of uncertainty regarding the choice. As the choice becomes more certain (i.e., the probabilities approach zero or one), the effect of a change in an observed variable lessens.

One can also determine the extent to which the probability of choosing a particular alternative changes when an observed variable relating to *another* alternative changes. Let z_{nj} denote an attribute of alternative j. How does the probability of choosing alternative i change as z_{nj} increases?

$$\begin{aligned}
\frac{\partial P_{ni}}{\partial z_{nj}} &= \frac{\partial[(e^{V_{ni}})/(\sum_k e^{V_{nk}})]}{\partial z_{nj}} \\
&= -[(e^{V_{ni}})/(\sum e^{V_{nk}})^2](e^{V_{nj}})(\frac{\partial V_{nj}}{\partial z_{nj}}) \\
&= -\frac{\partial V_{nj}}{\partial z_{nj}} P_{ni} P_{nj}.
\end{aligned}$$

When V_{nj} is linear in z_{nj} with coefficient β_z, then this cross-derivative becomes $-\beta_z P_{ni} P_{nj}$. If z_{nj} is a desirable attribute such that β_z is positive, then raising z_{nj} decreases the probability of choosing each alternative other than j. Furthermore, the decrease in probability is proportional to the value of the probability before z_{nj} was changed.

A logically necessary aspect of derivatives of choice probabilities is that, when an observed variable changes, the changes in the choice probabilities sum to zero. This is a consequence of the fact that the probabilities must sum to one before and after the change; it is demonstrated for logit models as follows:

$$\begin{aligned}
\sum_{i=1}^{J} \frac{\partial P_{ni}}{\partial z_{nj}} &= (\frac{\partial V_{nj}}{\partial z_{nj}}) P_{nj}(1 - P_{nj}) + \sum_{i \neq j}(-\frac{\partial V_{nj}}{\partial z_{nj}}) P_{nj} P_{ni} \\
&= (\frac{\partial V_{nj}}{\partial z_{nj}}) P_{nj} \left[(1 - P_{nj}) - \sum_{i \neq j} P_{ni} \right] \\
&= (\frac{\partial V_{nj}}{\partial z_{nj}}) P_{nj}[(1 - P_{nj}) - (1 - P_{nj})]
\end{aligned}$$

$$= 0.$$

In practical terms, if one alternative is improved so that the probability of its being chosen increases, the additional probability is necessarily "drawn" from other alternatives. To increase the probability of one alternative necessitates decreasing the probability of another alternative. While obvious, this fact is often forgotten by planners who want to improve demand for one alternative without reducing demand for other alternatives.

Economists often measure response by elasticities rather than derivatives, since elasticities are normalized for the variables' units. An elasticity is the percent change in one variable that is associated with a percent change in another variable. The elasticity of P_{ni} with respect to z_{ni}, a variable entering the utility of alternative i, is

$$\begin{aligned} E_{iz_{ni}} &= (\partial P_{ni}/\partial z_{ni})(z_{ni}/P_{ni}) \\ &= \frac{\partial V_{ni}}{\partial z_{ni}} P_{ni}(1-P_{ni})(z_{ni}/P_{ni}) \\ &= \frac{\partial V_{ni}}{\partial z_{ni}} z_{ni}(1-P_{ni}). \end{aligned}$$

If representative utility is linear in z_{ni} with coefficient β_z, then $E_{iz_{ni}} = \beta_z z_{ni}(1-P_{ni})$.

The cross-elasticity of P_{ni} with respect to a variable entering alternative j is

$$\begin{aligned} E_{iz_{nj}} &= (\partial P_{ni}/\partial z_{nj})(z_{nj}/P_{ni}) \\ &= -\frac{\partial V_{nj}}{\partial z_{nj}} z_{nj} P_{nj}. \end{aligned}$$

which in the case of linear utility reduces to $E_{iz_{nj}} = -\beta_z z_{nj} P_{nj}$. As discussed in section (3.3.2), this cross elasticity is the same for all i: a change in an attribute of alternative j changes the probabilities for all other alternatives by the same percent. This property of the logit cross-elasticities is a manifestation, or re-statement, of the IIA property of the logit choice probabilities.

3.7 Estimation

Manski and McFadden (1981) and Cosslett (1981) describe estimation methods under a variety of sampling procedures. We discuss in this section estimation under the most prominent of these sampling schemes. We first describe estimation when the sample is exogenous and all alternatives are used in estimation. We then discuss estimation on a subset of alternatives and with certain types of choice-based (i.e., non-exogenous) samples.

3.7.1 Exogenous sample

Consider first the situation in which the sample is exogenously drawn, that is, is either random or stratified random with the strata defined on factors that are exogenous to the choice being analyzed. If the sampling procedure is related to the choice being analyzed (for example, if mode choice is being examined and the sample is drawn by selecting people on buses and pooling them with people selected at toll booths), then more complex estimation procedures are generally required, as discussed in the next section. We also assume that the explanatory variables are exogenous to the choice situation. That is, the variables entering representative utility are independent of the unobserved component of utility.

A sample of N decision-makers is obtained for the purposes of estimation. Since the logit probabilities take a closed form, the traditional maximum likelihood procedures can be applied. The probability of person n choosing the alternative that he was actually observed to choose can be expressed as

$$\prod_i (P_{ni})^{y_{ni}},$$

where $y_{ni} = 1$ if person n chose i and zero otherwise. Note that since $y_{ni} = 0$ for all nonchosen alternatives and P_{ni} raised to the power of zero is 1, this term is simply the probability of the chosen alternative.

