Discrete Choice Analysis

Moshe Ben-Akiva Michel Bierlaire Daniel McFadden
Joan Walker

January 2018

Do not distribute and do not quote

This is a preliminary and incomplete draft of a textbook on discrete choice analysis, updating the book by Ben-Akiva and Lerman (1985, MIT Press). This document is a working copy. Please send comments and suggestions to michel.bierlaire@epfl.ch

Contents

Part I

Background

Chapter 1

Introduction

"*Everything should be made as simple as possible but not simpler*", Albert Einstein.

Contents

1.1 The Context of Behavioral Analysis and Demand Forecasting

Many aspects of engineering, planning, policy, and business involve a human element, be it consumers, businesses, governments, or other organizations. Effective design and management requires understanding this human response. This textbook focuses on behavioral theories and the use of quantitative methods to analyze human response. A mix of theory and practical

tools are covered, with applications drawn from infrastructure investment and use, urban growth and design, health, sustainability, and marketing.

Take, for example, civil infrastructure systems, such as transportation, water, and energy. These are developed, designed, implemented, and managed with the objective of serving the needs of people and businesses. People and businesses are at the heart of most of the issues in metropolitan studies, in which prime objectives are to better understand and improve the urban environment. The effectiveness of infrastructures and urban environments will depend on how users respond to such systems. Humans are not automatons and it is not possible to force their actions. Therefore, it is essential to anticipate the actions of users in order to effectively design and manage civil infrastructure systems. Consider high-speed rail. Design issues include determining the route and station locations as well as the capacity, frequency, and speed of the trains. Policy decisions include finance, pricing, and zoning. Studying any one of these dimensions requires estimating the demand for the system. To do that, we need to understand the behavior of the users: Under what conditions (price, travel time, frequency, comfort, etc.) will high speed rail be competitive with existing modes? For what segments of travelers (trip purpose, age, levels of income, etc.)?

The methods described in this text aim to answer such questions by employing behavioral research and modeling methods to predict the behavior of users of systems. Predicting behavior is challenging. Humans are heterogeneous with varying preferences, motivations, experiences, and decision-making processes. Indeed, predicting demand has proven to be elusive, with a long history of poor predictions. However, ignoring it is not the answer. It is necessary to develop useful tools for modeling behavior.

Behavioral research is interdisciplinary in nature and includes fields such as economics, psychology, sociology, marketing research, urban planning, and transportation engineering, and planning. In this text we focus on the quantitative methods used to analyze behavior, which necessarily draw both on methods from statistics and theories from behavioral science.

While there are a variety of techniques used to model behavior, the most widely used technique is discrete choice analysis (DCA), the focus of this book. These are statistical tools that are used when the variable being explained is discrete or categorical in nature (unlike in regression where the dependent variable is continuous). They are applicable to model behavioral situations in which a decision-making agent (for example, a person, household, or business) is making a choice from a finite set of discrete alternatives. A classic transportation example is a person deciding what mode (car, transit, walk, bike, etc.) to take to work. Or, in marketing, which particular product an individual buys within some product class. Or in health, what

health insurance one chooses. DCA is used in a wide array of disciplines, so much so that Daniel McFadden was awarded the 2000 Nobel Prize in Economics due to his pioneering work in DCA.

The basic problem confronted by discrete choice analysis is the modeling of choice from a set of mutually exclusive and collectively exhaustive alternatives. Discrete choice problems have been of interest to researchers for a long time in a variety of disciplines. The origins of probabilistic choice models are in mathematical psychology (see Thurstone, 1927, Luce, 1959, Marschak, 1960, Luce and Suppes, 1965, Bock and Jones, 1968, Tversky, 1972). But discrete or qualitative response models had also been used early on in biometric applications (see Berkson, 1944, Finney, 1971, Cox, 1970). The models were later linked to economic consumer theory (McFadden, 1974). The origins of DCA are based on simple functional forms, for example the logit (as well as probit) models that are emphasized in the first part of this text. However, increases in computational power and new developments in flexible functional forms have led to models with greater explanatory power, and such models are highlighted in the later parts of this text. An outstanding history of the field as well as future directions can be found in Daniel McFadden's 2000 Nobel lecture (McFadden, 2001).

We frequently use the principle of utility maximization to motivate the mathematical form of the choice models that are the subject of this text. Briefly, a decision maker is modeled as selecting the alternative with the highest utility among those available at the time a choice is made. An operational model consists of parameterized utility functions in terms of observable independent variables and unknown parameters, and their values are estimated from a sample of observed choices made by decision makers when confronted with a choice situation. It is impossible to specify and estimate a discrete choice model that will always succeed in predicting the chosen alternatives by all individuals. Therefore we adopt the concept of random utility, an idea that first appeared in psychology (Thurstone, 1927). The true utilities of the alternatives are considered random variables, so the probability that an alternative is chosen is defined as the probability that it has the greatest utility among the available alternatives. However, while utility theory is often a useful reference point for choice models, it is by no means a necessary assumption. The mathematical form stands alone as a statistical model. Rationality is also not a necessary assumption. Indeed, much of the advancement in this field is aimed at going beyond the random utility model in its narrowest formulation to incorporate elements of cognitive process that have been identified as important, whether rational or not. The methods described in this text are far from rigid, but rather when combined with the more advanced topics emphasized later in this text, represent

a flexible statistical framework able to capture rich behavioral processes in all their complexities.

The objective of this text is to cover the main concepts and methods of discrete choice analysis and describe the applications of the methodology, starting from the most basic material through to advanced techniques and providing theoretical grounding as well as discussion of practical issues that come about in estimation and application.

For the bulk of this chapter, we use a simple example to convey the key concepts, terminology and process used throughout the text. Following this example, we present an outline of the material in the book.

1.2 Simple Example

In this simple example, the basic components of the choice modeling process are described, including definition of the choice problem, collection of data, model specification, model estimation, and model application. The problem being studied is the market penetration of smartphones relative to other (non-smart or "feature") mobile phones. In today's market, the line between smartphones and feature phones is tricky to define. Here we define smartphones as mobile phones that have both internet access (email and web) and the ability to download and install applications or apps.

1.2.1 Choice problem

For this simple example, we analyze consumer's choice of what kind of mobile telephone to own, and in particular whether to own a smartphone or a feature mobile phone. The questions we are interested in answering are: what is the current market penetration of smartphones in the US relative to feature phones and how will the penetration change in the future. We expect that the likelihood of an individual choosing a smartphone over a feature phone will vary based on the characteristics of the individual. In particular, we hypothesize that level of education is the only explanatory variable (for the sake of keeping the example simple).

1.2.2 Survey

In order to study this problem, it is too costly to perform a census of the population, and such a census is also unnecessary. Rather, we use data collected from a sample of the population to estimate parameters of the full population. For this simple example, we use made up data inspired by a

Pew Internet & American Life Project survey (Brenner, 2013). In 2012, Pew conducted a survey of 2261 households, asking each household a wide range of questions about mobile phone ownership and use. We massaged the actual data to keep our example simple, but the results are consistent with the Pew data. Further, we make the simplifying assumption that our dataset consists of responses from 2000 mobile phone owners who were randomly selected from the population of all mobile phone owners in the US population. Random selection means that every member of the population that is surveyed had an equal probability of being in the sample. For each respondent, we will make use of the answers to two questions:

- Is your mobile phone a smartphone?
  - Yes,
  - No.
- What is your level of educational attainment?
  - No high school diploma,
  - High school graduate,
  - College graduate.

The use of only these two questions is built on our simplifying assumption that education is the only explanatory variable for the choice of whether to own a smartphone or a feature phone. Of course later we will work with more explanatory variables as well as more complicated choices. From the first question, we have a yes or no response from each decision-maker. From the second question, we have a categorical variable of the level of education, which we can characterize by the following labels: Low (no high school diploma), Medium (high school graduate), or High (college graduate) education. The data from this survey can be summarized in a Contingency Table, and our made up data is shown in Table 1.1. All the variables in this survey are discrete and therefore all the available information contained in the survey records can be included in a contingency table. Each of the 2000 respondents falls into one of the six (Smartphone, Education) cells in this contingency table. Therefore, the counts of the respondents in these six cells summarize, without any loss of information, all the data available from our simple survey.

1.2.3 Model Specification

From the survey data, we can estimate the current penetration of smartphones relative to all mobile phones, which is equal to 1085/2000 or 54.3%.

	Education			
Smartphone	Low ($k = 1$)	Medium ($k = 2$)	High ($k = 3$)	
Yes ($i = 1$)	75	500	510	1085
No ($i = 2$)	175	500	240	915
	250	1000	750	2000

Table 1.1: Contingency Table of Survey Responses

However, we want to know more than the current penetration; we want to predict how it will change in the future.

In order to do so, we need to develop a model that explains who chooses a smartphone over a feature phone. In describing the model for our smartphone case, we will introduce terminology and notation.

There are two different types of variables in our model. The first is the dependent or endogenous variable, which is what we are explaining. In the example, the decision either to own a smartphone or a feature phone is our dependent variable. We use the index i to denote the different categories of the dependent variable. In this case $i = 1$ denotes a "yes" response to the smartphone ownership question and $i = 2$ denotes a "no" response[1]. This is a discrete decision, and such discrete decisions are the focus of this book. Continuous dependent variables, such as the amount of hours talking on the smartphone, are modeled using regression analysis and this is not the focus of this book.

The other type of variables are called various equivalent names, including the independent, exogenous, or explanatory variables. In the methods emphasized in this book, these variables can be either discrete or continuous. There can be any number of explanatory variables. In this example we have one discrete explanatory variable of three levels of education. We use the index k to denote these categories, where $k = 1$ denotes low education, $k = 2$ denotes medium education, and $k = 3$ denotes high education. Therefore, the pair (i, k) denotes a particular cell in the contingency table.

In this example, we come across several types of probabilities of interest. First, one of the measures of interest is the frequency of smartphone ownership relative to feature phone ownership in the population (not the sample). This share, or probability, is denoted as $P(i = 1)$. With a random sample, we can use the frequency in the sample to provide an estimate of the frequency in the population. Thus, we can estimate $P(i = 1)$ as follows:

[1]Note that these values are arbitrary, and do not represent any ranking, or any magnitude. Coding "yes" with 2 and "no" with 1 would have been equally fine.

$\widehat{P}(i = 1) = 1085/2000 = 0.543$. We use the sample to make inference on the characteristics of the population and use a hat to indicate an estimate. Note that this is a marginal probability in this context because we have two random variables, smartphone ownership status and level of education.

We also may be interested in joint probabilities, which are the frequency in the population of a particular combination (i, k) of the two random variables. For example, $P(i = 1, k = 2)$ is the frequency in the population of persons who are both owners of smartphones and fall into the medium level of education category. The data from the survey can be used to provide an estimate of this frequency: $\widehat{P}(i = 1, k = 2) = 500/2000 = 0.25$. We can calculate the marginal probability $\widehat{P}(i = 1)$ estimated above by summing over joint probabilities, that is $\widehat{P}(i = 1) = \sum_{k=1}^{3} \widehat{P}(i = 1, k) = 75/2000 + 500/2000 + 510/2000 = 0.543$.

The interest for our study is the penetration of smartphones under conditions other than the present. In order to obtain such information we need to uncover stable behavioral (or causal) relationships between the variables, and for this we need a conditional probability, $P(i|k)$. Such conditional probabilities are the focus of this textbook. For example $P(i = 1|k = 1)$ is the probability that an individual is a smartphone owner given that s/he falls into the low education category. Any joint probability can be expressed as a product of a marginal probability and a conditional probability, in our case $P(i, k) = P(i)P(k|i)$ or $P(i, k) = P(k)P(i|k)$. For example, $P(i = 1, k = 2) = P(i = 1)P(k = 2|i = 1)$ or $P(i = 1, k = 2) = P(k = 2)P(i = 1|k = 2)$. Note in the case of independence when the two random variables are independently distributed, the joint probability equals the product of marginal probabilities, $P(i, k) = P(i)P(k)$, but our fundamental hypothesis in this example is that we do not have independence, that level of education is related to the decision to own a smartphone. We have already defined the joint and marginal probabilities, so from these we can obtain the conditional probabilities by dividing the joint probability by the marginal probability of the conditioning event: $P(k|i) = P(i, k)/P(i)$ (for $P(i) \neq 0$) and $P(i|k) = P(i, k)/P(k)$ (for $P(k) \neq 0$). Of these two conditional probabilities, $P(i|k)$ is the stable behavioral relationship that we have in mind for our study, and so this is the decomposition of interest. $P(k|i)$ also conveys useful information, for example discriminant analysis methods focus on modeling the distribution of characteristics of sub-populations, say, levels of education of smartphone owners, to understand the characteristics of customers (perhaps useful for a service provider). However, this is not the focus of this textbook.

$P(i|k)$ is the behavioral model of interest, and a key assumption of the modeling performed in this book is that $P(i|k)$ has some invariability, or

stability, over time. Therefore, it will serve as the main basis for forecasting. With our smartphone model, we expect $P(i)$ to vary over time, we also expect that levels of education $P(k)$ will vary over time. However, we expect that $P(i|k)$ is stable over time.

Now we are ready to specify our model and estimate its unknown parameters. In our case, the behavioral model of interest is simple, and it is fully specified with the following 3 equations:

$$\begin{aligned} P(i = 1|k = 1) &= \pi_1, \\ P(i = 1|k = 2) &= \pi_2, \\ P(i = 1|k = 3) &= \pi_3, \end{aligned}$$

where π_1, π_2, and π_3 are unknown parameters. Note that $P(i = 2|k) = 1 - P(i = 1|k)$ and additional parameters for $i = 2$ would be redundant.

1.2.4 Model Estimation and Testing

Now that we have collected data and specified the behavioral model, the next step is to estimate the model, that is inferring values for the unknown parameters of the model using the survey data. There are a number of general approaches to develop these estimates. Below, we introduce a technique called Maximum Likelihood Estimation, but for now we can do the estimation more trivially using the entries in the contingency table, as follows:

$$\begin{aligned} \widehat{\pi}_1 &= 75/250 &&= 0.300, \\ \widehat{\pi}_2 &= 500/1000 &&= 0.500, \\ \widehat{\pi}_3 &= 510/750 &&= 0.680. \end{aligned}$$

The first question after any estimation to ask is, do these estimates make sense? Do they match our a priori expectations? These estimates suggest that as years of education increases, there is a higher penetration of smartphones, which does match our expectation. There is increasing use of email and internet use with increasing education and also more disposable income to use to purchase the more expensive devices.

In addition to estimating the parameters, we also want to know something about how good our parameter estimates are. In other words, we would like to know how far are the estimates from the true population values. However, the true values are unknown and therefore the properties of estimated values can only be described in terms of their sampling distribution. What is a sampling distribution? In our example, we sampled 2000 individuals. Suppose that we did the survey again and again; we would get a different set of 2000 people each time and also a different estimate of π_1, π_2, and π_3 each time. From

these multiple samples we would obtain a distribution of estimates for each of our parameters. Such a distribution for $\widehat{\pi}_2$ is shown in Figure 1.1. It is centered on the true value of the parameter that is *assumed* to be 0.48 in the figure (the true value is unknown). The estimate from our single sample (Table 1.1) is also indicated on the figure. The spread (or standard deviation) of the sampling distribution is an indication of how precise our estimate will be. In practice it is obviously infeasible to perform this process and repeat the sampling multiple times. Therefore, the derivation of such sampling distributions is based on theoretical results or simulation.

In our case, the parameter estimates are sample averages of Bernoulli random variables (binary 0/1 variables). From the property of the sample average, we know that this sampling distribution would be centered on the true value (the estimate is unbiased, which is defined below). The sample size affects the spread of this distribution, so the sample size of 2000 has a wider spread than a sample size of 4000, going all the way to a full census of the population, which would have zero spread. The larger the sample size, the better the estimate. Since the sample observations are independent, the standard error of a sample average is equal to $\sqrt{\sigma^2/N}$, where σ^2 is the variance of the variable that is being averaged, and N is the number of observations used to calculate the average.

The variance σ^2 of a Bernoulli variable is $\pi(1-\pi)$, where π is the probability of occurrence, or the probability of the Bernoulli variable being a 1. Therefore, using our estimate $\widehat{\pi}$ for π in the variance equation, the estimated standard errors of our estimates are $\widehat{s}_{\pi_k} = \sqrt{\widehat{\pi}_k(1-\widehat{\pi}_k)/N_k}$. This is an estimate of the standard error of the sampling distribution, and can be calculated for each of our three estimated parameters, $k = 1, 2, 3$, from a single sample alone. The standard errors can be compared relative to the estimate to generate confidence intervals: plus/minus 2 times the standard error is approximately[2] a 95% confidence interval. These standard errors can be used to calculate a t-ratio equal to the estimated parameter divided by its standard error. A large absolute t-value is taken as an indication that the parameter is significantly different from 0. P-values are also often calculated, which is estimated probability of rejecting that the parameter is 0 when it actually is 0. A small p-value is an indication that the parameter is significantly different from 0. Applying a 95% confidence level for testing the significance of parameters (testing whether or not they are different than 0) results in a critical t-ratio value of 1.96 and a critical p-value of 0.05 (reject that the parameter is equal to 0 if the |t-ratio| $>$ 1.96 or, equivalently, the p-value$<$ 0.05). The estimation results for our model are shown in typical

[2]A more precise 95% interval would be obtained with 1.96 times the standard error.

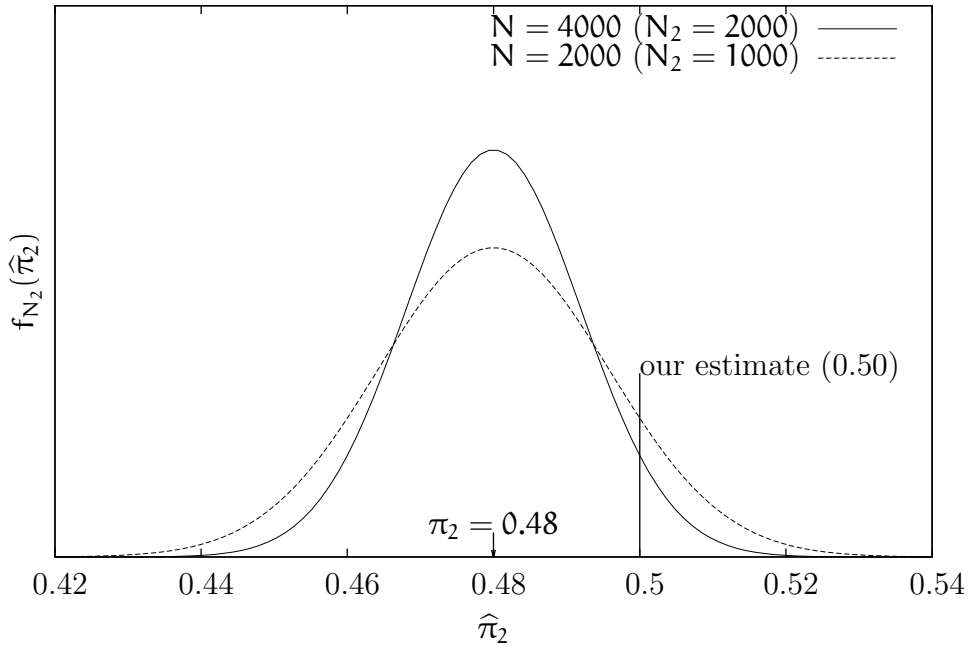


Figure 1.1: Sampling distributions

format in Table 1.2.

There are two particularly important properties of estimators, one is unbiasedness, which means that the sampling distribution of the estimate is centered on the true value of the parameter. The other is efficiency, which means that the estimator achieves the smallest spread possible (standard error) given the sample size. That is, is our estimate the best we could do given the data, or is there a better estimator? These are called small sample properties, because they hold for a finite sample size. Small sample properties are often obtained for linear models such as regression and for the simple averages that we calculate in this example. However, for non-linear models, such as discrete choice, there are no known small sample properties. Therefore, asymptotic properties, which hold in the limit as the sample size goes to infinity, are used. The asymptotic equivalents to unbiasedness and efficiency are consistency (as $N \to \infty$ the sampling distribution collapses on the true value) and asymptotic efficiency (the standard error of the asymptotic sampling distribution is less than or equal to that of any other consistent estimator), respectively.

Maximum likelihood estimation is probably the most general and straightforward procedure for finding estimators. Stated simply, a maximum likeli-

hood estimator is the value for the parameters for which the observed sample is most likely to have occurred. We define $\mathcal{L}^*$ as the likelihood function and the likelihood for our smartphone example is

$$\mathcal{L}^* = \prod_{n=1}^{N} P(i_n|k_n). \tag{1.1}$$

This is simply the product over all decision-makers in the sample ($n = 1, ..., N$ total respondents) of the probability of observing each decision-makers choice i_n, where we condition the probability on the explanatory variable education k_n. The probability function is the behavioral model and this is a function of the parameters that we want to estimate from the data. The parameters of interest are π_1, π_2, π_3 as defined above, and inserting these into the likelihood for our observed sample, we obtain:

$$\mathcal{L}^* = (\pi_1)^{75}(1-\pi_1)^{175}(\pi_2)^{500}(1-\pi_2)^{500}(\pi_3)^{510}(1-\pi_3)^{240}. \tag{1.2}$$

The maximum likelihood estimates of π_1, π_2, π_3 are then those which maximize $\mathcal{L}^*$. It has been proven that maximum likelihood estimates have the desirable properties of consistency and asymptotic efficiency, and therefore they are attractive estimators to use. The small sample properties of the maximum likelihood estimator need to be investigated on a case by case basis. For numerical reasons, we typically maximize the logarithm of $\mathcal{L}^*$ rather than $\mathcal{L}^*$ itself, where we denote $\mathcal{L} = \ln(\mathcal{L}^*)$. This does not change the values of the parameter estimates because the logarithmic function is strictly monotonically increasing. Thus we solve

$$\begin{aligned} \max \mathcal{L}(\widehat{\pi}_i) \quad = \quad & 75\ln(\pi_1) + 175\ln(1-\pi_1) \\ & +500\ln(\pi_2) + 500\ln(1-\pi_2) \\ & +510\ln(\pi_3) + 240\ln(1-\pi_3). \end{aligned} \tag{1.3}$$

$\mathcal{L}$ is continuous and differentiable in π_i and concave, so the parameter estimates are obtained via solving the first-order conditions:

$$\frac{\partial \mathcal{L}}{\partial \widehat{\pi}_i} = 0 \quad \text{for } i = 1, 2, 3. \tag{1.4}$$

Recall from calculus that the derivative is the slope of a function, and the slope equals zero at maxima and minima. For example the first-order condition for π_2 is

$$\frac{\partial \mathcal{L}}{\partial \widehat{\pi}_2} = \frac{500}{\widehat{\pi}_2} - \frac{500}{1-\widehat{\pi}_2} = 0 \quad \Rightarrow \quad \widehat{\pi}_2 = \frac{500}{1000} = 0.500. \tag{1.5}$$

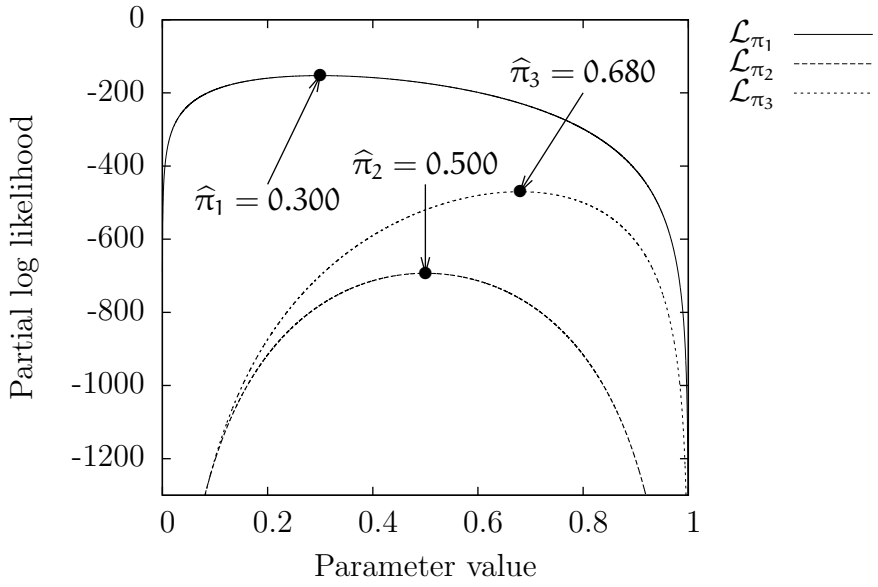


Figure 1.2: Maximizing the log likelihood

This is the identical estimate for π_2 that we obtained above, and the same can be shown for the maximum likelihood estimates of π_1 (estimate is 0.300) and π_3 (estimate is 0.680). The maximization is shown graphically in Figure 1.2. As the log likelihood is separable along the three parameters, the log likelihood associated with each parameter is plotted separately. For example, $\mathcal{L}_{\pi_2} = 500 \ln(\pi_2) + 500 \ln(1 - \pi_2)$. The log likelihoods are plotted as the parameters vary from 0 to 1. You'll note that the log likelihood is negative as the log of a probability (a value between 0 and 1) is always negative and summing up the individual $\ln(p_n)$ over the sample $n = 1, ..., N$ leads to a large negative number. The log likelihood for π_2 reaches a maximum at -693 when $\widehat{\pi}_2 = 0.500$ and this is when the slope of the log likelihood is zero (the first order condition). From the graph it is apparent that the solution is a maximum and not a minimum, although more generally the second-order conditions would be used to verify this if the likelihood function happened to be non concave.

In this book we often use statistical tests to make inferences about various parameters. Above the presentation of t-ratios and p-values were examples of statistical tests of whether a parameter is significantly different than 0. In this smartphone example, we may want to test whether the three parameters are significantly different than one another, that is whether or not education has an effect on smartphone ownership. We assume readers have a basic

understanding of hypothesis testing and introduce here a test statistic you will become very familiar with by the end of this text: the likelihood ratio test.

In the language of hypothesis testing, our null hypothesis is that $\pi_1 = \pi_2 = \pi_3$ and our alternative hypothesis is the opposite, that is at least one of the parameters does not equal the others. The null hypothesis is a restricted version of our model of interest, where the restriction is that the three parameters are constrained to equal each other. The alternative hypothesis is called the unrestricted model, in which the parameters are allowed to be different. Employing maximum likelihood estimation, there is a fairly simple test we can apply to test our null hypothesis. We estimate two different models: a restricted model and an unrestricted model. Conceptually, if imposing the restriction does *not* lead to a large loss of fit (as measured by a decrease in log likelihood), then we do not reject the null hypothesis.

The unrestricted model is what we estimated above, where we found $\widehat{\pi}_1 = 0.300$, $\widehat{\pi}_2 = 0.500$, and $\widehat{\pi}_3 = 0.680$ and the log likelihood for this model is -1316.0, which you can verify by substituting the estimated values of the parameters into equation (1.3). Now we need to estimate the restricted model.

The restricted model has only a single parameter π. The likelihood of the restricted model is $\mathcal{L}^* = (\pi)^{1085}(1-\pi)^{915}$; the log likelihood is $\mathcal{L} = 1085\ln(\pi) + 915\ln(1-\pi)$. Solving the first order condition (now only one equation since there is only one parameter), leads to $\widehat{\pi} = 0.543$.[3] We obtain the maximum value of the restricted log likelihood function by plugging this estimate back into the log likelihood function: $\mathcal{L} = 1085\ln(0.543)+915\ln(1-0.543) = -1379.0$. So, clearly there has been a loss of fit from -1316.0 (denoted as $\mathcal{L}^U$, the log likelihood of the unrestricted model) to -1379.0 (denoted as $\mathcal{L}^R$, the log likelihood of the restricted model). The loss of fit is expected as we restrict parameters, but the question is whether this loss is statistically significant. For this we need a test statistic.

It can be shown (See Theil, 1971, p. 396 for a derivation) that, under the null hypothesis, the test statistic $-2(\mathcal{L}^R - \mathcal{L}^U)$ is asymptotically distributed as χ^2 with degrees of freedom equal to the number of restrictions (in our case, 2). If this statistic is "large" in the statistical sense, we reject the null hypothesis that the restrictions are true. In our case, the value of the statistic is $-2(-1379.0 + 1316.0) = 126.1$. We need to use the χ^2 distribution to determine whether this is large enough to reject the null hypothesis. Using a level of significance of 1% (that is, the probability of rejecting the null

[3]Note that it is the same value as the market share of the smartphone. Can you explain why?

parameter	estimate	standard error	t-ratio	p-value
π_1	0.300	0.0290	10.3	< 0.001
π_2	0.500	0.0158	31.6	< 0.001
π_3	0.680	0.0170	39.9	< 0.001

Table 1.2: Estimation Results for Smartphone Model

hypothesis when it is true), the critical value of the χ^2 distribution with 2 degrees of freedom is 9.210 (see Table D.1). As our test statistic is well above this critical value, we reject the null hypothesis with at least 99% confidence and conclude that education *does* influence smartphone ownership.

1.2.5 Application

Now that we have an estimated model and tested the specification, what do we do with it? Sometimes we just want to perform hypothesis tests such as whether education influences smartphone ownership. In many cases we want to test a future scenario. The survey indicates that at the present time the smartphone penetration within the mobile phone market is 54.3%. What will happen to smartphone penetration if the distribution of education changes, currently there are 13% with low education, 50% with medium education, and 38% with high education. Suppose that in either another population or a future scenario that the distribution is 10% low, 40% medium, and 50% high. We can use our model to predict smartphone penetration given this distribution. It is equal to $0.300 \times 0.10 + 0.500 \times 0.40 + 0.680 \times 0.50 = 0.570$ or 57.0%. This forecast was calculated using the following equation:

$$P(i) = \sum_{k=1}^{3} P(i|k)P(k) \tag{1.6}$$

or, substituting our estimated parameters:

$$\widehat{P}(i) = \sum_{k=1}^{3} \widehat{\pi}_k P(k). \tag{1.7}$$

The marginal probability of education, $P(k)$ is an input to the model, the $\widehat{\pi}_k$ are the parameters from our model estimation, and $\widehat{P}(i)$ is the output from the model. Our model says that as education increases from a 13/50/38 to a 10/40/50 split, smartphone penetration increases from 54.3% to 57.0%.

Further, we may want to know something about the confidence of these estimates, both in terms of the base year penetration rate as well as the

forecast. The confidence interval is generated based on the standard error of (1.7), which in turn is a function of the standard errors of the parameter estimates (denoted as $\widehat{s}_{\pi_k}$) as reported in Table 1.2. The three $\widehat{s}_{\pi_k}$ are independent (they are estimated separately on different subpopulations) and so the estimated standard error of the estimated penetration rate $\widehat{P}(i)$ is equal to

$$\widehat{s}_{P(i)} = \sqrt{\text{Var}(\widehat{P}(i))} = \sqrt{\sum_{k=1}^{3} P(k)^2 \widehat{s}_{\pi_k}^2}. \tag{1.8}$$

A 95% confidence interval of the estimated penetration rate is the range within which we are 95% certain the true value of the population parameter lay. This range is calculated as the estimated penetration rate plus or minus 1.96 standard errors as calculated from (1.8). For the base case penetration estimate of 54.3%, the 95% confidence interval is (52.1%, 56.4%). For the forecast penetration estimate of 57.0%, the 95% confidence interval is (54.8%, 59.2%). Confidence intervals provide useful intuition on the precision of the estimates that are generated from the model.

1.2.6 Discrete choice model formulation

Our simple example involved one choice, own a smartphone or a feature phone, and one explanatory variable, the level of education. The resulting formulation of the choice model $P(i|k)$ was trivial, captured by the three parameters π_1, π_2, and π_3. In almost all applications, the problem will be more complex, often involving more choices and almost always involving more explanatory variables. Therefore, such a trivial specification will not be possible. Here we briefly introduce a common choice model, called logit. This model and its variants are covered in excruciating detail in the rest of the book. For now, we present the model mechanically so that you get a sense of the model formulation and how it relates to this simple example.

Let's say we can represent the attractiveness of each alternative by a function V_{in}, which will vary across alternatives $i = 1, ..., J$ (J is the number of alternatives) and decision-makers n. We will later motivate this "attractiveness" function from microeconomics and refer to it as the (systematic) utility. A natural form of the probability equation would then be

$$P_n(i) = \frac{V_{in}}{\sum_{j=1}^{J} V_{jn}} \text{ for all } i = 1, ..., J \tag{1.9}$$

Let's examine this equation to see if it is sufficient. As the attractiveness of the alternative increases, the probability of choosing that alternative in-

creases and vice versa, which is as expected. We also expect that the sum of probabilities over all of the choices $i = 1, ..., J$ for any individual n would equal 1, which is also the case (even for negative V_{in}). However, in order for the $P_n(i)$ in equation (1.9) to represent a proper probability (i.e., be non-negative and <= 1), we need to impose that $V_{in} \geq 0$ and at least one $V_{in} > 0$ (to not divide by 0). This is *not* reasonable as V_{in} can be negative (as we'll show later).

Fortunately, there is a simple fix. To allow for negative V_{in}, we can transform the equation as follows:

$$P_n(i) = \frac{e^{V_{in}}}{\sum_{j=1}^{J} e^{V_{jn}}} \text{ for all } i = 1, ..., J, \tag{1.10}$$

and this equation exhibits proper probabilities. This is the logit equation, a common discrete choice model. A more formal derivation and discussion is presented in Chapter 4 for a two alternative case and Chapter 5 for a multiple alternative case.

The next issue is what is the formulation of V_{in}. The attractiveness of any alternative will be a function of the explanatory variables x_{in}, which also may vary across i and n. These variables may include both attributes of the alternatives and characteristics of the decision-maker. It will also be a function of unknown parameters β, which will be estimated from data.