Assuming that each decision-maker's choice is independent of that of other decision-makers, the probability of each person in the sample choosing the alternative that he was observed actually to choose is

$$L(\beta) = \prod_{n=1}^{N} \prod_i (P_{ni})^{y_{ni}}$$

where β is a vector containing the parameters of the model. The log-likelihood function is then

$$LL(\beta) = \sum_{n=1}^{N} \sum_{i} y_{ni} ln(P_{ni}) \tag{3.11}$$

and the estimator is the value of β that maximizes this function. McFadden (1974) shows that $LL(\beta)$ is globally concave for linear-in-parameters utility, and many statistical packages are available for estimation of these models. When parameters enter representative utility non-linearly, the researcher might need to write her own estimation code using the procedures described in Chapter 8.

Maximum likelihood estimation in this situation can be re-expressed and re-interpreted in a way that assists in understanding the nature of the estimates. At the maximum of the likelihood function, its derivative with respect to each of the parameters is zero:

$$\frac{dLL(\beta)}{d\beta} = 0. \tag{3.12}$$

The maximum likelihood estimates are therefore the values of β that satisfy this first-order condition. For convenience, let representative utility be linear in parameters: $V_{nj} = \beta' x_{nj}$. This specification is not required, but makes the notation and discussion more succinct. Using (3.11) and the formula for the logit probabilities, we show at the end of this subsection that the first order condition (3.12) becomes:

$$\sum_{n} \sum_{i} (y_{ni} - P_{ni}) x_{ni} = 0. \tag{3.13}$$

Rearranging and dividing both sides by N, we have

$$\frac{1}{N} \sum_{n} \sum_{i} y_{ni} x_{ni} = \frac{1}{N} \sum_{n} \sum_{i} P_{ni} x_{ni} \tag{3.14}$$

This expression is readily interpretable. Let $\bar{x}$ denote the average of x over the chosen alternatives by the sampled individuals: $\bar{x} = (1/N) \sum_n \sum_i y_{ni} x_{ni}$. Let $\hat{x}$ be the average of x over the predicted choices of the sampled decision-makers: $\hat{x} = (1/N) \sum_n \sum_i P_{ni} x_{ni}$. The observed average of x in the sample is $\bar{x}$, while $\hat{x}$ is the predicted average. By (3.14), these two averages equal each other at the maximum likelihood estimates. That is: the maximum likelihood estimates of β

are those that make the predicted average of each explanatory variable equal to the observed average in the sample. In this sense, the estimates induce the model to reproduce the observed averages in the sample.

This property of the maximum likelihood estimator for logit models takes a special meaning for the alternative-specific constants. An alternative-specific constant is the coefficient of a dummy variable that identifies an alternative. A dummy for alternative j is a variable whose value in the representative utility of alternative i is $d_i^j = 1$ for $i = j$ and zero otherwise. By (3.14), the estimated constant is the one that gives

$$\begin{aligned} \frac{1}{N}\sum_n \sum_i y_{ni} d_i^j &= \frac{1}{N}\sum_n \sum_i P_{ni} d_i^j \\ S_j &= \hat{S}_j \end{aligned}$$

where S_j is the share of people in the sample who chose alternative j and $\hat{S}_j$ is the predicted share for alternative j. With alternative-specific constants, the predicted shares for the sample equal the observed shares. The estimated model is therefore correct on average within the sample. This feature is similar to the function of a constant in a linear regression model, where the constant assures that the average of the predicted value of the dependent variable equals its observed average in the sample.

The first-order condition (3.13) provides yet another important interpretation. The difference between a person's actual choice, y_{ni}, and the probability of that choice, P_{ni}, is a modeling error, or residual. The left hand side of (3.13) is the sample covariance of the residuals with the explanatory variables. The maximum likelihood estimates are therefore the values of the β's that make this covariance zero, that is, make the residuals uncorrelated with the explanatory variables. This condition for logit estimates is the same as applies in linear regression models. For a regression model $y_n = \beta' x_n + \varepsilon_n$, the ordinary least squares estimates are the values of β that set $\sum_n (y_n - \beta' x_n) x_n = 0$. This fact is verified by solving for β: $\beta = (\sum_n x_n x_n')^{-1} (\sum_n x_n y_n)$, which is the formula for the ordinary least squares estimator. Since $y_n - \beta' x_n$ is the residual in the regression model, the estimates make the residuals uncorrelated with the explanatory variables.

Under this interpretation, the estimates can be motivated as providing a sample-analog to population characteristics. We have assumed that the explanatory variables are exogenous, meaning that they are uncorrelated in the population with the model errors. Since the variables and errors are uncorrelated in the population, it makes sense to choose estimates that make the variables and residuals uncorrelated in the sample. The estimates do exactly that: they provide a model that reproduces in the sample the zero covariances that occur in the population.

Estimators that solve equations of the form (3.13) are called method-of-moments estimators, since they use moment conditions (correlations in this case) between residuals and variables to define the estimator. We will return to these estimators when discussing simulation-assisted estimation in Chapter 10.