So let's look at the formulation for the smartphone example. In this case, $J = 2$ and there are two utilities: V_{1n} for the attractiveness of smartphone and V_{2n} for the attractiveness a feature mobile phone. The probability of person n choosing to own a smartphone is then

$$P_n(1) = \frac{e^{V_{1n}}}{e^{V_{1n}} + e^{V_{2n}}}. \tag{1.11}$$

Dividing the numerator and denominator by $e^{V_{1n}}$ leads to

$$P_n(1) = \frac{1}{1 + e^{-(V_{1n} - V_{2n})}} \tag{1.12}$$

Note that in this equation only the differences in attractiveness matters, and so we only need to write an equation for this difference $V_{1n} - V_{2n}$. V_{in} is a function of explanatory variables and unknown parameters. The only explanatory variable in the simple example is education, which has three categories of low, medium and high. If we assume a linear in parameter form, the difference in attractiveness is written as follows

$$V_{1n} - V_{2n} = \beta_1 x_{(k=1)n} + \beta_2 x_{(k=2)n} + \beta_3 x_{(k=3)n}, \tag{1.13}$$

where $x_{(k=1)n}$ is equal to 1 if person n has low education and 0 otherwise, and $x_{(k=2)n}$ and $x_{(k=3)n}$ are defined similarly for medium and high education. β_1, β_2, and β_3 are parameters to be estimated from data. Using the data we presented above in and, for example, the maximum likelihood technique described above, the estimated parameters are as follows

$$\begin{aligned}\widehat{\beta}_1 &= -0.850,\\ \widehat{\beta}_2 &= 0.000,\\ \widehat{\beta}_3 &= 0.750.\end{aligned}$$

One can verify that plugging these estimated values back into Equations 1.12 and 1.13 and will result in the same probabilities estimated above of $P(1|k=1) = 0.300$, $P(1|k=2) = 0.500$, and $P(1|k=3) = 0.680$. The parameters β capture how ones education level impacts the attractiveness of owning a smartphone (relative to a feature mobile phone). The parameters increase with education, indicating that attractiveness owning a smartphone (relative to a feature mobile phone) increases with increasing education. We could produce an estimation results table similar to Table 1.2 and perform statistical tests and do forecasting, but now with the model as formulated in Equations 1.12 and 1.13 and with the parameters β_1, β_2, and β_3.

1.3 Summary and Outline of the Book

At this point we have used the simple smartphone example to cover the basic concepts of the material in the book: modeling context, data, specification of a behavioral model, estimation, statistical inference, and model application. The remainder of the book will address how to perform such analysis when there is more complexity, including more complicated choices and more explanatory variables. For example, in this analysis we may want to consider different types of mobile phones (smart and other), other attributes of these alternatives such as price and features, and to include more characteristics of the decision maker. We will show how the logit model presented above is extended for more alternatives (e.g., where there is one utility for each mobile phone on the market) and more explanatory variables (e.g., including other characteristics of the decision maker, such as income and age, as well as attributes of the alternatives, such as mobile phone features and price). Furthermore, the techniques touched on here will be derived rigorously using a behavioral theory, maximum likelihood estimation, and hypothesis testing. While the material will get more complicated as we spend time on the formulation of a choice model, keep in mind that the underlying fundamental objectives and processes are as simple as laid out in this chapter.

The following chapters of this book develop the basic concepts and methods of discrete choice analysis and describe the applications of the methodology. The coverage includes choice theories, alternative model formulations, properties of choice models, estimation methods and their properties, tests used in the process of model development, and procedures to use these models in applications for forecasting. There are three sections in this text. This first section covers background material, beginning with key concepts covered by the simple smartphone example presented in this chapter. Chapter 2 focuses on issues pertaining to the data needed for discrete choice analysis, including necessary components of the dataset, examples of datasets, types of data including revealed and stated preferences, data design and collection, and sampling issues. Chapter 3 presents the various conceptual approaches of choice theories that are the foundations of the operational probabilistic choice models developed in the subsequent chapters.

After this background section, Section II presents the basic methods, including the derivations and estimation methods of binomial (Chapter 4) and choices with multiple alternatives (Chapter 5) of different forms with an emphasis on the logistic and the logit models. These are historically the most widely applied discrete choice models and still the workhorses of modern discrete choice analysis. The properties of the models and of the procedures for estimating their parameters are described in detail. Chapter 6 focuses on specification testing, including the informal and formal tests that have proved to be most useful during the process of estimating a discrete choice model. Chapter 7 is concerned with the properties and the applications of the nested logit model, a generalization of logit, and Chapter 8 the multivariate extreme value model, a further generalization of logit. Chapter 10 addresses the issue of prediction and aggregate forecasting with disaggregate models. In principle, this problem is straightforward, but in practice, it requires an application of numerical procedures some of which require different assumptions in order to be of use.

The third section covers more advanced methodological topics that are important for behavioral modeling. These include issues of implementation such as implications of sampling on the estimation of discrete choice models (Chapter 11) and large choice sets (Chapter 12), advanced models such as mixture models (Chapter 13) and combining data (Chapter 15), and estimation techniques such as simulation-based estimation (Chapter 14).

Chapter 2

Data for choice modeling

> "*It is our choices that show what we truly are, far more than our abilities*", Albus Dumbledore.

The objective of this chapter is to overview data necessary for choice modeling and to cover key concepts in data collection, including topics of experimental design and sampling. In Chapter 1, we describe the problem of discrete choice analysis as the modeling of choice from a set of mutually exclusive and collectively exhaustive alternatives. To develop the necessary dataset, we of course first need to define the choice problem of interest, including the decision maker, the choice being made, and the alternatives from which the choice is selected. In this book, we focus on models developed from observations of individual decision makers (as opposed to aggregate market share data). In particular, we collect data from a sample of decision-makers, and then use the sample to estimate parameters of the full population.

In this chapter, we describe the components of such a sample, present sample datasets, and introduce the two primary types of data: revealed preferences and stated preferences. We also cover two technical topics: experimental design for stated preference surveys and sampling theory.

2.1 Introduction to discrete choice data

First we provide a general overview of discrete choice data, starting with the components of a discrete choice dataset, then providing examples of datasets, and discussing data types such as stated and revealed preferences.

2.1.1 Data components

In order to infer parameters of a behavioral model, we must have data on individual preferences and the choice environment from which these preferences were made. This choice environment includes the alternatives from which the choice is selected as well as the attributes of the alternatives. As often we use characteristics of the decision maker to explain the behavior, we also need data on individual characteristics. Therefore, the data necessary for choice modeling consists of the following elements, which were introduced in Chapter 1:

1. a sample of decision makers,
2. characteristics of each decision maker,
3. information on their preferences (for example, their chosen alternative),
4. alternatives available to each decision maker, and
5. attributes of alternatives faced by each decision maker.

In our simple example, the choice environment had two alternatives: either *own a smartphone* or *own a non-smart mobile phone*. We used no other attributes of the alternatives other than their names. The decision maker was the individual respondent, and we used education as a characteristic to explain their preferences. The preferences were gathered via the survey question "Is your mobile phone a smartphone?", and the education through the survey question "What is your level of educational attainment?". A portion of this individual survey smarphone dataset, which would be used to develop the discrete choice models of this book, is shown in Table 2.1. The critical aspects to a dataset for discrete choice analysis are (1) that it contains actual preferences for a sample of individuals and (2) the attributes are known for all of the alternatives faced by the individual (that is, it is *not* enough to have attributes for only the chosen alternative). In the smartphone case, the only attribute used was the so-called name of the alternatives. However, attributes that may make sense to use in this case would be price structure and some measure of quality. We also assumed that smartphones are available to everyone in the sample, but this is not necessary (as long as it is available for some respondents). Indeed, predicting a selection from a choice set with a single alternative would make our job trivial.

Preference data as in our smartphone example are known as revealed preferences, because they are actual behavior from the market place. There are also data known as stated preferences, which are observed or expressed

preferences in response to hypothetical scenarios or experiments. Note that while the focus of this book is on modeling preferences, the methods described here can also be used for non-preferences data; an example of such an application is included in the next section.

2.1.2 Dataset examples

Appendix E provides detail on a host of datasets that we make use of throughout this text. To provide an idea of the variety and content of discrete choice datasets, a subset of these datasets is briefly summarized in Table 2.2 and Table 2.3. The first key point is that the discrete choice models of this book are applied in a wide array of fields. A handful of domains are shown in the table, and the areas of application are, indeed, much broader. Try searching key words such as "discrete choice analysis" or "logit" (the most commonly used discrete choice model) and your field of interest. Nutrition? yes. Elections? yes. Marriage? yes. Terrorism? yes. And so on.

The richness of the datasets can vary tremendously. The observed preference can be a simple binary choice as in the yes/no decision of the smartphone example. It can be a choice from multiple alternatives such as the Swissmetro transport mode choice of auto, traditional rail, or high speed rail. It can also consist of a large number of alternatives, such as making a decision of what car to purchase from the hundreds of available models. And preferences can come in forms other than a single choice from a set of alternatives. For example, in the Boeing dataset, the respondents ranked the three alternatives from most preferred to least preferred. As we will see later, this is one of the advantages of stated preferences data, which is that more information can be obtained regarding the preferences than just a single choice.

The agent making the decision also varies from individuals to households to firms. Some of our datasets are revealed preferences (choices made in a real market situation) and some are stated preferences (responses to hypothetical choice environments). The facial expression dataset is an example of discrete choice analysis being applied to *non*-preference data and therefore there is no decision maker, per se. The objective of the facial expression model is to classify the facial expression of a person in a photograph, such as happy, sad, angry, and so on. This is a classification problem, which is another typical application of discrete choice analysis.

Table 2.3 describes the choice set (that is, set of alternatives) in each dataset and also provides examples of the attributes of these alternatives, which must be known for all alternatives in the choice set. Finally, the table also provides examples of the available characteristics of the decision makers.

Further information on all of these datasets is available in the Appendix.

Respondent	Choice	Education
1	Smartphone	Medium
2	Smartphone	High
3	Non-smartphone	Medium
.	.	.
.	.	.
.	.	.
1999	Non-smartphone	Low
2000	Smartphone	Medium

Table 2.1: Smartphone dataset for choice modeling

2.1.3 Data Types

We have already introduced above the concepts of revealed preferences and stated preferences. Revealed preferences can be collected by either observing actual behavior (for example, consumer data collected electronically at the point of purchase) or by gathering reports from decision-makers on choices they have made in the past (for example, collecting a diary of travel made throughout a day). This is behavior that is happening or has already happened in a real marketplace. There are numerous ways to collect stated preference behavior, and the key here is that these are observed or expressed preferences in response to hypothetical scenarios (or *experiments*). As such, these are statements of *intention*. At first glance, it may seem that one should always collect revealed preferences. In a perfect world, this may be the case. However, revealed preferences may be too expensive and/or do not have the information that is required for the application of interest. Indeed, there are advantages and disadvantages to each type of data. Table 2.4 provides a more detailed comparison, citing advantages and disadvantages of each among a number of dimensions: preference, alternatives, attributes, choice set, number of responses, and preference elicitation. We spend a good part of the rest of this chapter focusing on issues of stated preference data collection, which will expound on this list.

Note that given the relative advantages and disadvantages of each type of data, techniques to combine stated and revealed preferences (covered in Chapter 15) are becoming increasingly popular.

Name of dataset	Domain	Decision maker	Decision	Datatype
Smartphone	Telecom	Individual	Whether mobile phone is smart or not	Revealed
Residential telephone services	Telecom	Household	Choice of residential calling plan	Revealed
Quebec energy	Energy	Household	Choice of residential heating plan	Revealed
Swissmetro	Transport (Ground)	Individual	What ground mode to travel between cities	Stated
Boeing	Transport (Air)	Individual	Choice of flight for a trip	Stated
Choice-lab fashion	Fashion	Firm	Whether to remain as a client of a B2B marketing firm	Revealed
Facial expression	Psychology	None	Classification of facial expression	Classification

Table 2.2: Example datasets

Name of dataset	Alternatives (Number: Description)	Example of Characteristics of decision maker	Examples of Attributes of Alternatives
Smartphone	2: Own smartphone, Own non-smart mobile phone	Education	Name of alternative
Residential telephone services	5: Different calling plans for household telephone service	Household income, number of household members	Monthly cost of plan and expected usage cost
Quebec energy	9: Different configurations (combinations of gas, electric, oil, wood)	Household income, Age of household head, Age of house, Size of house	Annual operating cost, Annual fixed cost
Swissmetro	3: Car, Conventional Train, High-speed Train	Traveler income, gender, and travel purpose	Travel time, travel cost
Boeing	3: Hypothetical flight itineraries between a given origin and destination	Income, Education, Gender, Desired departure time	Airfare, Travel time, Number of transfers, Airline
Choice-lab fashion	2: Stay as a client, Leave as a client	Time as customer, Number of employees, Credit rating	Name of alternative
Facial expression	7: Happiness, Surprise, Fear, Disgust, Sadness, Anger, Neutral	Facial measures such as width and height of eyes and mouth	Name of alternative

Table 2.3: Example Datasets (continued)

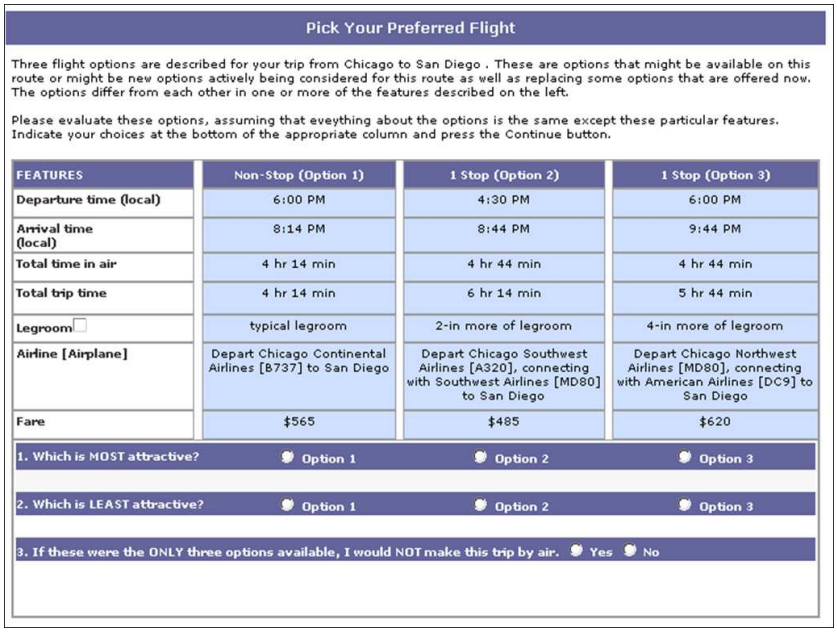
	Revealed Preferences	Stated Preferences
Preference	Choice behavior in actual market.	Preference statement for hypothetical scenarios.
	Cognitively congruent with actual behavior.	May be cognitively incongruent with actual behavior.
	Market and personal constraints are accounted for.	Market and personal constraints may not be considered.
Alternatives	Actual alternatives.	Alternatives generated by analyst.
	Responses to non-existing alternatives are unobservable.	Can elicit preference for new (non-existing) alternatives.
Choice Set	Ambiguous in many cases.	Pre-specified by analyst.
Attributes	May include measurement errors.	No measurement errors.
	Correlated attributes.	Multicollinearity can be controlled by experimental design.
	Ranges are limited.	Ranges can be extended by analyst.
Number of responses	Often difficult to obtain multiple responses from an individual.	Repetitive questioning is easily implemented.
Preference elicitation	Only choice is available.	Various preference formats are available (for example, ranking, rating, matching).

Table 2.4: Comparison of Revealed Preference and Stated Preference Data

2.2 Stated Preferences

Due to a variety of limitations in revealed preference data, market researchers have long used stated preferences data to provide insight on preferences. Stated preference analysis originates from mathematical psychology (Thurstone, 1931 and Luce and Tukey, 1964) and was further developed for demand analysis by marketers, economists, and engineers (Johnson, 1974, Green et al., 1981, Louviere and Woodworth, 1983, McFadden, 1986, and Morikawa et al., 2002). The idea is to obtain a rich form of data on behavior by studying preferences under hypothetical scenarios designed by the researcher. This is distinctly different from revealed preferences, which are observed or reported actual behavior (for example, through travel diary, consumer purchases, etc.). The term *conjoint* analysis is often used synonymously with stated preference analysis. This terminology came about to emphasize that one has to measure the effects of a set of attributes together in order to capture the relative tradeoffs that a consumer makes (e.g. the *combination* of time and cost for an alternative) rather than examining each attribute separately and then summing up the individual effects as was done previously in market research. There is a large literature on stated preference analysis. We review the basics here and point the reader to Green et al. (2001), Moscati (2007), and McFadden (2013) for historical perspectives, Carroll and Green (1995) and Louviere et al. (2000) for discussion of the methods, and Carson and Louviere (2011) for discussion regarding nomenclature nuances.

An example of a stated preferences experiment is shown in Figure 2.1. This is from a survey conducted by Boeing Commercial Airplanes in 2004 and 2005. The survey was designed by Boeing staff with the assistance of Jordan Louviere of the University of Technology, Sydney. Boeing was interested in understanding the sensitivity that air passengers have toward the attributes of an airline itinerary. The survey was conducted by intercepting customers of an internet airline booking service that searches for low-cost travel deals. While waiting for the search engine to return the real itineraries, customers were asked to complete a survey tailored to their origin and destination. A typical page of the survey instrument is shown in Figure 2.1, in this case for a customer that desired to travel from Chicago to San Diego. Each respondent is offered three hypothetical flights, each described by departure time, travel time, legroom, airline, airplane, and fare. The respondent was asked to rank the available choices as well as given the option to decline all of the stated options. In addition to these preferences data, demographic data collected included age, gender, income, occupation, and education. Further, situational variables that were collected included the desired departure time, trip purpose, who is paying for the trip, and the number in the travel party.