We asserted without proof that (3.13) is the first-order condition for the maximum likelihood estimator of the logit model. We give that proof now. The log-likelihood function (3.11) can be re-expressed as

$$
\begin{aligned}
LL(\beta) &= \sum_n \sum_i y_{ni} ln(P_{ni}) \\
&= \sum_n \sum_i y_{ni} ln(\frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}}) \\
&= \sum_n \sum_i y_{ni}(\beta' x_{ni}) - \sum_n \sum_i y_{ni} ln(\sum_j e^{\beta' x_{nj}}).
\end{aligned}
$$

The derivative of the log-likelihood function then becomes

$$
\begin{aligned}
\frac{dLL(\beta)}{d\beta} &= \frac{\sum_n \sum_i y_{ni}(\beta' x_{ni})}{d\beta} - \frac{\sum_n \sum_i y_{ni} ln(\sum_j e^{\beta' x_{nj}})}{d\beta} \\
&= \sum_n \sum_i y_{ni} x_{ni} - \sum_n \sum_i y_{ni} \sum_j P_{nj} x_{nj} \\
&= \sum_n \sum_i y_{ni} x_{ni} - \sum_n (\sum_j P_{nj} x_{nj}) \sum_i y_{ni} \\
&= \sum_n \sum_i y_{ni} x_{ni} - \sum_n (\sum_j P_{nj} x_{nj}) \\
&= \sum_n \sum_i (y_{ni} - P_{ni}) x_{ni}
\end{aligned}
$$

Setting this derivative to zero gives the first-order condition (3.13).

Estimation on a subset of alternatives

In some situations, the number of alternatives facing the decision-maker is so large that estimating model parameters is very expensive or even impossible. With a logit model, estimation can be performed on a subset of alternatives without inducing inconsistency. For example, a researcher examining a choice situation that involves 100 alternatives can estimate on a subset of 10 alternatives for each sampled decision-maker, with the person's chosen alternative included as well as 9 alternatives randomly selected from the remaining 99. If all alternatives have the same chance of being selected into the subset, then estimation proceeds on the subset of alternatives as if it were the full set. If alternatives have unequal probability of being selected, more complicated estimation procedures may be required. The procedure is described as follows.

Suppose that the researcher has used some specific method for randomly selecting alternatives into the subset that is used in estimation for each sampled decision-maker. Denote the full set of alternatives as F and a subset of alternatives as K. Let $q(K \mid i)$ be the probability under the researcher's selection method that subset K is selected given that the decision-maker chose alternative i. Assuming that the subset necessarily includes the chosen alternative, $q(K \mid i) = 0$ for any K that does not include i. The probability that person n chooses alternative i from the full set is P_{ni}. Our goal is to derive a formula for the probability that the person chooses alternative i *conditional* on the researcher selecting subset K for him. This conditional probability is denoted $P_n(i \mid K)$.

This conditional probability is derived as follows. The joint probability that the researcher selects subset K and the decision-maker chooses alternative i is $Prob(K, i) = q(K \mid i)P_{ni}$. The joint probability can also be expressed with the opposite conditioning as $Prob(K, i) = P_n(i \mid K)Q(K)$ where $Q(K) = \sum_{j\in F} P_{nj}q(K \mid j)$ is the probability of the researcher selecting subset K marginal over all the alternatives that the person could choose. Equating these two expressions and solving for $P_n(i \mid K)$, we have

$$\begin{aligned} P_n(i \mid K) &= \frac{P_{ni}q(K \mid i)}{\sum_{j\in F} P_{nj}q(K \mid j)} \\ &= \frac{e^{V_{ni}}q(K \mid i)}{\sum_{j\in F} e^{V_{nj}}q(K \mid j)} \end{aligned}$$

$$= \frac{e^{V_{ni}} q(K \mid i)}{\sum_{k \in K} e^{V_{nk}} q(K \mid j)} \tag{3.15}$$

where the second line has canceled out the denominators of P_{ni} and P_{nj} $\forall j$, and the third equality uses the fact that $q(K \mid j) = 0$ for any j not in K.

Suppose that the researcher has designed the selection procedure such that $q(K \mid j)$ is the same for all $j \in K$. This property occurs if, for example, the researcher assigns an equal probability of selection to all non-chosen alternatives, such that the probability of selecting j into the subset when i is chosen by the decision-maker is the same as for selecting i into the subset when j is chosen. McFadden (1978) calls this the "uniform conditioning property" since the subset of alternatives has a uniform (equal) probability of being selected conditional on any of its members being chosen by the decision-maker. When this property is satisfied, $q(K \mid j)$ cancels out of the above expression, and the probability becomes

$$P_n(i \mid K) = \frac{e^{V_{ni}}}{\sum_{j \in K} e^{V_{nj}}}$$

which is simply the logit formula for a person who faces the alternatives in subset K.

The conditional likelihood function under the uniform conditioning property is

$$CLL(\beta) = \sum_n \sum_{i \in K_n} y_{ni} ln \frac{e^{V_{ni}}}{\sum_{j \in K_n} e^{V_{nj}}}$$

where K_n is the subset selected for person n. This function is the same as the log-likelihood function given in (3.11) except that the subset of alternatives K_n replaces, for each sampled person, the complete set. Maximization of CLL provides a consistent estimator of β. However, since information is excluded from CLL that LL incorporates (i.e., information on alternatives not in each subset), the estimator based on CLL is not efficient.

Suppose that the researcher designs a selection process that does not exhibit the uniform conditioning property. In this case, the probability $q(K \mid i)$ can be incorporated into the model as a separate variable. The expression in (3.15) can be re-written as

$$P_n(i \mid K) = \frac{e^{V_{ni} + ln(q(K|i))}}{\sum_{j \in K} e^{V_{nj} + ln(q(K|j))}}.$$

A variable z_{nj} calculated as $ln(q(K_n \mid j))$ is added to the representative utility of each alternative. The coefficient of this variable is constrained to 1 in estimation.

The question arises, why would a researcher ever want to design a selection procedure that does not satisfy the uniform conditioning property, since satisfying the property makes estimation so straightforward? An illustration of the potential benefit of non-uniform conditioning is provided by Train, McFadden and Ben-Akiva (1987) in their study of telecommunications demand. The choice situation in their application included an enormous number of alternatives representing portfolios of calls by time of day, distance and duration. The vast majority of alternatives were hardly ever chosen by anyone in the population. If alternatives had been selected with equal probability for each alternative, it was quite likely than the resulting subsets would consist nearly entirely of alternatives that were hardly ever chosen, coupled with the person's chosen alternative. Comparing a person's chosen alternative to a group of highly undesirable alternatives provides little information about the reasons for a person's choice. To avoid this potential problem, alternatives were selected in proportion to the shares for the alternatives in the population (or, to be precise, estimates of the population shares). This procedure increased the chance that relatively desirable alternatives would be in each subset of alternatives that was used in estimation.

3.7.2 Choice-based samples

In some situations, a sample drawn on the basis of exogenous factors would include few people who have chosen particular alternatives. For example, in the choice of water heaters, a random sample of households in most areas would include only a small number who had chosen solar water heating systems. If the researcher is particularly interested in factors that affect the penetration of solar devices, a random sample would need to be very large to assure a reasonable number of households with solar heat.

In situations such as these, the researcher might instead select the sample, or part of the sample, on the basis of the choice being analyzed. For example, the researcher examining water heaters might supplement a random sample of households with households that are known (perhaps through sales records at stores if the researcher has access to these records) to have recently installed solar water heaters.

Samples selected on the basis of decision-makers' choices can be purely choice-based or a hybrid of choice-based and exogenous. In a purely choice-based sample, the population is divided into those that choose each alternative and decision-makers are drawn randomly within each group, though at different rates. For example, a researcher who is examining the choice of home location and is interested in identifying the factors that contribute to people choosing one particular community might draw randomly from within that community at the rate of one out of L households, and draw randomly from all other communities at a rate of one out of M, where M is larger than L. This procedure assures that the researcher has an adequate number of people in the sample from the area of interest. A hybrid sample is like the one drawn by the researcher interested in solar water heating, in which an exogenous sample is supplemented with a sample drawn on the basis of the households' choices.

Estimation of model parameters with samples drawn at least partially on the basis of the decision-maker's choice is fairly complex in general, and varies with the exact form of the sampling procedure. For interested readers, Ben-Akiva and Lerman (1985, pp.234-244) provide a useful discussion. One result is particularly significant, since it allows researchers to estimate logit models on choice-based samples without becoming involved in complex estimation procedures. This result, due to Manski and Lerman (1977), can be stated as follows. If the researcher is using a *purely* choice-based sample and includes an alternative-specific constant in the representative utility for each alternative, then estimating a logit model as if the sample were exogenous produces consistent estimates for all the model parameters except the alternative-specific constants. Furthermore, these constants are biased by a known factor and can therefore be adjusted so that the adjusted constants are consistent. In particular, the expectation of the estimated constant for alternative j, labeled $\hat{\alpha}_j$, is related to the true constant α_j^*,

$$E(\hat{\alpha}_j) = \alpha_j^* - ln(A_j/S_j)$$

where A_j is the share of decision-makers in the population who chose alternative j and S_j is the share in the choice-based sample who chose alternative j. Consequently, if A_j is known (that is, if population shares are known for each alternative), then a consistent estimate of the alternative-specific constant is the constant $\hat{\alpha}_j$ that is estimated on the choice-based sample *plus* the log of the ratio of the population share to the sample share.

3.8 Goodness of Fit and Hypothesis Testing

We discuss goodness of fit and hypothesis testing in the context of logit models, where the log likelihood function is calculated exactly. The concepts apply to other models, with appropriate adjustment for simulation variance, when the log likelihood function is simulated rather than calculated exactly.

3.8.1 Goodness of fit

A statistic called the "likelihood ratio index" is often used with discrete choice models to measure how well the models fit the data. Stated more precisely, the statistic measures how well the model, with its estimated parameters, performs compared with a model in which all the parameters are zero (which is usually equivalent to having no model at all). This comparison is made on the basis of the log likelihood function, evaluated at both the estimated parameters and at zero for all parameters.

The likelihood ratio index is defined as

$$\rho = 1 - \frac{LL(\hat{\beta})}{LL(0)},$$

where $LL(\hat{\beta})$ is the value of the log likelihood function at the estimated parameters and $LL(0)$ is its value when all the parameters are set equal to zero. If the estimated parameters do no better, in terms of the likelihood function, than zero parameters (that is, if the estimated model is no better than no model), then $LL(\hat{\beta}) = LL(0)$ and so $\rho = 0$. This is the lowest value that ρ can take (since if $LL(\hat{\beta})$ is less than $LL(0)$, then $\hat{\beta}$ would not be the maximum likelihood estimate).