Pick Your Preferred Flight

Three flight options are described for your trip from Chicago to San Diego . These are options that might be available on this route or might be new options actively being considered for this route as well as replacing some options that are offered now. The options differ from each other in one or more of the features described on the left.

Please evaluate these options, assuming that eveything about the options is the same except these particular features. Indicate your choices at the bottom of the appropriate column and press the Continue button.

FEATURES	Non-Stop (Option 1)	1 Stop (Option 2)	1 Stop (Option 3)
Departure time (local)	6:00 PM	4:30 PM	6:00 PM
Arrival time (local)	8:14 PM	8:44 PM	9:44 PM
Total time in air	4 hr 14 min	4 hr 44 min	4 hr 44 min
Total trip time	4 hr 14 min	6 hr 14 min	5 hr 44 min
Legroom	typical legroom	2-in more of legroom	4-in more of legroom
Airline [Airplane]	Depart Chicago Continental Airlines [B737] to San Diego	Depart Chicago Southwest Airlines [A320], connecting with Southwest Airlines [MD80] to San Diego	Depart Chicago Northwest Airlines [MD80], connecting with American Airlines [DC9] to San Diego
Fare	$565	$485	$620

1. Which is MOST attractive? Option 1 Option 2 Option 3

2. Which is LEAST attractive? Option 1 Option 2 Option 3

3. If these were the ONLY three options available, I would NOT make this trip by air. Yes No

Figure 2.1: Example Stated Preferences Survey

Boeing chose to use stated preferences due to difficulties in conducting a comparable revealed preferences study. Revealed preferences are harder to collect in this instance because of the proprietary nature of such data: airlines and booking services do not easily (or cheaply) part with such data. In addition, the choice set involved in real airline itinerary choice is extremely complicated in that the true set of available itineraries is huge, it is specific to the time at which the flight is purchased (availability and price vary), and is sensitive to algorithm used by the seller to present a selection of itineraries (for example which is chosen to display at the top of the screen).

This Boeing survey is just one example of a stated preference survey instrument. There are many ways to collect stated preference data. The remainder of this chapter will cover issues of designing and using stated preference data.

2.2.1 Motivation for Using Stated Preferences Data

Revealed preferences are in many ways more attractive than stated preferences, simply because they represent real choice in the market. However, the use of stated preferences is necessary in some cases and useful in others. Advantages of stated preferences come from the disadvantages of revealed

preference data, which are that RP data does not always have the information that you need or it may be too expensive to collect. More specifically, the list below expands on reasons one may want to use stated preference data.

Identification: There are cases where certain effects cannot be identified from revealed preference data. For example, it is not possible in revealed preferences data to capture responses to products not yet on the market (such as civilian space travel or new forms of bike share programs) or for attributes that are not yet in existing products (such as a brand new feature for a cellular phone). Sometimes the attribute may exist in the market, but one is interested in attribute levels that are beyond what exist in the market (such as computer speeds or video resolution), and this is also an issue of identification.

Efficiency: There are other instances where the attributes may exist in the attribute levels of interest, but the effect is difficult to identify using revealed preference data either because of limited variability of the attribute (such as transit fare) or collinearity of attributes (such as price and quality).

Choice Set Definition: In the real market, the choice set that any decision maker faces as well as the attributes of these choices are often not clearly specified. It is difficult to impute the choice set and attributes; people are not good at articulating their choice set and attributes (particularly of non-chosen alternatives) and, even if they were, it would significantly increase the length of the questionnaire thereby increasing costs and decreasing response rates.

Data Collection Resources: In some cases, collection of revealed preference data is simply too expensive and/or time consuming to collect and process. It is often cheaper to design a hypothetical experiment that can be administered in a computer lab or over the web. Going into the field to collect revealed preference data can be costly both in terms of recruitment (either intercepting people at point of decision or asking about choices retrospectively) and collecting the necessary data about the choice environment (including alternatives available and considered, attributes for all available alternatives, and other information about the choice context).

Stated preferences can be used to address all of these issues simply because the analyst is directly in charge of the entire choice experiment: the choice set, the alternatives, the attributes and their values. Further, while in revealed preferences one can often only obtain a single choice, with stated preferences it is easy to obtain numerous responses per respondent and employ various response formats that are more informative than a single choice (for example, ranking, rating, or matching as described later). The primary drawback to stated preference data is that they are not real decisions in the

marketplace and may not be congruent with actual behavior.

2.2.2 Elicitation of Stated Preferences

In developing a stated preference survey, one needs to define the experimental setting, which consists of the context of the hypothetical scenario as well as the alternatives or (in the language of experimental design) *profiles* that make up this scenario. These profiles are defined as bundles of attributes. The respondent is presented with limited sets of alternatives to evaluate. From this evaluation, we obtain an expression of the respondent's preferences. There are many different elicitation methods that can be used to obtain information on preferences. In a discrete choice experiment, a respondent would be asked to make a choice between two or more discrete alternatives. However "choosing" is just one such elicitation method. One of the advantages of stated preference data is that preferences can be expressed in a number of forms, including:

Choice: This is analogous to what is collected with revealed preferences data, where the respondent is asked to state a preference of *one* alternative relative to each of the others. It can be a choice from among two or more alternatives.

Ranking: With ranking data, respondents state the preference of *each* alternative relative to each of the others. That is, they provide complete rankings from most to least preferred. The Boeing example above is an example of ranking data, because information is obtained on both the most attractive and least attractive of the three alternatives.

Best-worst: Asking about best and worse choices can be used for more than three alternatives even though it does not provide a full ranking. The rationale is that after choosing one's favorite alternative, the next easiest choice to make is one's least favorite. Therefore, this is not particularly burdensome for the respondent (certainly less so than a full ranking) and it provides greater information on preferences.

Pick any: The pick any approach asks respondents to divide a set of alternatives into those that are acceptable and those that are unacceptable.

Rating: Rating data asks respondents to put each alternative on a scale from good to bad. In this way cardinal information on the strengths of preferences is obtained. A special case of rating data is *pairwise comparison*. For example say two alternatives are offered, i and j, and the respondent is to choose from among definitely choose i, probably choose i, neutral, probably choose j, definitely choose j.

Matching: Matching is when one asks direct questions regarding the relative value of attributes that are being traded off. This is most often

done with a price attribute in which respondents are asked to state the price that will make them indifferent between alternatives. Take an example from transportation where travel time and travel cost distinguish the different modal alternatives and therefore a choice of mode represents a trade-off between time and cost (faster modes cost more and vice versa). In a matching exercise, the respondent is presented with two alternatives, the first with a specified time and cost and the second with a specified time but no cost. The respondent is then asked: what cost of alternative 2 would make you indifferent between alternatives 1 and 2? Or, what is the most you will be willing to pay for alternative 2? There are many variants of matching, including bidding and payment card.

Allocation: Gathering even more information, one may ask the respondent to divide a fixed amount of resources (for example, money or time) to a set of alternatives.

There are any other number of other variants of elicitation approaches. For example presenting gambles is common in the decision under risk and ambiguity literature. Further, any particular elicitation method may be the presentation of one stand-alone question to a respondent versus the presentation of a sequence of questions to the respondent. In some cases this is done to gather more information from a given respondent. In other cases it is used to ask successive questions to narrow in on the preference parameter for the individual. And hybrid methods can be used that combine two or more different elicitation approaches.

In the list above, note that one is gradually getting more information with each successive indicator listed above. More information is better, right? Not necessarily. You also must worry about the quality of the information that is obtained. Each successive indicator above is also getting less congruent with what people do in reality and this may cause data quality issues. In the market, consumers are used to facing alternatives with well-defined prices and then making a choice, and so elicitation methods that differ from this market norm (such as matching and allocation) can be problematic.

2.2.3 Potential Sources of Bias

There are many potential sources of bias in stated preference data that one should be aware of. While there is a significant literature on this topic, we briefly introduce a number of biases below and provide limited citations as examples and entries into the literature.

Indifference to the experimental task: As SP surveys are hypothetical without real ramifications to the respondent, respondents may not take the survey seriously and do something to answer the questions as quickly as

possible without considering the choice task. For example, a respondent may focus on a dominant attribute and always choose the alternative that is best for that attribute irrespective of the rest of the description of the alternative.

Policy response bias or protest behavior: Some respondents may give the response they think the surveyor wants to hear, for example by selecting the more politically correct answer. Respondents also may make choices to try to influence the outcome, for example choosing a public transit mode because it is good for society and may reduce road congestion even if the respondent does not intend to take transit. (The satirical journal *the Onion* wrote an article titled "98% of U.S. Commuters Favor Public Transportation For Others," November 29, 2000). Further, some may give extreme responses (for example preferences amounting to an extremely high or an extremely low willingness to pay) or refuse to respond because they protest against some aspect of the survey. The latter has been found in contingent valuation studies that investigate willingness to pay for public goods.

Justification bias: Respondents may respond in ways that justify choices they have made in the past in the real market, without necessarily processing the alternatives as they are defined in the stated preference experiment. This is an inertia effect.

Omission of situational constraints: In the real market, decision makers have various situational constraints that impact their choice, one example being an income constraint. There could also be other constraints like whether they own a bicycle or how many bags they are carrying. While such situational constraints impact choices in the real market, the respondent may not consider these factors in the hypothetical choice experiment.

Incomplete descriptions of alternatives: In designing stated preference experiments, there is always an issue of how completely to describe the alternatives. Because there may be a limit to how many attributes a respondent can process in a stated preference experiment, there is a tendency to limit the number of attributes. However, one has to be careful; the respondent has a tendency to fill in missing information based on the information that *is* provided. For example, say in a residential choice model the quality of schools and level of crime is not described. In these case, respondents may associate high price housing alternatives with high quality schools and/or low crime and vice versa. One should not underestimate the ability of the respondent to be able to handle complex (well designed) tasks.

Framing effects: The presentation of the survey questions can influence the behavioral response, even when there is absolutely no difference in the content of the questions. For example, the order in which questions are presented, the use of absolute versus relative values of attributes, and the presentation of changes in states as either gains or losses.

Cognitive incongruity with actual behavior: Because the setting is significantly different from a revealed choice setting, respondents may use all sorts of different decision protocols that are not consistent with what they would do in reality. This can lead to questions regarding the validity of the response protocol. One example is that respondents may take short cuts such as choosing the attribute that is most important to them (such as price) and ranking the alternatives based on this single variable alone rather than considering all attributes as they would in a real choice setting.

2.2.4 SP Experimental Design

Developing an SP experiment involves defining the context, the number of alternatives, the way the alternatives will be described (attributes and their values), and the choice question. So for the Boeing experiment, the context is a domestic flight that the respondent searched for at the time s/he was intercepted, three alternatives are provided, each alternative is described by specific attributes (number of stops, departure time, arrival time, etc.), and two choice questions (most and least preferred) are used to obtain a respondent's full ranking of the three alternatives. In this section, we focus on a specific aspect of the design, which is, given a particular context and format of the choice question (number of alternatives and attributes), how does one determine the specific values of each attribute that are presented to a respondent? For example, in Figure 2.1, it has to be determined that for this respondent flying from Chicago to San Diego, the choice scenario consists of Option 1; which is nonstop, departs at 6:00 PM, arrives at 8:14 PM, has typical legroom, is operated by Continental on a B737 airplane, and costs $565; and Option 2 and Option 3 as shown in the figure. The choice scenarios will vary over the respondents, sometimes to the point where each individual is provided a unique choice scenario. That is, different respondents will see different values of the attributes for each alternative. Further, these attributes may be customized to the respondent; in the Boeing case, it is customized to the origin and destination that the respondent has entered into the booking website.

The experimental design question is how to determine which specific alternatives to present to each respondent. Each presented alternative is essentially a bundle of attribute values. For a very simple choice experiment, the number of possible combinations may be small enough that the design issue is trivial. For example, a choice of balls where the balls can be red or blue and they can be big or small leads to the 4 bundles (alternatives): small red, small blue, large red and large blue. The choice presented to the respondent can then be among all 4 options, or a random selection of 2 or 3 of the possible

4 types of balls. Even in more complicated situations, domain knowledge and judgment can be used to determine which bundles of attributes are presented to the respondents. However, typically, the choice situation is too complex to use all possible combinations or rely on domain knowledge to determine the presented alternatives. For example, in the choice of balls, the price may be set anywhere from 50 cents to 4 dollars, so what prices will be offered to the respondents?

In cases where the choice context is complex and domain knowledge is not sufficient, there are three general approaches that can be made to determine these values: random draws, factorial designs, and efficient designs.

The *random approach* is the most straightforward to explain and implement. One assumes a distribution of values for each attribute, and alternatives are created by making random draws from these distributions. The distributions can be continuous (for example, a uniform or triangular distribution between two values) or discrete (with either equal or unequal probabilities assigned to each value). Using the ball example, the design may be that the price is uniformly distributed between 50 cents and 4 dollars, there is a 50% chance of a red ball and a 50% chance of a blue ball, and a 25% chance of a small ball and 75% chance of a large ball.

The *factorial approach* is based on the idea that there may be advantages to introducing structure into how the alternatives are generated rather than a pure random approach. Factorial designs aim to increase the *efficiency* of the design, meaning increasing the precision of the parameter estimates (i.e., decreasing standard errors and increasing t-stats) or, equivalently, needing a smaller sample size to obtain a desired level of precision. Factorial design is based on combinatorics and makes use of only discrete distributions with (implicitly) equal probabilities assigned to each value within each distribution. Distributions for the ball example may be: 50 cents, 2 dollars, and 4 dollars with equal probability; red and blue with equal probability; small and large with equal probability. Note the equal distributions can be relaxed by making more categories, for example small, large, large, and large with equal probability results in 25% small and 75% large ball. The so-called *full factorial* is a list of unique alternatives that includes each possible combination of the attribute values. In the ball example, this would be $3 \times 2 \times 2 = 12$ different unique alternatives that capture every possibility. For most designs employed for discrete choice analysis, the full factorial is too large and so a selection of alternatives must be made, a so-called *fraction* of the full factorial.

Factorial design provides methods for selecting a good fraction in terms of identification and efficiency of parameter estimates. The overriding premise is that a good fraction has zero correlation across the attributes, a so-called

orthogonal design. Intuitively, one can think of the extreme opposite, which is perfectly collinear attributes. For example, say the cost of a car trip is calculated as a factor (in terms of cost per distance) times the distance trip and emissions from a car trip is also a factor (in terms of emissions per distance) times the distance of the trip. In this case, the effects are confounded and it is not possible to identify the effects of cost and emissions separately. So avoiding perfect collinearity is essential, whether in linear regression or discrete choice. The orthogonal design pushes this to an extreme by designing for zero correlation across attributes. For linear regression, it can be proven that an orthogonal design is optimal in terms of efficiency (leads to the smallest standard errors of the parameter estimates), which has led to the widespread implementation of orthogonal designs for a wide variety of models, including discrete choice. However, the findings from regression are not applicable to discrete choice, and orthogonal designs are not optimal for discrete choice models.

Therefore, *efficient design* approaches have been developed in order to try to generate optimal designs for discrete choice. While this sounds ideal, the issue is that in order to generate the optimal efficient design, one needs to know the full model specification including parameter values. As the parameter estimates are not known (after all, this is why one is collecting the data), estimates have to be used to generate the efficient design. More recently, *bayesian efficient design* approaches have been developed to reflect the inherent uncertainty of the parameter estimates.

In addition to the three general classifications above, there can also be a combination of approaches. For example, random draws can be made from alternatives generated in a full factorial, rather than using an orthogonal fraction. Also, domain knowledge is often used to edit the designs produced by automated methods, for example by removing alternatives that are unrealistic.

There are advantages and disadvantages to each of these approaches, and the suite should be thought of as a bag of tools to be employed as the situation calls. We turn our attention now to providing more detail on factorial designs as they are a useful and relatively straightforward method for generating SP surveys, and unlike the random design approach require more explanation. We then revisit (much more briefly) efficient design approaches, and end with a discussion of a number of practical issues in SP experimental design.

2.2.5 Factorial experiments

Factorial experiments have been used at least since the 1920s for industrial field experiments in, for example, agriculture (see Fisher, 1926 and Fisher,

1935). This is a big topic and our objective here is to provide the basic information (both theoretical and practical) that is needed to conduct stated preference surveys.