At the other extreme, suppose the estimated model were so good that each sampled decision-maker's choice could be predicted perfectly. In this case, the likelihood function at the estimated parameters would be one, since the probability of observing the choices that were actually made is one. And, since the log of one is zero, the log likelihood function would be zero at the estimated parameters. With $LL(\hat{\beta}) = 0$,

$\rho = 1$. This is the highest value that ρ can take. In summary, the likelihood ratio index ranges from zero, when the estimated parameters are no better than zero parameters, to one, when the estimated parameters perfectly predict the choices of the sampled decision-makers.

It is important to note that the likelihood ratio index is not at all similar in its interpretation to the R-squared used in regression, despite both statistics having the same range. R-squared indicates the percent of the variation in the dependent variable that is "explained" by the estimated model. The likelihood ratio has no intuitively interpretable meaning for values between the extremes of zero and one. It is the percent increase in the log likelihood function above the value taken at zero parameters (since $\rho = 1 - (LL(\hat{\beta})/LL(0)) = (LL(0) - LL(\hat{\beta}))/LL(0)$.) However, the meaning of such a percent increase is not clear. In comparing two models estimated on the same data and with the same set of alternatives (such that $LL(0)$ is the same for both models), it is usually valid to say that the model with the higher ρ fits the data better. But this is saying no more than that increasing the value of the log likelihood function is preferable. Two models estimated on samples that are not identical or with a different set of alternatives for any sampled decision maker cannot be compared via their likelihood ratio index values.

Another goodness-of-fit statistic that is sometimes used, but should actually be avoided, is the "percent correctly predicted." This statistic is calculated by identifying for each sampled decision-maker the alternative with the highest probability, based on the estimated model, and determining whether or not this was the alternative that the decision-maker actually chose. The percent of sampled decision-makers for which the highest probability alternative and the chosen alternative are the same is called the percent correctly predicted.

This statistic incorporates a notion that is opposed to the meaning of probabilities and the purpose of specifying choice probabilities. The statistic is based on the idea that the decision-maker is predicted by the researcher to choose the alternative for which the model gives the highest probability. However, as discussed in the derivation of choice probabilities in Chapter 2, the researcher does not have enough information to predict the decision-maker's choice. The researcher has only enough information to state the probability that the decision-maker will choose each alternative. In stating choice probabilities, the researcher is saying that if the choice situation were repeated numerous times (or faced by numerous people with the same attributes), each alternative would be chosen a certain proportion of the time. This is quite different from saying that the alternative with the highest probability will be chosen each time.

An example might be useful. Suppose an estimated model predicts choice probabilities of .75 and .25 in a two-alternative situation. Those probabilities mean that if 100 people faced the representative utilities that gave these probabilities (or one person faced these representative utilities 100 times), the researcher's best prediction of how many people would choose each alternative are 75 and 25. However, the percent correctly predicted statistic is based on the notion that the best prediction for each person is the alternative with the highest probability. This notion would predict that one alternative would be chosen by all 100 people while the other alternative would never be chosen. The procedure misses the point of probabilities, gives obviously inaccurate markets shares, and seems to imply that the researcher has perfect information.

3.8.2 Hypothesis testing

As with regressions, standard t-statistics are used to test hypotheses about individual parameters in discrete choice models, such as whether the parameter is zero. For more complex hypotheses, a likelihood ratio test can nearly always be used, as follows. Consider a null hypothesis H that can be expressed as constraints on the values of the parameters. Two of the most common hypotheses are (1) several parameters are zero, and (2) two or more parameters are equal to each other. The constrained maximum likelihood estimate of the parameters (labeled $\hat{\beta}^H$) is that value of β that gives the highest value of LL without violating the constraints of the null hypothesis H. Define the ratio of likelihoods, $R = L(\hat{\beta}^H)/L(\hat{\beta})$ where $\hat{\beta}^H$ is the (constrained) maximum value of the likelihood function (not logged) under the null hypothesis H and $\hat{\beta}$ is the unconstrained maximum of the likelihood function. As in likelihood ratio tests for models other than those of discrete choice, the test statistic defined as $-2logR$ is distributed chi-squared with degrees of freedom equal to the number of restrictions implied by the null hypothesis. Therefore, the test statistic is $-2(LL(\hat{\beta}^H) - LL(\hat{\beta}))$. Since the log likelihood is always negative, this is simply two times the (magnitude of the) difference between the constrained and unconstrained maximums of the log likelihood function. If this value exceeds the critical value of chi-squared with the appropriate degrees of freedom, then the null hypothesis is rejected.

Examples

Null Hypothesis I: The Coefficients of Several Explanatory Variables are Zero

To test this hypothesis, estimate the model twice: once with these explanatory variables included and a second time without them (since excluding the variables forces their coefficients to be zero). Observe the maximum value of the log likelihood function for each estimation; two times the difference in these maximum values is the value of the test statistic. Compare the test statistic with the critical value of chi-squared with degrees of freedom equal to the number of explanatory variables excluded from the second estimation.

Null Hypothesis II: The Coefficients of the First Two Variables are the Same

To test this hypothesis, estimate the model twice: once with each of the explanatory variables entered separately including the first two; then with the first two variables replaced by one variable that is the sum of the two variables (since summing the variables forces their coefficients to be equal). Observe the maximum value of the log likelihood function for each of the estimations. Multiply the difference in these maximum values by two and compare this figure with the critical value of chi-squared with one degree of freedom.

3.9 Case Study: Forecasting for a new transit system

One of the earliest applications of logit models, and a prominent test of their capabilities, arose in the mid 1970's in the San Francisco Bay Area. A new rail system, called the Bay Area Rapid Transit (BART), had been built. Daniel McFadden obtained a grant from the National Science Foundation to apply logit models to commuters' mode choices in the Bay Area and to use the models to predict BART ridership. I was lucky enough to serve as his research assistant on this project. A sample of commuters was taken before BART was open for service. Mode choice models were estimated on this sample. These estimates provided important information on the factors that enter commuters' decisions, including their value of time savings. The models were then used to forecast the choices that the sampled commuters would make once BART became available. After BART had opened, the commuters were re-contacted and their mode choices were observed. The predicted share taking BART was compared with the observed share. The models predicted quite well, far more accurately than the procedures used by the BART consultants, who had not used discrete choice models.

The project team collected data on 771 commuters before BART was opened. Four modes were considered to be available for the trip to work: (1) driving a car by oneself, (2) taking the bus and walking to the bus stop, (3) taking the bus and driving to the bus stop, and (4) carpool. The time and cost of travel on each mode was determined for each commuter, based on the location of the person's home and work. Travel time was differentiated as walk time (for the bus/walk mode), wait time (for both bus modes), and on-vehicle time (for all the modes.) Characteristics of the commuter were also collected, including income, household size, number of cars and drivers in the household, and whether the commuter was the head of the household. A logit model with linear-in-parameters utility was estimated on these data.

The estimated model is shown in Table 3.1, which is reproduced from Train (1978). The cost of travel was divided by the commuter's wage to reflect the expectation that workers with lower wages are more concerned about cost than higher-paid workers. On-vehicle time enters separately for car and bus travel to indicate that commuters might find time spent on the bus to be more, or less, bothersome than time spent driving in a car. Bus travel often involves transfers, and these transfers can be onerous for travelers. The model therefore includes the number of transfers and the expected wait time at the transfers. The headway (i.e., the time between scheduled buses) for the first bus line that the commuter would take is included as a measure of the maximum amount of time that the person would need to wait for this bus.

The estimated coefficients of cost and the various time components provide information on the value of time. By definition, the value of time is the extra cost that a person would be willing to incur to save time. Utility takes the form: $U_{nj} = \alpha c_{nj}/w_n + \beta t_{nj} + \ldots$ where c is

Table 3.1: Logit Model of Work Trip Mode Choice

Modes: 1. Auto alone, 2. Bus with walk access, 3. Bus with auto access, 4. Carpool

Explanatory variable (Variable enters modes in parentheses and is zero in other modes.)	Coefficient	t-statistic
Cost divided by post-tax wage, in cents per minute (1-4)	-.0284	4.31
Auto on-vehicle time, in minutes (1,3,4)	-.0644	5.65
Transit on-vehicle time, in minutes (2,3)	-.0259	2.94
Walk time, in minutes (2,3)	-.0689	5.28
Transfer wait time, in minutes (2,3)	-.0538	2.30
Number of transfers (2,3)	-.1050	0.78
Headway of first bus, in minutes (2,3)	-.0318	3.18
Family income with ceiling of 7500 (1)	.00000454	0.05
Family income minus 7500 with floor of 0 and ceiling of 3000 (1)	-.0000572	0.43
Family income minus 10,500 with floor of 0 and ceiling of 5000 (1)	-.0000543	0.91
Number of drivers in household (1)	1.02	4.81
Number of drivers in household (3)	.990	3.29
Number of drivers in household (4)	.872	4.25
Dummy if worker is head of household (1)	.627	3.37
Employment density at work location (1)	-.0016	2.27
Home location in or near central business district (1)	-.502	4.18
Autos per driver with a ceiling of one (1)	5.00	9.65
Autos per driver with a ceiling of one (3)	2.33	2.74
Autos per driver with a ceiling of one (4)	2.38	5.28
Auto alone dummy (1)	-5.26	5.93
Bus with auto access dummy (1)	-5.49	5.33
Carpool dummy (1)	-3.84	6.36
Likelihood ratio index		0.4426
Log likelihood at convergence		-595.8
Number of observations		771
Value of time saved as a percent of wage:		t-statistic
Auto on-vehicle time	227	3.20
Transit on-vehicle time	91	2.43
Walk time	243	3.10
Transfer wait time	190	2.01

cost and t is time. The total derivative with respect to changes in time and cost is $dU_{nj} = \alpha/w_n dc_{nj} + \beta dt_{nj}$, which we set to zero and solve for dc/dt to find the change in cost that keeps utility unchanged for a change in time: $dc/dt = -(\beta/\alpha)w_n$. The value of time is therefore β/α proportion of the person's wage. The estimated values of time are reported at the bottom of Table 3.1. Time saved from riding on the bus is valued at 91 percent of wage $((-.0259/-.0284) \cdot 100)$, while time saved from driving in a car is worth more than twice as much: 227 percent of wage. This difference suggests that commuters consider the hassles of driving to be considerably more onerous than riding the bus, when evaluated on a per-minute basis. Commuters apparently choose cars not because they like driving *per se* but because driving is usually quicker. Walking is considered more bothersome than waiting for a bus (243 percent of wage versus 190 percent), and waiting for a bus is more bothersome than riding the bus.