These techniques are employed to analyze the joint effect of several attributes. In the terminology of factorial design, attributes are called *factors*. Factors of interest in the early agricultural field experiments (from Fisher, 1926) included type of fertilizer, timing of application, and quantity applied. Factors in a transport mode experiment might be fare, travel time, and frequency of service. Rather than treating the value which any attribute can take on as a continuous variable, factorial design simplifies the problem by assigning a discrete number of values to each attribute. These values are called *levels*. So, in the transport example, the levels may be: time = fast, slow; fare = high, medium, low; and frequency = infrequent, frequent. The possible alternatives are constructed by combining each level of attribute with every other level of all other attributes; in the language of factorial design these combinations of attributes are called *profiles* or *treatments*. In our transport example, one profile is {fast, high fare, infrequent} and another profile is {slow, medium fare, frequent}. Considering all the combinations in our transport case leads to 12 different profiles: (2 levels of time) $\times$ (3 levels of fare) $\times$ (2 levels of frequency) = 12 unique combinations. In the language of factorial design, this complete list of profiles is called a *full factorial*. Typically, the situations we are dealing with have many more factors and many more levels, leading to a large full factorial. The size of the factorial is what leads to complications in design. However, we stick to a very simple example in order to explain the theory of factorial design. We return at the end to discuss real world applications in which this theory is applied via computer programs.

For our simple example, let's continue with a hypothetical study of a public transportation service, but with 2 levels of each of the three attributes: low/high fare, fast/slow travel time, and infrequent/frequent service. We denote K as the number of attributes, in this case 3, and L as the number of levels, in this case 2. The full factorial has $L^K = 2^3$ profiles, and these are enumerated in Table 2.5. To generate the design, the actual name of attributes and values of the levels are not significant, and so the factorial can be converted to generic numerical coding as shown in Table 2.6 where -1 indicates a poor value of an attribute and 1 indicates a good value. The use of $-1/1$ rather than, say, $0/1$, has a mathematical advantage: if two attributes are orthogonal (meaning, independent), then the inner product of their values, e.g. A^TB, will equal 0. In Table 2.6, attribute A is orthogonal to attribute B: $(1 \times 1) + (1 \times 1) + (1 \times -1) + (1 \times -1) + (-1 \times 1) + (-1 \times 1) + (-1 \times -1) + (-1 \times -1) = 0$. We will see later that orthogonality is a

	Attributes		
Profile	Travel Cost	Travel Time	Service Frequency
1	Low	Fast	Infrequent
2	Low	Fast	Frequent
3	Low	Slow	Infrequent
4	Low	Slow	Frequent
5	High	Fast	Infrequent
6	High	Fast	Frequent
7	High	Slow	Infrequent
8	High	Slow	Frequent

Table 2.5: Simple 2^3 Example: Full factorial with descriptive representation

	Attributes		
Profile	A	B	C
1	1	1	-1
2	1	1	1
3	1	-1	-1
4	1	-1	1
5	-1	1	-1
6	-1	1	1
7	-1	-1	-1
8	-1	-1	1

Table 2.6: Simple 2^3 Example: Full factorial with numerical representation

principle design component in factorial design methods.

Now we have the concepts of factors, levels, profiles and full factorials. The issue in survey design is that full factorials usually contain too many combinations to make use of all of them in the design. For example, we will see later that the definition of attributes and levels in the Boeing experiment led to $360,448$ unique alternatives and $360,448^3 = 4.7 \times 10^{16}$ unique choice sets. The survey targeted a sample size of only $3,000$ respondents, and so we need a method for pairing down this tremendous number of alternatives to those that will actually be used in the survey. This is what fractional factorial design is. Fractions contain usually much fewer combinations. However, in determining the fraction some information needs to be sacrificed. How do we select a good fraction? What is a good fraction? We want to design a fraction that guarantees satisfaction of certain desirable statistical properties, namely the ability to identify the behavioral parameters of utmost interest and to obtain precisely estimated parameters (that is, estimates with small

	Travel Cost	Travel Time	Service Headway
Respondent's Actual Train Trip	\$4.00	60 minutes	30 minutes
Stated Preferences Alternatives			
As absolute changes:			
level -1	+\$1.00	+15 minutes	+10 minutes
level 1	−\$0.50	−15 minutes	−15 minutes
As proportional changes:			
level -1	+25%	+25%	+33.3%
level 1	−12.5%	−25%	−50%
Presentation of Alternatives			
Recent trip	\$4.00	60 minutes	30 minutes
Alternative 1 (Profile 3)	\$5.00	75 minutes	40 minutes
Alternative 2 (Profile 6)	\$3.50	45 minutes	15 minutes

Table 2.7: Attribute levels dependent on respondent's observed behavior

standard errors).

First, what kind of information do we want to retain and what are we willing to sacrifice? Here we need to define the terms *main effect* and *interaction effect*. A main effect is the effect of one variable that is independent of the values of the other variables. An interaction effect occurs when the effect of one variable depends on the value of another variable. In Chapter 1 we introduced an equation representing the attractiveness of an alternative, denoted U, and in Chapter 3 we referred to this function as *utility*. Let's consider this equation in the context of the Boeing example shown in Figure 2.1. The attractiveness of a particular flight itinerary will be a function of all of the attributes presented in the survey. Honing in on in-flight time ("time") and legroom, the attractiveness of a particular flight i for a passenger n may take the form:

$$U_{in} = \beta_1 \text{Time}_{in} + \beta_2 \text{Legroom}_{in} + \beta_3 \text{Time}_{in}\text{Legroom}_{in} + ... \qquad (2.1)$$

The main effects are represented by β_1 and β_2 and the interaction effect is represented by β_3. The interaction term may be a significant behavioral parameter as one would expect comfort issues to be more important for longer flights ($\beta_3 > 0$). However, typically the main effects are considered to be more important than the interaction terms and, given that something needs to be sacrificed in going from the full factorial to a fractional factorial, we often sacrifice being able to make inferences on such interactions.

So, let's continue working with our simple 2^3 example and try to determine a good subset of alternatives (that is, a good fraction) to use. Again, this is a trivial case that will allow us to explicitly present the theory and terminology. In consumer survey research, there would be no need to pair down 8 alternatives; however, you can imagine an industrial experiment in, say, education where you can only have 4 schools in your field test.

First, to explore the concept of main effects and interaction effects, we display the interactions (both 2-way and 3-way) of the attributes in Table 2.8. The interactions are simply products of the attribute values. First note that in this full factorial, all of the columns (whether main effects or interactions) are independent (or orthogonal). It means that each main effect and interaction effect is measurable. If we can only use 4 of the profiles, which four should we choose? Choosing profiles 1, 2, 3, and 4 is a bad choice because the effect of attribute A could not be determined as it does not vary across the chosen profiles. Choosing the profiles 1, 4, 6, and 7 (as displayed in Table 2.9) is a good selection. Why do we say this? First, we have stated that it is important to identify the main effects. This fraction allows this because we have variability in the values of A, B, and C across the four profiles (some values of 1 and some values of -1 in each column) and no pair of A, B, or C is perfectly correlated (no two main effect columns are identical).

In fact, our design is better than merely having variability and no perfect correlation: first, there is an even distribution of high and low values for each attribute and, second, the main effects are independent (inner product of any two main effect columns is equal to 0). We call such a factorial a *balanced* and *orthogonal* design. These are highly desirable (in fact, optimal) properties for regression as it leads to precise estimates of the main effects with minimum standard errors for the given sample size. While not optimal for discrete choice, it is still valuable to generate designs that have little or no correlation. While we obtain excellent estimates for the main effects, what about for the interaction terms? The 2-way interactions are confounded with the main effects and therefore not estimable: AB is identical to C, AC is identical to B, and BC is identical to A. Further, the 3-way interaction ABC is not estimable because it does not have variation over the profiles. In fact, we used the 3-way interaction as a means for selecting our fraction. Using such higher order interactions is a means for selecting fractions that have good properties. In summary, the design in Table 2.9 provides excellent inference for the main effects, although none of the interactions will be observable.

Complete factorial experiments can be divided into blocks, where the block size is smaller than the number of profiles. Then different blocks can be given to different respondents in the sample. Table 2.10 provides an example of blocking for our simple 2^3 example where the balance and orthogonal

design defined above is block 1, and the remaining alternatives are a balanced and orthogonal design defined as block 2. The advantage of this is that it reduces the number of profiles per respondent while maximizing coverage of the full factorial in the sample among respondents.

So far we have employed a simple example to introduce the key concepts and terminology of experimental design, including: factors, levels, profiles, orthogonal coding, full factorial, partial factorial, balanced design, and orthogonal design. Obviously attributes can have more than two levels, there can be more than three alternatives, and different attributes can have different number of levels. Nonetheless, the basic theory described above applies. For example, Table 2.12 provides the full factorial for a 3^3 case (three attributes, three levels each) and Table 2.13 provides a 3^{3-1} balanced and orthogonal fraction. Montgomery (2008) provides detail on how this fraction is generated and the resulting confounding. Note that while the coding for a two level factor was $(-1, 1)$ the coding for a three level factor is $(-1, 0, 1)$. In general, the coding is such that factor levels sum to zero and consist of only odd numbers (and zero). A 0 is included for an odd number of levels and is not included for an even number of levels. Therefore, a four level factor would be coded (-3,-1,1,3), a five level coded (-3,-1,0,1,3), and so forth. With such coding, the orthogonality condition can be trivially checked by verifying the inner products are zero.

Another wrinkle is that while we have discussed only how to generate profiles for a given alternative, in many stated preference surveys used for choice modeling, the respondent is presented with a number of alternatives (2, 3, 4, etc.) from which to make a choice. Therefore the design needs to produce choice sets rather than individual alternatives. One option is to create each choice set by randomly selecting the desired number of alternatives from the generated fraction. Alternatively, the process of generating a fraction above can be trivially extended to represent a choice set. Instead of creating a design based on a single set of attributes (with associated levels), we create a design based on multiples of these attributes. Define A to be the number of attributes, L to be the number of levels per attribute and M to be the number of alternatives in the choice set. To generate a choice set with M alternatives, we generate attribute values M times. The number of all possible choices sets is then L^{MA} and this is often called an $L-M-A$ design. (Note that here we stick to the notation used in factorial design, but in the rest of the book we use K rather than A for number of attributes, J rather than M for number of alternatives.) Table 2.11 displays an L^{MA} design for the case of a choice set containing two alternatives (car versus bus), where each alternative has two attributes (travel time and travel cost), and each attribute has two levels (good and bad). The full factorial is shown with

	Attributes			Interactions 2-way			3-way
Profile	A	B	C	AB	AC	BC	ABC
1	1	1	1	1	1	1	1
2	1	1	-1	1	-1	-1	-1
3	1	-1	1	-1	1	-1	-1
4	1	-1	-1	-1	-1	1	1
5	-1	1	1	-1	-1	1	-1
6	-1	1	-1	-1	1	-1	1
7	-1	-1	1	1	-1	-1	1
8	-1	-1	-1	1	1	1	-1

Table 2.8: Simple 2^3 Example: Interactions in Factorial Design

	Attributes			Interactions 2-way			3-way
Profile	A	B	C	AB	AC	BC	ABC
1	1	1	1	1	1	1	1
4	1	-1	-1	-1	-1	1	1
6	-1	1	-1	-1	1	-1	1
7	-1	-1	1	1	-1	-1	1

Table 2.9: Simple 2^3 Example: Fractional Factorial Design

$2^{2\times2} = 16$ profiles, which are now choice sets. The same process of selecting balanced and orthogonal fractions can be conducted to determine a good selection of profiles to use in the experiment.

2.2.6 Selecting attributes, levels, and attribute values

So given the concepts of attributes and levels, how does one decide the number of attributes, number of levels and the attribute values? This is as much an art as it is a science. There are a number of things to keep in mind. In selecting the number of attributes and levels per attribute, there is a trade-off between realism and complexity. On the one hand, one wants to ensure that the alternatives are adequately described (so more attributes), because otherwise respondents will tend to fill in information that is not provided (see discussion on incomplete descriptions of alternatives). However, more attributes leads to increased complexity in the choice question being presented to the respondent (perhaps with more information than they can or will process) and also requires a larger experimental design (a larger frac-

				Interactions				Blocks
	Attributes			2-way			3-way	
Profile	A	B	C	AB	AC	BC	ABC	
1	1	1	1	1	1	1	1	1
2	1	1	-1	1	-1	-1	-1	2
3	1	-1	1	-1	1	-1	-1	2
4	1	-1	-1	-1	-1	1	1	1
5	-1	1	1	-1	-1	1	-1	2
6	-1	1	-1	-1	1	-1	1	1
7	-1	-1	1	1	-1	-1	1	1
8	-1	-1	-1	1	1	1	-1	2

Table 2.10: Simple 2^3 Example: Defining Blocks

	Alt 1 (Car)		Alt 2 (Bus)	
Choice Set	Time	Cost	Time	Cost
1	-1	-1	-1	-1
2	-1	-1	-1	1
3	-1	-1	1	-1
4	-1	-1	1	1
5	-1	1	-1	-1
6	-1	1	-1	1
7	-1	1	1	-1
8	-1	1	1	1
9	1	-1	-1	-1
10	1	-1	-1	1
11	1	-1	1	-1
12	1	-1	1	1
13	1	1	-1	-1
14	1	1	-1	1
15	1	1	1	-1
16	1	1	1	1

Table 2.11: Example of Constructing Choice Sets

	Attributes			Interactions			
Profile	A	B	C	AB	AC	BC	ABC
1	-1	-1	-1	1	1	1	-1
2	-1	-1	0	1	0	0	0
3	-1	-1	1	1	-1	-1	1
4	-1	0	-1	0	1	0	0
5	-1	0	0	0	0	0	0
6	-1	0	1	0	-1	0	0
7	-1	1	-1	-1	1	-1	1
8	-1	1	0	-1	0	0	0
9	-1	1	1	-1	-1	1	-1
10	0	-1	-1	0	0	1	0
11	0	-1	0	0	0	0	0
12	0	-1	1	0	0	-1	0
13	0	0	-1	0	0	0	0
14	0	0	0	0	0	0	0
15	0	0	1	0	0	0	0
16	0	1	-1	0	0	-1	0
17	0	1	0	0	0	0	0
18	0	1	1	0	0	1	0
19	1	-1	-1	-1	-1	1	1
20	1	-1	0	-1	0	0	0
21	1	-1	1	-1	1	-1	-1
22	1	0	-1	0	-1	0	0
23	1	0	0	0	0	0	0
24	1	0	1	0	1	0	0
25	1	1	-1	1	-1	-1	-1
26	1	1	0	1	0	0	0
27	1	1	1	1	1	1	1

Table 2.12: 3^3 Full factorial and interactions

Profile	A	B	C
1	-1	-1	-1
5	-1	0	0
9	-1	1	1
12	0	-1	1
13	0	0	-1
17	0	1	0
20	1	-1	0
24	1	0	1
25	1	1	-1

Table 2.13: 3^{3-1} Balanced and Orthogonal Fraction

tion). In terms of the number of levels, more levels provide more information but also require a larger fraction for the experiment. At least three levels are needed if one wants to capture non-linear impact of attributes, and it is recommended to use as many levels as is feasible. In terms of the values of the levels, typical values are chosen such that they are both plausible and relate to the respondent's experience. The values are often based on domain knowledge, for example in terms of the range of computer processor speed or laptop size.

It can be helpful to personalize the design based on the real-world choice context of the respondent. For example, in our public transport example, the attribute levels can be made dependent on the characteristics of the respondent's real world choice context such as an actual bus trip to, say, work. In this case, the levels are deviations (in either absolute changes or proportional changes) from the attributes of an actual trip that the respondent has taken recently and described as part of the survey. This concept is show in Table 2.7. The choices presented to the respondent are then presented in relation to this recent trip. Similarly, the design can be customized to account for situational constraints. For example, whether or not someone owns a car and/or has a membership to a car sharing service would impact the auto alternatives presented (or not) and the costs of the trip.

Another important issue is that the choice of attribute levels may place bounds on the tradeoffs that can be observed. This is best shown by example. Say our experiment involves a public transport mode choice, where the choice presented to the respondent is between a faster but more expensive mode and a cheaper but slower mode. In this case there are two factors: time and cost. The important tradeoff in such an example is referred to as the value of time, represented in units such as $/hour, which represents the amount of money a

traveler is willing to spend to save a unit of time (e.g., an hour) traveling on a trip (see the discussion in Section 3.6.1). Say we give each factor three levels: 40, 50 and 60 minutes for time and 0, 1, and 2 dollars for cost. Enumerating all combinations leads to 81 unique choice experiment profiles: $3 \times 3 = 9$ options for alternative 1 and $3 \times 3 = 9$ options for alternative 2, so $9 \times 9 = 81$ combinations of alternatives 1 and 2. However, once we remove combinations where the two alternatives are identical (for example {40 minutes, $1} versus {40 minutes, $1}), combinations where one alternative is clearly dominant over the other (for example {40 minutes, $1} versus {50 minutes, $1}), and combinations that are a repeat where the slow and expensive mode simply swap places (for example, alternative 1 = {40,1} and alternative 2 = {50,0} versus alternative 1 = {50,0} and alternative 2 = {40,1}), we are left with the 9 profiles shown in Table 2.2.6. Now if we look at the cost-time tradeoff represented by these profiles (also shown in the table), the range only varies from $3/hour to $12/hour. This is going to be an issue if values of time are outside of this range, because such values of time cannot be observed from this survey design. Figure 2.2 displays graphically the trade-offs represented by the profiles as the four dots on the graph and the observable range of value of time (the minimum and maximum possible slope). Therefore it is critical that one verifies that the trade-offs expected in the data are within this observable range, and make modifications if necessary by changing the number of levels, the range of values or the increments between levels. Using unequal increments between attributes levels will generate greater variability in the tradeoffs that are represented in the design. For example, the even increments shown in Table 2.2.6 (10 minute increments for travel time and $1 increments for travel cost) resulted in only 3 different tradeoffs in the design ($3, $6, and $12 per hour). Table 2.15 shows that with time levels of 40, 55, and 60 minutes and cost levels of $0, $0.75, and $2, then nine different tradeoffs (rather than three) are represented ranging from $3 to $24 per hour.

2.2.7 Efficient designs

While orthogonal designs are widely used and are a useful tool in the sampling toolbox, there are questions regarding their appropriateness for discrete choice analysis. The motivation behind orthogonal designs is to be able to identify (i.e., estimate) the most important parameters in the model (the main effects) and also to be able to estimate these parameters with precision (i.e., with minimal standard errors). While orthogonal designs are efficient (produce estimates with minimum variance) for linear regression, they are not necessarily efficient for discrete choice. There are a number of aspects of discrete choice models that complicate the issue.

Profile	[Alt. 1] Minutes	$	[Alt. 2] Minutes	$	Difference in time	Difference in cost	Tradeoff ($/hour)
1	40	1.00	60	0.00	-20	1.00	3.00
2	40	2.00	60	1.00	-20	1.00	3.00
3	40	1.00	50	0.00	-10	1.00	6.00
4	40	2.00	50	1.00	-10	1.00	6.00
5	50	1.00	60	0.00	-10	1.00	6.00
6	50	2.00	60	1.00	-10	1.00	6.00
7	40	2.00	60	0.00	-20	2.00	6.00
8	40	2.00	50	0.00	-10	2.00	12.00
9	50	2.00	60	0.00	-10	2.00	12.00

Table 2.14: Profiles for mode choice example

Table 2.15: Revised mode choice profiles with unequal level increments

Profile	[Alt. 1] Minutes	$	[Alt. 2] Minutes	$	Difference in time	Difference in cost	Tradeoff ($/hour)
1	40	0.75	55	0.00	-15	0.75	3.00
2	40	2.00	55	0.75	-15	1.25	5.00
3	40	0.75	50	0.00	-10	0.75	4.50
4	40	2.00	50	0.75	-10	1.25	7.50
5	50	0.75	55	0.00	-05	0.75	9.00
6	50	2.00	55	0.75	-05	1.25	15.00
7	40	2.00	55	0.00	-15	2.00	8.00
8	40	2.00	50	0.00	-10	2.00	12.00
9	50	2.00	55	0.00	-05	2.00	24.00

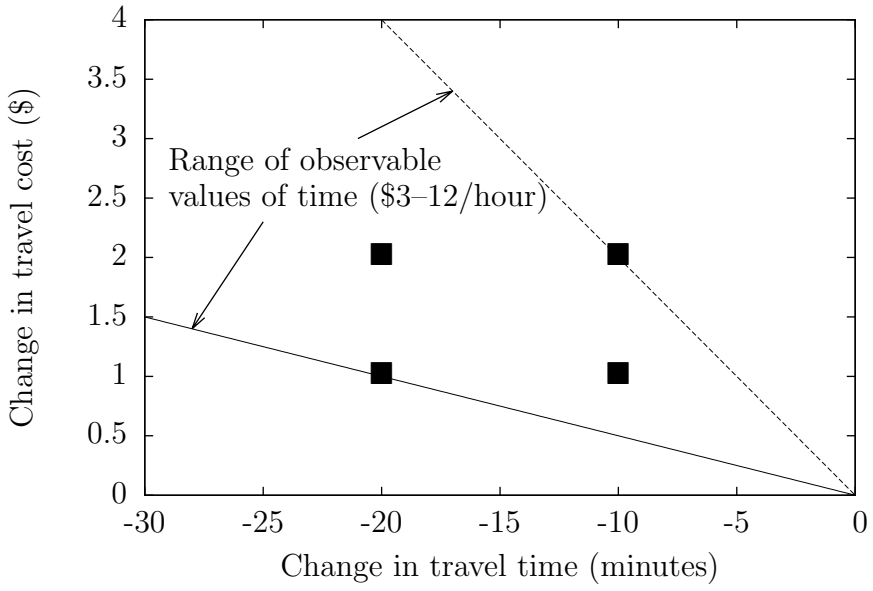


Figure 2.2: Choice of levels may limit range of observable behavior

One reason why the results are not transferable is that in discrete choice, the standard error of the parameter estimate is a function of the value of the parameter itself. Therefore, one cannot optimize the sample for discrete choice without knowing the value of the parameter that one is collecting the sample to estimate. This is in direct contrast to regression. To demonstrate this difference, we start the discussion with simple univariate models of regression and discrete choice. Consider a simple linear model as follows: Let

$$y_n = \alpha + \beta x_n + \xi_n, \tag{2.2}$$

where α, β and x are scalars and ξ is an independent disturbance with mean 0, variance σ^2. The least squares estimator of β is

$$\widehat{\beta} = \frac{\sum_{n=1}^{N_s}(x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^{N_s}(x_n - \bar{x})^2}, \tag{2.3}$$

where $\bar{x}$ and $\bar{y}$ denote the sample averages for x and y, respectively. The variance of $\widehat{\beta}$ for a sample stratified on the basis of the independent variable x is given by

$$\mathrm{Var}(\widehat{\beta}) = \frac{\sigma^2}{\sum_{n=1}^{N_s}(x_n - \bar{x})^2}. \tag{2.4}$$

In this case the variance of $\widehat{\beta}$ is not affected by the value of the parameter β. Therefore, minimizing $\mathrm{Var}(\widehat{\beta})$ does not require any knowledge of β. By examining this equation, one can see that choosing extreme values of x, that is, an equal number of very small and very large value of x_n, will maximize $\sum_{n=1}^{N_s}(x_n - \bar{x})^2$. The practical implication for SP design is that in determining the value of each level, one wants the values to span a large range of values, as large as is credible to the respondent. If the effect is expected to be non-linear, then one would include a number of levels to span the range between the largest and smallest values.

Now consider a simple binary logit model, where

$$P(i|x_n) = \frac{1}{1 + e^{-\beta x_n}}, \tag{2.5}$$

and β and x are scalars and x_n is defined as $x_{in} - x_{jn}$.

If the sample is an exogenous one and maximum likelihood estimation is used, then by the Cramér-Rao bound the asymptotic variance of $\widehat{\beta}$ is

$$\mathrm{Var}[\widehat{\beta}] = \frac{-1}{E[\partial^2 \mathcal{L}/\partial \beta^2]} = \frac{1}{N_s}\left(E\left[\frac{x_n^2 e^{\beta x_n}}{(1 + e^{\beta x_n})^2}\right]\right)^{-1}. \tag{2.6}$$

Now note that if we wish to minimize the asymptotic variance of $\widehat{\beta}$ for a given sample size N_s, we should maximize the term inside the expected value. The key here is that this is a function of the value of β, so the only way to generate an optimal design for x_n that minimizes $\mathrm{Var}[\widehat{\beta}]$ is to know the value of β.

If we were to adopt an analog to the standard linear model, we might try to do this by choosing x_{in} and x_{jn} as different as possible, thus moving x_n as close to $\pm\infty$ as practical. However, as long as $\beta \neq 0$,

$$\lim_{x \to \pm\infty} \frac{x_n^2 e^{\beta x_n}}{(1 + e^{\beta x_n})^2} = 0, \tag{2.7}$$

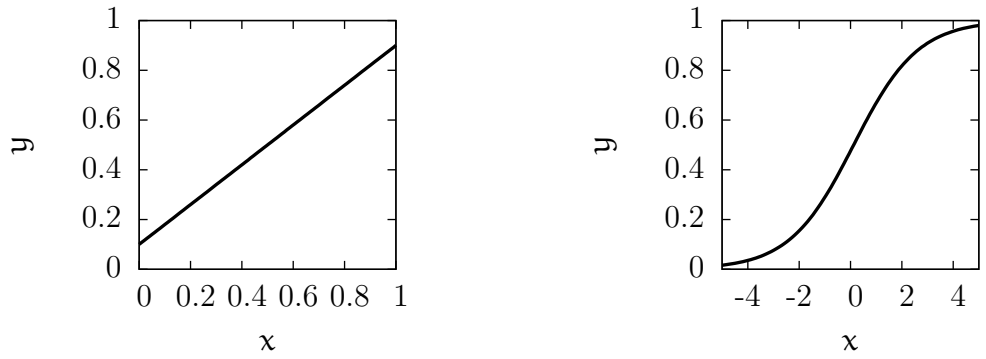
which is its minimal, not maximal, value. Thus not only is the sampling rule analogous to the one for the regression model not optimal when $\beta \neq 0$, it is the worst possible rule.

However, it is interesting to note that for the case where $\beta = 0$, this result changes entirely:

$$\frac{x_n^2 e^{\beta x_n}}{(1 + e^{\beta x_n})^2} = \frac{x_n^2}{4} \tag{2.8}$$

which is now maximized for large values of x_n^2. Thus the same experimental design that is the worst one for some values of the parameters is the best one for another value.

The intuition for these results can be seen by examining Figure 2.3. To infer the slope of the line in linear regression, the worst case is if the x are all clustered together (where the fitted line would pivot around the clustered points), and the best case is where the x are spread wide and the line can be fit through the extremes. Note that the actual slope of the line does not impact the analysis of the best design for x. For the discrete choice curve, extreme values of x give no information on the slope of the curve because the curve is flat at the extremes. For discrete choice, it is most critical to capture values of x that are located where the curve is sloping. Further, the range of the slope (that is, the values of x around this critical region) depends on the value of the β. When $\beta = 0$, the logit curve is a horizontal line, and therefore observing that it is flat will best be observed with extreme values of x.



(a) Example of linear regression

(b) Example of binary logit

Figure 2.3: Linear vs nonlinear models

This result generalizes as follows: There is no single sample design for discrete choice analysis that is unambiguously optimal for all values of the parameters. Rather, whether a sample design is good or bad (in the classical sense) depends on the unknown parameter values.

With this context in mind, we introduce *efficient designs*. The idea of the efficient design is to let the concept of orthogonality go, and to focus on generating a set of profiles such that estimation with the given profiles will minimize the standard errors of the estimated parameters. The optimality functions are based on the variance-covariance matrix of the estimated parameters, which is the multi attribute version of (2.4) for linear regression and of (2.6) for discrete choice. The optimality procedures generally aim to optimize over a summary index of these standard errors, such as the trace

or the determinant of the variance-covariance matrix. As in the univariate case worked through above, in minimizing the standard errors in regression it is useful to spread the values of all attributes as wide as credible but for discrete choice it is desirable to concentrate the values of the attributes in the region where the probability function is sloping (which is a function of the parameter values). Further, in multi attribute settings in regression, it has been proven that the standard errors are minimized when an orthogonal fraction is used, that is when there is no correlation among attributes in the resulting fraction. However, this is not true for discrete choice. Indeed, some correlation in the x variables is advantageous.

To demonstrate this, we refer to Figure 2.5. This builds on the simple mode choice experiment with time and cost attributes that was introduced in Section 3.6.1 to demonstrate parameter estimation and used above in Section 2.2.6 to demonstrate issues regarding the choice of attribute levels. Here we revisit the parameter estimation plot of Figure 3.9 within the context of efficient experimental design. First assume that the true value of the parameter β that we wish to estimate (the slope of the line) is known and is as plotted in each of the graphs for reference. An experimental design generates a number of profiles to present to survey participants, where each profile has specific values of t_1, t_2, c_1, and c_2 and the respondent is asked to make a choice between alternative 1 (with time t_1 and cost c_1) and alternative 2 (with time t_2 and cost c_2). For this example, we employ 8 levels for each of the four attributes and generate approximately 100 different profiles for each design method (a fraction size that is well beyond a minimal main effects design). On the graphs, each point represents a generated profile, where the difference in time $(t_1 - t_2)$ between the two alternatives is plotted on the horizontal access and difference in cost $(c_1 - c_2)$ on the vertical access.

Figure (a) shows results from an orthogonal design in which the t_1, t_2, c_1, and c_2 are uncorrelated. In this case, the points are generally scattered in all four quadrants of the figure, thus generating many choice sets that are uninformative because one alternative dominates the other (is both faster and cheaper). Figure (b) shows the same orthogonal design, but where such choice sets with a dominant alternative are removed from the fraction. The resulting fraction is no longer orthogonal, but improves the efficiency by only offering choice sets where the respondent must make a trade-off between time and cost.

The remaining four plots on the figure all are generated using efficient design methods, based on some optimization of the variance-covariance matrix. Plots (c) and (d) display fractions generated by an efficient design method called D-efficient design, where the optimality criterion is based on minimizing the determinant (thus "D") of the variance covariance matrix.

Conceptually, what the efficient design tries to do is to maximize the information that is obtained from each choice situation. In a choice context, more information is obtained when the alternatives are relatively closer in *overall* attractiveness. Therefore, an efficient design does not produce as many choice contexts in which one alternative is dominant, but instead produces more where the respondent has to be more discriminating in making the choice. However, knowing the region in which choices require more discrimination requires knowledge of both the utility specifications and the parameters. Since the parameters are not known, different assumptions of the parameter values are used as priors in creating efficient designs. These assumptions range from assuming zero-parameter vectors, to fixed parameter vectors, to sign constraints for some or all parameters, to the use of Bayesian techniques that employ priors recognizing the inherent uncertainty of the parameter estimate. The effectiveness of the efficient design rely on the accuracy of the estimates and the resemblance of the final model specification to the one used for design.

In figure (c), the profiles are generated by inputing the correct prior assumption on the values of the parameters. In this case, the generated profiles are all nicely scattered around the line. However, in figure (d), the correct prior is used for β_c, but an incorrect prior of 0 is entered for β_t. In this case, the prior information used for design is a value of time of 0/hour, and the generated profiles are scattered along this horizontal line. The key point is that the value of the prior has a significant impact on the design that is generated. Further, in a real application one does not know the true values of the parameters. To address the issue that the values of the parameters are unknown, Bayesian designs are used. Bayesian designs also use priors, but they are in the form of a distribution with known parameter values. Plots (e) and (f) present Bayesian designs. The prior for both cases is a normal distribution on the value of time with a correct mean of 18/hour. Figure (e) assumes a tighter distribution with a standard deviation of 6/hour and figure (f) a wider distribution with a standard deviation of 12/hour. In both cases the points are still clustered around the mean value of the distribution (the line on the plot), but are more spread out than the D-efficient design in plot (c). The larger the assumed standard deviation, the more spread the cloud of points.

While these are just a sampling of the types of designs that have been developed and used, our aim was to demonstrate in a simple example how different experimental design methods operate.

2.2.8 Practical Issues in SP Design

Above we presented the basics of generating fractional factorial designs, starting with describing orthogonal designs and then discussing efficient designs. So far we have dealt with the issue on a more theoretical level, and in this section we focus on a more practical level. Just like modeling is an art in which the best model is decided upon by balancing information from statistical tests with a priori beliefs of the behavior and the intended application of the model, generating an SP design is also an art. The theory presented above for both orthogonal and efficient designs should not be applied without careful thought by the modeler. In generating a design and developing an SP survey, there are three key guiding objectives:

Identification: Being able to estimate the parameters of the model of interest.

Efficiency: Generating estimates of the parameters that have relatively smaller standard errors on the parameter estimates given the size of the sample.

Credibility: Presenting choice alternatives that appear credible to the respondents.

Identification and efficiency were emphasized in the discussion above. Credibility is added here as a practical issue, because respondents may question the credibility of the experiment if they are given choices that do not seem reasonable. For example, the orthogonal design process may generate alternatives that have a very low price with very high quality on most or all dimensions or vice versa. Not only is this an efficiency issue (because it makes for a more obvious accept or reject of an alternative and therefore does not provide much information to the analyst on preferences) it is also a credibility issue because one is not giving the respondent alternatives that they may hope to see in the market. That said, the ability to stretch the alternatives beyond what is available in the existing market is a major advantage of SP, it just needs to be done without pushing the bounds of credibility.

Above, many different methods were introduced for generating SP designs, including random, orthogonal, and efficient. The discussion was focussed on the theory behind different designs and on defining terms using simple examples. While this theory is nice, most problems are too large to work through in such a simple manner, and the design process described above are simply guiding principles that do not need to be applied religiously. There exist design catalogs (for example, see Hahn and Shapiro, 1966) that provide suggested (orthogonal) fractions for many, many different attribute and level combinations. Today, most fractions, whether orthogonal or efficient, are designed via use of computer programs. For example, SAS has

numerous routines (see Kuhfeld, 2005 and Inc., 2010) and Ngene is targeted specifically for choice modeling (see Cho, 2012). In the days of paper and pencil surveys, it was fairly critical to be able to generate small samples to limit the proliferation of unique survey instruments. With computer generated surveys, it is less critical to generate small fractions. However, as we will see in the Boeing case below, the size of the problem often requires that a fraction of the full factorial be used. The principles described above are used to generate such fractions. The software programs produce designs of various sizes, leaving the analyst the choice of how large of a fraction to use. In the language of orthogonal designs, the smallest fraction that is generally considered is a *main effects* design, which enables estimation of all of the main effects in a linear regression. The number of profiles needed for such a main effects design is equal to $1 + \sum_{(a=1...A)} (L_a - 1)$, where L_a is the number of levels for attribute a. This calculation assumes that different parameters will be estimated for all but one level of each attribute (called *part-worths*), where the normalization of one level per attribute is required for identification (see Section 5.4). Often other specifications are used for discrete choice, for example the parameter values are constrained to be the same across alternatives (defined in later chapters as a generic specification) or the effect of the attribute is assumed to be linear and a single parameter is estimated for an attribute rather than estimating each part-worth. Unless there is a practical issue that requires the use of a very small fraction, it is best to use the largest fraction that is possible as more variability in the data is better. With computer generated surveys, each replication for each respondent can be a unique profile, so the maximum size fraction is the number of respondents times the number of replications.