Income enters the representative utility of the auto-alone alternative. It enters in a piece-wise linear fashion to allow for the possibility that additional income has a different impact depending on the overall level of income. None of the income variables enters significantly. Apparently dividing travel cost by wage picks up whatever effect income might have on the mode choice of a commuter. That is: higher wages induce the commuter to be less concerned about travel costs but does not induce a predilection for driving beyond the impact through cost. The number of people and the number of vehicles per driver in the household have a significant impact on mode choice, as expected. Alternative-specific constants are included, with the constant for the bus/walk alternative normalized to zero.

The model in Table 3.1 was used to predict the mode choices of the commuters after BART was open for service. The choice set was considered to be the four modes listed above plus two BART modes, differentiated by whether the person takes the bus or drives to the BART station. Table 3.2 presents the forecasted and actual shares for each mode. BART demand was forecast to be 6.3 percent compared with an actual share of 6.2 percent. This close correspondence is remarkable.

The figures in Table 3.2 tend to mask several complications that arose in the forecasting. For example, walking to the BART station was originally included as a separate mode. The model forecasted this option very poorly, over-predicting the number of people who would

Table 3.2: Predictions for after BART opened

	Actual share	Predicted share
Auto alone	59.90	55.84
Bus with walk access	10.78	12.51
Bus with auto access	1.426	2.411
BART with bus access	0.951	1.053
BART with auto access	5.230	5.286
Carpool	21.71	22.89

walk to BART by a factor of twelve. The problem was investigated and found to be primarily due to differences in the experience of walking to BART stations compared to walking to the bus, given the neighborhoods in which the BART stations are located. These issues are discussed at greater length by McFadden *et al.* (1977).

3.10 Derivation of Logit Probabilities

It was stated without proof in section (3.1) that if the unobserved component of utility is distributed iid extreme value for each alternative, then the choice probabilities take the form of equation (3.6). We now derive this result. From (3.5) we have

$$P_{ni} = \int_{s=-\infty}^{\infty} \left(\prod_{j \neq i} e^{-e^{-(s+V_{ni}-V_{nj})}} \right) e^{-s} e^{-e^{-s}} ds$$

where s is ε_{ni}. Our task is to evaluate this integral. Noting that $V_{ni} - V_{ni} = 0$ and then collecting terms in the exponent of e, we have:

$$\begin{aligned} P_{ni} &= \int_{s=-\infty}^{\infty} \left(\prod_{j} e^{-e^{-(s+V_{ni}-V_{nj})}} \right) e^{-s} ds \\ &= \int_{s=-\infty}^{\infty} exp \left(-\sum_{j} e^{-(s+V_{ni}-V_{nj})} \right) e^{-s} ds \\ &= \int_{s=-\infty}^{\infty} exp \left(-e^{-s} \sum_{j} e^{-(V_{ni}-V_{nj})} \right) e^{-s} ds \end{aligned}$$

Define $t = exp(-s)$ such that $-exp(-s)ds = dt$. Note that as s approaches infinity, t approaches zero, and as s approaches negative infinity, t becomes infinitely large. Using this new term,

$$\begin{aligned}
P_{ni} &= \int_{\infty}^{0} exp\left(-t\sum_j e^{-(V_{ni}-V_{nj})}\right)(-dt) \\
&= \int_{0}^{\infty} exp\left(-t\sum_j e^{-(V_{ni}-V_{nj})}\right) dt \\
&= \frac{exp\left(-t\sum_j e^{-(V_{ni}-V_{nj})}\right)}{-\sum_j e^{-(V_{ni}-V_{nj})}} \Bigg|_0^{\infty} \\
&= \frac{1}{\sum_j e^{-(V_{ni}-V_{nj})}} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}},
\end{aligned}$$

as required.

Chapter 4

GEV

4.1 Introduction

The standard logit model exhibits the independence from irrelevant alternatives (IIA) property, which implies proportional substitution across alternatives. As we discussed in Chapter 3, this property can be seen either as a restriction imposed by the model or as the natural outcome of a well-specified model that captures all sources of correlation over alternatives into representative utility such that only white noise remains. Often the researcher is unable to capture all sources of correlation explicitly, such that the unobserved portions of utility are correlated and IIA does not hold. In these cases, a more general model than standard logit is needed.

Generalized extreme value (GEV) models constitute a large class of models that exhibit a variety of substitution patterns. The unifying attribute of these models is that the unobserved portions of utility for all alternatives are jointly distributed as a generalized extreme value. This distribution allows for correlations over alternatives and, as its name implies, is a generalization of the univariate extreme value distribution that is used for standard logit models. When all correlations are zero, the generalized extreme value distribution becomes the product of independent extreme value distributions and the GEV model becomes standard logit. The class therefore includes logit but also includes a variety of other models. Hypothesis tests on the correlations within a GEV model can be used to examine whether the correlations are zero, which is equivalent to testing whether standard logit provides an accurate representation of the substitution patterns.