Whatever process is used to generate the design (random, orthogonal, efficient, bayesian), the fraction produced should be carefully cleaned and checked before the survey is conducted. For example, cleaning is necessary to remove alternatives or choice sets that are deemed not credible or do not provide much information on preferences (e.g., cases of a dominated or dominating alternative). The amount of cleaning needed will vary based on the profile generation process. Orthogonal designs tend to need significant cleaning, but are nonetheless a good starting point for the design process. Given that most surveys are now done by computer, one should generate a very large fraction in the design process, and then, if necessary, clean it substantially. For this reason, even with orthogonal designs, the resulting design used is rarely orthogonal and balanced. However, while for regression this would be considered degrading the design, in discrete choice it is an improvement. Indeed the part of the cleaning process that removes dominating and dominated alternatives (thereby introducing correlation among

the attributes) improves the efficiency of the parameter estimates (as demonstrated in Figure **??**. One also needs to verify that the range of observable trade-offs (as discussed with Figure 2.2) is sufficiently broad to not restrict estimation.

Once the generated fraction is checked and cleaned, another critical step is to do a Monte Carlo exercise to ensure that the generated design can be used to estimate the parameters of the model of interest. First, hypothesize the model of interest, including both the exact specification and specific values for the parameters. Second, generate a set of choice questions that mimic those that will be given to respondents in the experiment. This will be based on the cleaned fraction created above and how it will be applied to the sample collected (e.g., choice questions drawn at random for 400 respondents). Third, use the hypothesized model to generate a chosen alternative for each choice question generated in the second step. Finally, use the dataset generated in step three to estimate the parameters of the hypothesized model, and verify that the parameters can be estimated within a reasonable amount of precision. An ad-hoc measure is to check that the parameters are estimated within one (or, at most, two) standard errors of the true parameters. If this is not the case, then one should revisit the design process to check for issues such as collinearity and limited ranges of attribute values. This is not an assurance that the experiment will produce the desired results, but it is a good check that the design is sufficient to estimate the parameters of the hypothesized model.

Beyond the decisions made in generating the fraction, there are a number of other practical issues and design guidelines in designing and executing a stated preference survey and in the analysis that follows. Many of these guidelines are a result of trying to diminish the potential biases presented in Section 2.2.3.

First, there are a number of factors to consider in designing the context of the choice experiment. It is important to make an effort to get the respondent within the frame of mind of making a real choice in the market; the more the subject can relate the experiment to a real world context, the closer their responses will be to their real world behavior. This can be done in the manner the experimental task is introduced and in terms of the manner in which the choice environment is presented. Often the respondent is asked to think about a specific purchase or decision he has recently made that is related to the choice experiment. Sometimes the choice experiment is then personalized by pivoting the alternatives presented off this real world experience, for example as described in Section 2.2.6. The consideration to get subjects in the frame of mind of a real world decision suggests making choice experiments that are very realistic. However, realism typically comes

at the price of complexity, which is another important design issue. As realism and complexity are at conflict, one needs to find a balance between these that serves the needs of the analysis and will depend based on the application. One thing to keep in mind is that humans can handle very complex choice decisions. Starbucks has used in its advertising the statistic that customers who purchase a drink at one of their shops have made a choice from among $87,000$ different drink combinations. Another design issue is how many responses one obtains from each respondent. It is relatively easy to obtain multiple responses from each individual in a stated preference experiment. This leads to a lower cost per observation of each preference, however the quality of the data also diminishes as people may get fatigued and provide their preferences with less precision. The decision on the number of experiments per respondent depends on the budget, respondents' patience, and the context of the SP survey (onboard vs. home-based interview, self-completion survey).

- survey design
  - neutrality - suggest to ensure the survey doesn't come across as an opinion type of survey to reduce policy response bias
  - realism - think of possible constraints and build into survey (bags, bike ownership... via details on recent and specific trip). fully describe alternatives and constraints. size of survey question not as critical. people can process a lot if well designed and motivated. no obviously impossible or unlikely alternatives.
  - qualitative investigation to guide design e.g. Klojgaard, Bech, and Sogaard journal of choice modeling 5(2). focus groups, debriefing interviews, beta tests (bennett and adamowicz 2001, louviere, hensher and swait 2000, kanninen 2006, kaplowitz, lupin and hoehn 2004... from)
  - incentive compatible designs
- augment design
  - vary experimental design - vary questions. use different elicitation methods. framing:shuffle order of presentation of attributes across respondents; frames attribute levels the way people receive information about them the marketplace; don't create unrealistic reference points. (paper maya sent 6/8/13 kraft and bennett). hoehn, lupi and kaplowtitz vary question format (info format table vs text)

  - ask additional questions re process (attribute non-attendance) self reported indicators of non-attendance to attributes to constrain to zero the parameters of the non-attended attributes. strong political leanings, e.g. protest behavior.

- examine data
  - unreasonable survey completion times (too fast), choice of dominated alternatives, choice along prominent attribute (choice is always based on value of a prominent attribute, e.g. least cost alternative), look for observations with unreasonable willingness to pay values

- econometric analysis
  - latent class, non-attribute attendance(ref: hess, hensher, rose), protest response (Meyerhoff and Liebe) Attribute non-attendance: latent class model, with one or more classes corresponding to non-attended attributes with zero parameters. issue is that can't tell whether sp non-attendance reflects true rp behavior or is a pure sp artifact. can use rp choice as a dummy variable in the choice model to account for justification bias (need to consider endogeneity)

2.2.9 Example revisited

Now we return to the Boeing survey introduced at the beginning of the chapter to work through the process of experimental design using a real case. Recall that there are three alternatives in the choice set: a non-stop, a one-stop with no airline change and a one-stop with an airline change. Each respondent is presented with three itineraries and are asked to select both the best and worst alternatives. An L^{MA} design was used with a fraction of 256 choice sets. This was a main-effects-only design.

One of the first steps in design is to determine the factors and levels for the design. Those used in the Boeing case are shown in Table 2.16. The attribute levels were customized to the characteristics of the trip/market. For example, markets were defined based on the flight length and direction. Departure time values were assigned by market to ensure that all alternative flights depart and arrive at reasonable times. There are 7 factors with the number of levels per factor ranging from 4 levels (base fare, legroom, stop penalty) to 11 levels (airline). This leads to a full factorial of $4^3 \times 8^3 \times 11 = 360,448$ alternatives and (given 3 alternatives per choice set) of $360,448^3 = 4.7 \times 10^{16}$ unique choice sets. As the survey was conducted on around 3,000 respondents, each

receiving just one treatment, this amounts to less than 1% of the full factorial being used. The question is how to determine which choice set profiles to include in the survey? First, we can look at the minimum necessary for a main effects design. Using the equation above, this equals $3\times(4-1)+3\times(8-1)+(11-1)=40$ main effects per alternative. Since we have three alternatives, in order to estimate alternative-specific main effects we have 40×3 plus 2 alternative specific constants leads to 122 main effects. Therefore the design should have at least 122 choice sets. Typically one employs more than the absolute minimum and this is what was done in this case as 256 was used. Often sample sizes of powers of 2 are generated simply because it simplifies the process of creating a nice fraction (orthogonal, balanced, etc.). In this case 256 is a near power of 2 to the necessary 122 choice sets. One choice set from the 256 was randomly drawn for each respondent in the sample. Note that a larger fraction, one up to the 3,000 respondents, could have been used in this case.

Additional data were collected along with the choice responses. These data include socio-economic variables, such as age, income, gender, home zip code, occupation, and education. Situational variables were also collected, including travel dates and times requested to the internet booking service, whether departure time or arrival time is more important and the ideal departure or arrival time, the trip purpose, who is paying for the trip, and the number in the travel party. Finally, the flight history in terms of number of round trip flights the respondent took in the last year including the total and for the requested origin-destination pair.

These data were then used to develop models of air traveler itinerary choice for Boeing (Garrow et al., 2006 and Brey and Walker, 2011).

2.3 Statistical inference and sampling

Above we discussed what data to collect and how to design the survey. In this section we discuss the issue of from whom to collect the data. Our purpose of sampling is to make statements about the values of unknown population characteristics based on characteristics of a sample drawn from the full population. Conducting a census of the population to measure these characteristics is both prohibitively costly and also unnecessary. The power of statistics and statistical inference is that it allows us to provide reasonable estimates of the population based on a sample.

In this section we describe the sampling process and discuss issues of sample size. We define a random sample as well as discuss other types of samples that may be used and why. In this discussion, we primarily review

Table 2.16: Factors and levels for Boeing Experiment

Factor	# Levels	Values of Levels	Notes
1: Base fare	4	-25%, 0%, 25%, 50%	Represents the percent change in the average fare for the market. The same for all alternatives in the choice set for any respondent. Varies across respondents.
2: Premium fare	8	-15%, -10%, -5%, 0%, 5%, 10%, 15%, 20%	Percent change from the base fare. Varies across each alternative and respondent.
3: Departure time	8	time 1, 2, 3, 4, 5, 6, 7, 8	Origin-destination specific to ensure reasonable departure and arrival times
X: Arrival time	–		Determined by departure time, stop time, and minimum flight time for a nonstop in that origin-destination market.
4: Stop penalty	4	60, 90, 120, 150 minutes	Represents additional trip time for alternatives with a stop
5: Legroom	4	typical legroom, 2 inches less, 2 inches more, 4 inches more	
6: Airplane type	8	airplane 1, 2, 3, 4, 5, 6, 7, 8	Specific airplane names held back for proprietary reasons
7: Airline	11	airline 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	Specific airline names held back for proprietary reasons

basic concepts that were originally developed for estimating simple population characteristics, and we first present them in this context. At the end, we draw connections with discrete choice analysis. The sampling problem for discrete choice analysis is conceptually no different. Instead of seeking to estimate population means or variances (as emphasized in basic sampling theory), we are trying to infer the vector of parameters of the choice model (for example, the β vector in Equation 2.1) and resulting distributions of the characteristics of the population such as the demand function.

2.3.1 The sampling process

Sampling theory addresses how to estimate characteristics of a population (such as average income) with adequate accuracy while saving money, saving time, reducing survey administration problems, and minimizing intrusion.

Any discussion of sampling must necessarily begin with the definition of the *population* being studied. In many situations this may be far more complicated than one would first think. For example, the pool of potential travelers in an urban context consists of more than just all the residents of the metropolitan area; it also includes possible visitors and individuals passing through the city. In addition we typically adopt a rather arbitrary definition of the geographical extent of the area under study, often using jurisdictional boundaries such as county or town lines.

Populations can be treated as having either a *finite* or *infinite* number of members. Of course no population is truly infinite. The distinction here is whether or not the population is large enough—when compared to the size of the sample taken from it—that it can be treated as infinite relative to the size of the sample. Since most of the populations we use in concrete studies are quite large and the samples we use are small, we focus here almost exclusively on the simpler case of infinite populations.

A second major concept is the *sampling unit*. The sampling units must be defined to be mutually exclusive and must collectively exhaust the population. It can be households, firms, individuals, or any relevant entity.

Taken collectively, an implicit or explicit list of all the sampling units constitutes the *sampling frame*. Such lists are often derived from utility records, telephone listings, motor vehicle registry lists, or private directories of firms. Very often the creation of an appropriate sampling frame may be extremely difficult. The proliferation of mobile phones has made the use of landline telephone records (an old staple for generating sampling frames) inadequate, requiring other sources for listings of residences. In many developing countries there are no complete listings of residents or addresses, and it may be necessary to visit every residence in an area to establish a reliable sampling

frame.

Proceeding from our definition of the sampling unit and a sampling frame, we specify the rules by which observations are drawn. These rules define a *sampling strategy*. Implementation of the sampling strategy produces an *outcome*, which is the set of observations that is drawn from the population via the sampling process. The most widely analyzed class of strategies are termed *probability samples*. Such strategies have the following properties (adapted from Cochran, 1977):

1. The possible outcomes of the sampling process are defined.
2. A probability (or probability density) can be assigned to each possible outcome for the given process.
3. One of the possible outcomes is selected at random with these assigned probabilities.
4. There is a well-defined estimator that uses the sample and yields a unique estimate for each possible outcome.

The advantages of probability samples are that they can be used to obtain statistically valid estimates of population characteristics and they allow calculations of the magnitudes of sampling errors. *Sampling errors* are the error in an estimate of a population characteristic that is based on a sample rather than a census. Non-probability sampling strategies are also used, which is any sampling method in which the probability of any population element's inclusion in the sample is unknown. An example of a non-probability sample is a convenience sample, such as when respondents are recruited by targeting email lists. We focus here on probability sampling.

Sampling strategies can be implemented either *with or without replacement*. In sampling without replacement, once a sampling unit is drawn, it cannot be drawn again. In contrast, when sampling with replacement is used, a sampling unit, once drawn, is "returned" to the population and can be drawn again. This distinction is only relevant when the population is assumed to be finite because for an infinite population the distribution of characteristics within the pool of sampling units is unaffected by the removal of any finite number of them.

We now employ a simple example to illustrate the sampling terminology described above and also to demonstrate the relationship between the sampling strategy and the properties of estimators. Suppose we have a well-defined population of N households and we are interested in estimating the mean household income Y. We use each household as a sampling unit. Our

sampling frame therefore consists of all households in the population, and we define the sample to consist of $N_s < N$ households drawn without replacement from the population. This would imply that the number of possible outcomes of the sampling process is.

$$N^* = \frac{N!}{N_s!(N - N_s)!}, \tag{2.9}$$

where each outcome is a different set of N_s households.

We now define our sampling strategy to consist of drawing N_s households *at random* from the entire population so that each sampling unit has equal probability of being drawn. In this strategy, each outcome has equal probability $1/N^*$ of being drawn. We call this sampling method *simple random sampling*.

Finally, suppose we use the sample average as an estimator of the mean household income:

$$\widehat{Y}^R = \frac{1}{N_s} \sum_{n=1}^{N_s} y_n, \tag{2.10}$$

where y_n is the income of the n^{th} sampled household. The superscript R denotes that this is the estimator for a random sample. Later we introduce a different estimator and use a different superscript. Denoting σ^2 as the variance of the distribution of y, the variance of this estimator is then

$$\text{Var}[\widehat{Y}^R] = \frac{\sigma^2}{N_s}. \tag{2.11}$$

Whether or not the population N is considered finite or infinite, it can be shown that this estimator has all of the desirable properties of an estimator. First, it is unbiased, meaning $E[\widehat{Y}^R] = Y$. Second, it is efficient, meaning the variance (2.11) is the minimum amongst all unbiased estimators. See Chapter 1 for an introduction to these terms.

We can easily define sampling strategies for which $\widehat{Y}^R$ as defined in equation (2.10) has only some or even none of these desirable properties. To see this, suppose we divide our population into two subpopulations, each consisting of half of the N households. For the purposes of exposition, define the subpopulations as consisting of central city and suburban residents, and let Y_1 and Y_2 denote their respective mean incomes. Assume the groups have different mean income, that is, $Y_1 \neq Y_2$. We now redefine our sampling strategy so that for each suburban resident in the sample we draw at random two central city residents. Such a strategy is called *stratified sampling*. Under this sampling strategy and using the estimator $\widehat{Y}^R$ as defined in (2.10) the

sample average income is written as

$$\widehat{Y}^R = \frac{1}{N_s}\left(\sum_{n=1}^{N_s/3} y_n + \sum_{n=(N_s/3)+1}^{N_s} y_n\right), \tag{2.12}$$

where the data are assumed to be ordered such that the first $N_s/3$ observations are suburban residents. In this case

$$E[\widehat{Y}^R] = \frac{1}{N_s}\left(\frac{N_s}{3}Y_2 + \frac{2N_s}{3}Y_1\right) = \frac{Y_2 + 2Y_1}{3} \neq Y. \tag{2.13}$$

Thus, as long as $Y_2 \neq Y_1$, under the new sampling strategy the sample average is biased, whereas under the simple random sampling strategy it is unbiased. In this simple example it should be obvious that one way to obtain an unbiased estimator for the new sampling strategy is to "reweight" the observations to reflect the oversampling of central city residents. Thus we could use

$$\widehat{Y}^S = \frac{1}{N_s}\left(\frac{3}{2}\sum_{n=1}^{N_s/3} y_n + \frac{3}{4}\sum_{n=(N_s/3)+1}^{N_s} y_n\right), \tag{2.14}$$

where the S superscript denotes this is the estimator for a stratified sample. While the estimator $\widehat{Y}^R$ was biased when applied to the stratified sample, the estimator $\widehat{Y}^S$ is unbiased:

$$E[\widehat{Y}^S] = \frac{1}{N_s}\left(\frac{3}{2}\frac{N_s}{3}Y_2 + \frac{3}{4}\frac{2N_s}{3}Y_1\right) = \frac{Y_2}{2} + \frac{Y_1}{2} = Y. \tag{2.15}$$

This example has demonstrated that the properties of estimators are dependent on the way in which the sample is drawn.