The most widely used member of the GEV family is called "nested logit." This model has been applied by many researchers in a variety of situations, including energy, transportation, housing, telecommunications and a host of other fields; see, for example, Ben-Akiva (1972), Train (1986, Ch. 8), Train, McFadden and Ben-Akiva (1987), Forinash and Koppelman (1993), and Lee (1999). Its functional form is relatively simple compared to other types of GEV models, and it provides a rich set of possible substitution patterns. Sections 4.2 and 4.3 describe the specification and estimation of nested logit models. This description is useful in itself, since nested logit models are so prominent, but also as background for understanding more complex GEV models. In section 4.4, we turn to other GEV models that researchers have implemented, with special emphasis on one of the most promising of these, namely, the paired combinatorial logit (PCL) and generalized nested logit (GNL). The chapter's final section describes the entire class of GEV models and how new specifications within the class are generated.

Only a small portion of the possible models within the GEV class have ever been implemented. This means that the full capabilities of this class have not yet been fully exploited and that new research in this area has the potential to find even more powerful models than those already used. An example of this potential is evidenced by Karlstrom (2001), who specified a GEV model of a different form than had ever been used before and found that it fit his data better than previously implemented types of GEV models. GEV models have the advantage that the choice probabilities usually take a closed form such that they can be estimated without resorting to simulation. For this reason alone, GEV models will continue to be the source of new and powerful specifications to meet researchers' needs.

4.2 Nested logit

4.2.1 Substitution patterns

A nested logit model is appropriate when the set of alternatives faced by a decision-maker can be partitioned into subsets, called "nests," in such a way that the following properties hold. (1) For any two alternatives that are in the *same* nest, the ratio of probabilities is independent of the attributes or existence of all other alternatives.

That is, IIA holds within each nest. (2) For any two alternatives in *different* nests, the ratio of probabilities can depend on the attributes of other alternatives in the two nests. IIA does not hold in general for alternatives in different nests.

An example can best explain whether a set of alternatives can be so partitioned. Suppose the set of alternatives available to a worker for his commute to work consists of driving an auto alone, carpooling, taking the bus, and taking rail. If any alternative were removed, the probabilities of the other alternatives would increase (e.g., if the worker's car was being repaired such that he could not drive to work by himself, then the probability of carpool, bus, and rail). The relevant question in partitioning these alternatives is: by what proportion would each probability increase when an alternative is removed? Suppose the changes in probabilities occur as set forth in Table 4.1. Note that the probabilities for bus and rail always rise by the same proportion whenever one of the other alternatives is removed. IIA therefore holds between these two alternatives. Let us put these alternatives in a nest and call the nest "transit." Similarly, the probability of auto alone and carpool rise by the same proportion whenever one of the other alternatives is removed. IIA holds between these two alternatives, and so we put them into a nest called "auto." IIA does not hold between either of the auto alternatives compared with either of the transit alternatives. For example, when the auto alone alternative is removed, the probability of carpool rises proportionately more than the probability of bus or rail. With our two nests, we can state the patterns of substitution succinctly as: IIA holds within each nest but not across nests. A nested logit model with the two auto alternatives in one nest and the two transit alternatives in another nest is appropriate to represent this situation.

Table 4.1: Example of IIA holding within nests of alternatives.

Change in probabilities when one alternative is removed.

	Original	Alternative removed:			
		Auto alone	Carpool	Bus	Rail
Auto alone	.40	-	.45 (+12.5%)	.52 (+30%)	.48 (+20%)
Carpool	.10	.20 (+100%)	-	.13 (+30%)	.12 (+20%)
Bus	.30	.48 (+60%)	.33 (+10%)	-	.40 (+33%)
Rail	.20	.32 (+60%)	.22 (+10%)	.35 (+70%)	-

A convenient way to picture the substitution patterns is with a tree diagram. In such a tree, each branch denotes a subset of alternatives within which IIA holds, and every leaf on each branch denotes an alternative. For example, the tree diagram for the worker's choice of mode described above is given in Figure 4.1. The (upside down) tree consists of two branches, labeled "auto" and "transit," for the two subsets of alternatives, and each of the branches contains two twigs for the two alternatives with in the subset. There is proportional substitution across twigs within a branch but not across branches.

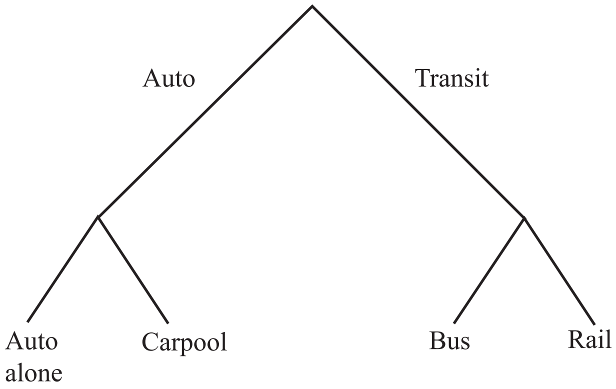

Figure 4.1: Tree diagram for mode choice.

4.2.2 Choice probabilities

Daly and Zachary (1978), McFadden (1978), and Williams (1977) showed, independently and using different proofs, that the nested logit model is consistent with utility maximization. Let the set of alternatives j be partitioned into K non-overlapping subsets denoted $B_1, B_2, \ldots, B_K$ and called nests. The utility that person n obtains from alternative j in nest B_k is denoted, as usual, as $U_{nj} = V_{nj} + \varepsilon_{nj}$, where V_{nj} is observed by the researcher and ε_{nj} is a random variable whole value is not observed by the researcher. The nested logit model is obtained by assuming that the vector of unobserved utility, $\varepsilon_n = \langle \varepsilon_{n1}, \ldots, \varepsilon_{nJ} \rangle$