2.3.2 Overview of Common Sampling Strategies

In this section we review some of the common sampling strategies used for estimating population characteristics such as population totals, means, or variances. Although most of these strategies are also useful for obtaining data to estimate the parameters of discrete choice models, we hold off a detailed discussion of the implications on estimation until later in the text. This review is not intended to be rigorous or complete. Readers interested in a far more comprehensive treatment of standard sampling theory are referred to Deming (1960), Cochran (1977), Fink (1995), or Lohr (1999).

Simple Random Sampling

As defined above, simple random sampling refers to the case where each sample outcome has an equal probability of being drawn. Equivalently, each individual sampling unit has an equal probability of being drawn.

Stratified Random Sampling

In addition to simple random sampling, the most widely used design is some form of *stratified random sampling*. In this approach, the sampling frame is divided into G mutually exclusive and collectively exhaustive groups, each called a stratum. Simple random sampling is then used to sample N_{sg} observations from each stratum, and therefore $N_s = \sum_{g=1}^{G} N_{sg}$.

Our example of a nonrandom sampling strategy in section 2.3.1 was a stratified random sample in which $G = 2$ (the central city and suburban residents) and $N_{s1} = 2N_s/3$ and $N_{s2} = N_s/3$. Geographic stratification, however, is only one basis for partitioning the population. Other possibilities include size of firm, housing type, number of automobiles owned, age, or sex.

As with all forms of nonrandom sample design, stratification is done for a number of not necessarily mutually exclusive reasons. First, it may be useful to know the characteristics of certain subpopulations as well as those of the whole population. For example, we may want to know something about the travel behavior of elderly or immigrant populations. In cases where these groups of interest would be a very small proportion of a random sample, we may intentionally "oversample" households with such individuals in order to obtain more data on their travel choices. Second, we may find it less expensive to obtain a stratified random sample when compared to a simple random sample of equal size. For example, per household sampling in high density neighborhoods is cheaper than sampling in rural areas. A third motivation for sampling nonrandomly is to increase the efficiency with which certain characteristics of the population are estimated.

To see how a stratified sample can improve the efficiency of estimation, let us again assume we are interested in estimating the mean income Y of a population. However, suppose it is in an urban area that has areas with strong economic stratification (i.e., low variance of income within a neighborhood) and areas that are more diverse (i.e., higher variance of income within a neighborhood). Suppose we define G strata based on these geographic boundaries, each of which has a different unknown mean and variance denoted as Y_g and σ_g^2, respectively. Suppose we know the proportion of the population in each stratum and denote it by W_g.

Following the line of reasoning motivating the weighted estimator (2.14)

in the previous section, one unbiased estimator for the population mean can be constructed as follows. First, we estimate the within-stratum means as

$$\widehat{Y}_g = \frac{1}{N_{sg}} \sum_{n=1}^{N_{sg}} y_{gn}, \quad g = 1, \ldots, G, \tag{2.16}$$

where y_{gn} is the n^{th} observation of y from stratum g. Then we estimate the population mean as the weighted average of the strata estimates. In other words, we define an estimator $\widehat{Y}^S$ of Y as foIlows:

$$\widehat{Y}^S = \sum_{g=1}^{G} \widehat{Y}_g W_g. \tag{2.17}$$

The reader should note that $\widehat{Y}^S$ is not fully efficient because it weights high variance observations the same as those with low variance. We use it here primarily for expository purposes. The expected value of this estimator is equation (2.17),

$$E[\widehat{Y}^S] = Y, \tag{2.18}$$

which indicates that it is an unbiased estimator. The variance of the estimator is

$$\text{Var}[\widehat{Y}^S] = \sum_{g=1}^{G} W_g^2 \left(\frac{\sigma_g^2}{N_{sg}} \right). \tag{2.19}$$

Since the variance of $\widehat{Y}^S$ depends on values of N_{sg}, $g = 1, \ldots, G$, we can ask, for a given total sample size N_s, what strata sample sizes minimize the variance of this estimator of the population mean? To solve this, we find $(N_{s1}, N_{s2}, \ldots, N_{sG})$ which are a solution to

$$\min_{N_{s1}, \ldots, N_{sG}} \sum_{g=1}^{G} W_g^2 \left(\frac{\sigma_g^2}{N_{sg}} \right) \tag{2.20}$$

$$\text{s.t.} \sum_{g=1}^{G} N_{sg} = N_s.$$

This problem can be solved by setting up a Lagrangian and solving for the first order conditions, which leads to:

$$N_{sg} = \left(\frac{W_g \sigma_g}{\sum_{g'=1}^{G} W_{g'} \sigma_{g'}} \right) N_s. \tag{2.21}$$

This result suggests that by altering the sample sizes for each of the strata, we can alter the variance of the estimated mean. Moreover there exists a stratified random sampling strategy yielding an estimator of Y with minimum variance, in which we sample each stratum proportionately to the product of its standard deviation and the fraction of the population it represents. For example, when the strata are of equal size, one should "oversample" those that are heterogeneous as this reduces the variance of the estimated mean.

We can also explore the relative efficiency of stratified and simple random sampling. To calculate the variance for the stratified sample, we evaluate equation (2.19) at the optimal sample sizes (2.21). This results in the variance

$$\operatorname{Var}[\widehat{Y}^S] = \frac{\left(\sum_{g=1}^{G} W_g \sigma_g\right)^2}{N_s}. \tag{2.22}$$

The variance of the estimator using a simple random sample is given by (2.11) where the variance of the distribution of y (σ^2) is a function of the within strata variances. This results in the variance

$$\operatorname{Var}[\widehat{Y}^R] = \frac{\sigma^2}{N_s} = \frac{\sum_{g=1}^{G} W_g \sigma_g^2}{N_s}. \tag{2.23}$$

Taking the ratio of (2.23) and (2.22), we obtain

$$\frac{\operatorname{Var}[\widehat{Y}^R]}{\operatorname{Var}[\widehat{Y}^S]} = \frac{\sum_{g=1}^{G} W_g \sigma_g^2}{\left(\sum_{g=1}^{G} W_g \sigma_g\right)^2}. \tag{2.24}$$

The numerator is the mean of the variances (i.e., standard deviations squared) across all strata and the denominator is the square of the mean across strata of the standard deviations. Thus by Jensen's inequality (see (B.113) with $f(x) = x^2$ convex) we get

$$\operatorname{Var}[\widehat{Y}^R] \geq \operatorname{Var}[\widehat{Y}^S]. \tag{2.25}$$

The two sampling strategies will yield estimators with equal variance when all the strata have equal variance; otherwise, the optimal stratified sample will be better than a random one.

One can also show that using stratified random sampling with poorly chosen strata sample sizes can actually produce samples for which the estimated mean has higher variance than the estimate under simple random sampling. For example, suppose there are two strata, each representing half of the entire population with variances σ_1^2 and σ_2^2 where $\sigma_2^2 = 9\sigma_1^2$; that is, the second

stratum has nine times higher variance than the first. The optimal stratified sample is to select 25% of the total sample from the first stratum and 75% from the other. By equation (2.22) this yields a variance of

$$\frac{(\sigma_1/2 + 3\sigma_1/2)^2}{N_s} = \frac{4\sigma_1^2}{N_s}. \tag{2.26}$$

By equation (2.23) a simple random sample would have variance

$$\frac{\sigma_1^2/2 + 9\sigma_1^2/2}{N_s} = \frac{5\sigma_1^2}{N_s}. \tag{2.27}$$

As expected, the variance of the optimal stratified sample outperforms a random sample. However, if we chose to draw 3/4 of our observations from group 1 and 1/4 of our observations from group 2, the variance of our stratified sample estimator (2.17) would be (by equation (2.19))

$$\frac{(1/2)^2\sigma_1^2}{3N_s/4} + \frac{(1/2)^2 9\sigma_1^2}{N_s/4} = \frac{9.33\sigma_1^2}{N_s}, \tag{2.28}$$

which is greater than the variance of $\widehat{Y}^R$ from the simple random sample.

While the above discussion focused on the case where the variance of the characteristic of interest varies across strata, another issue is when the per observation unit cost of sampling varies across strata. The optimization problem of (2.20) can be extended to reflect varying costs. Suppose that the costs of any one observation taken from stratum g is c_g and that the fixed costs of conducting the sample is c_s. What is the best allocation of a total budget of B? In this case, the objective function remains the same as in (2.20), that is the goal is to minimize the variance of the estimator. However, the question is not how to allocate a specific number of observations N_s but how to allocate a particular budget B, which is reflected in the constraint of the optimization problem as follows:

$$\min \sum_{g=1}^{G} W_g^2 \left(\frac{\sigma_g^2}{N_{sg}} \right) \tag{2.29}$$

$$\text{s.t. } c_s + \sum_{g=1}^{G} N_{sg} c_g = B.$$

Again, the Lagrange and first order conditions can be used to derive the following optimal allocation:

$$\frac{N_{sg}}{N_s} = \frac{W_g \sigma_g / \sqrt{c_g}}{\sum_{g'=1}^{G} W_{g'} \sigma_{g'} / \sqrt{c_{g'}}} \quad g = 1, \ldots, G \tag{2.30}$$

and optimal total sample size

$$\text{and } N_s = (B - c_s)\frac{\sum_{g=1}^{G} W_g \sigma_g / \sqrt{c_g}}{\sum_{g=1}^{G} W_g \sigma_g \sqrt{c_g}}. \tag{2.31}$$

Thus the optimal fraction of the sample from any stratum is proportional to the standard deviation of the characteristic being measured and the stratum size and inversely proportional to square root of the cost per sample.

More detailed discussions of stratified random sampling carry this analysis still further by exploring the properties of stratified random sample designs when ratios are estimated and when different non optimal strata sample sizes are used. Readers interested in such topics are referred to any standard text on sample design. For our purposes we highlight only the central message of this literature: *Stratification, if done appropriately, can potentially reduce sampling costs and increase the efficiency of estimators; however, if done inappropriately, it can make estimators worse than if simple random sampling is used.*

Stratified Random Sampling for Discrete Choice Analysis

When using stratified sampling for discrete choice analysis datasets, there is a further caveat that is of importance. The population distribution is defined along both the choice dimension (the endogenous variable) and the explanatory variables (the socio-economic characteristics and attributes of the alternatives), and therefore the definition of the strata can involve both the dependent, or explained, or endogenous variable i and the independent, or explanatory, or exogenous variables x. This gives rise to a rich class of sampling strategies which Manski and McFadden (1981a) term *general stratified sampling.*

Formally a general stratified sample is drawn as follows:

Step 1: Partition the population into G collectively exhaustive strata, each defined in terms of combinations of choices and attributes.

Step 2: Select sampling fractions $H_1, H_2, \ldots, H_G$ as the fractions of the sample to be drawn from the G strata. Then select the total sample size N_s.

Step 3: Draw $N_{sg} = H_g N_s$ observations at random from stratum g for all $g = 1, \ldots, G$.

Step 4: For each observation n, observe their choice (i_n) and attributes (x_n).

Figure 2.6 provides an illustration of the possible dimensions for general stratified sampling in the context of a three-mode choice model. Each of the rows corresponds to possible choices and each of the columns corresponds to a range of values of a variable in a hypothetical vector x. Any single observation would be a pair consisting of i and x.

The class of stratified sampling rules, in general, and the most relevant special cases, in particular, offer an enormous range of sample design possibilities to the analyst. In considering these options, it is important to distinguish what aspects of the sampling process the analyst does and does not control. What this person *does* control is the stratification and the number of decision makers sampled, N_{sg}. What he or she *does not* control are the identities of the decision makers then drawn. These drawings are to be independent and at random.

The simplest possible stratification would be to define only one stratum consisting of the entire population, and the corresponding sample would be drawn randomly from this single stratum. This is by definition a simple random sample. Three more interesting stratifications are often employed in discrete choice analysis: *exogenous sampling*, endogenous or *choice-based sampling*, and *enriched samples*. These are each defined next.

Exogenous Sampling: In an exogenous sample we define the strata by segmenting only on the exogenous variables that is, attributes x and not on the actual choices. In other words, we divide the possible attribute vectors into collectively exhaustive sets X_1, X_2,..., X_G, and the pair (i, x) belongs to stratum g if and only if $x \in X_g$. This is illustrated by Figure 2.7, where the strata are defined by the travel time. Within each stratum, individuals are drawn irrespectively of their mode of transportation i.

Choice-Based Sampling: Any strategy where the endogenous variable (the choice) affects the stratum where the individual belongs is called an *endogenous sampling*, also known as *choice-based sampling*. In a choice-based sample we partition the full choice set $\mathcal{C}$ into collectively exhaustive subsets $\mathcal{C}_1$,...,$\mathcal{C}_G$. The pair (i, x) then belongs to stratum g if $i \in \mathcal{C}_g$. It is typically useful when we have access to a customer database, or when we analyze products with a low market share. Pure choice-based sample is the case where each choice in $\mathcal{C}$ corresponds to a separate stratum. In this case $G = J$, the size of the choice set. This is illustrated by Figure 2.8, where the strata are defined by choice. Within each stratum, individuals are drawn irrespectively of the travel time.

In our mode choice example, the case of stratification by modal groups corresponds to a choice-based sample. For example, on-board transit surveys and roadside interviews are both choice-based samples for modal choice analysis.

It is important to stress that whether a sample is choice-based or exogenous depends on the aspect of model under study. For example, stratification by residence is exogenous if we are analyzing mode choice, but choice-based if we are analyzing households' residential location decisions.

Enriched Sampling: Cosslett (1981a) defines *enriched sampling* as the pooling of exogenously stratified samples with one or more choice-based samples. For example, a simple random sample might be merged with a transit on-board survey to analyze mode choice.

This type of sample design is often an attractive way to increase the number of observations choosing alternatives with low aggregate population shares. For example in the United States, even in a large random home interview survey we might find that very few people chose to bicycle to work. If we wanted to increase the number of bicycle riders in our sample we could "enrich" the random sample with a special survey of bicycle users.

Cluster Sampling

In our examples of sampling we have so far assumed that the sampling unit is also the basic unit of interest in the analysis. However, in many cases each sampling unit actually consists of relevant subunits. For example, if we are interested in the trips made by individuals, each household sampled is a group, or *cluster*, of individuals. A strategy that uses sampling units that have subunits of interest is called *cluster sampling*.

The most common reason for using cluster sampling is its lower cost per observation. It is, for example, generally easier to sample five individuals from the same household than five individuals from different households. Similarly, it is easier to sample all the shipments of a single firm than the same number of shipments each from a different firm.

In general, the reduced cost of sampling in clusters comes at the expense of increased variance in the estimators of population characteristics. This is because individuals within the same cluster will tend to have characteristics that are positively correlated. For example, if one shipment of a particular commodity used by a firm is large (as compared to the mean size for the entire population), then subsequent shipments of the same firm will tend to be large also.

To see why this affects the variance of the estimated mean, suppose we have two different samples of N_s observations. In the first case the observations are taken randomly from the population (i.e., by simple random sampling). If the population variance is σ^2, then the variance of the estimated mean is σ^2/N_s. Now consider the case where the N_s observations are from M equal-size clusters in which there is a correlation ρ between any

two values from the same cluster. For simplicity assume N_s is a multiple of M, so that N_s/M is an integer. The variance of the mean estimated by the sample average will be $\sigma^2(1 + [M-1]\rho)/N_s$ which, for $\rho > 0$, exceeds σ^2/N_s (Cochran, 1977). Of course, if ρ is negative, the variance of using a cluster sample will actually be less than for a random one.

Cluster sampling strategies offer a greater number of options. For example, we can sample clusters and then sample subunits within clusters. This procedure is called *two-stage cluster sampling* or *subsampling* (as opposed to one-stage cluster sampling in which all subunits in the sampling unit are observed). The distinction between cluster and stratified sampling is somewhat subtle. In stratified sampling the groups from which random samples are taken are not themselves samples; in cluster sampling we first draw a subset of the groups.

With cluster sampling we can also use different sampling strategies for choosing the clusters. For example, if firms are defined as clusters, we can select firms entirely at random, or we can select them with probability proportionate to some measure of their size. These two strategies may have very different costs and result in estimators with different variances. Moreover the estimators that are consistent and unbiased for one strategy may not be for the other.

There are also different estimators appropriate to different forms of cluster sampling. In cases where the within-cluster variance differs from cluster to cluster, it is possible to gain efficiency by appropriately correcting for the differing within-cluster variances. To complicate matters further, we must typically estimate the within-cluster variances to apply some of the estimators of the mean. Again a wide range of methods has been developed. Readers are referred to Cochran (1977) for a full treatment of these issues.

Double Sampling

To design a sample effectively, it is often useful for an analyst to have some information about the population. For example, to design an effective stratified sample, we must generally know the relative sizes and variances of all the strata. This data may not be available from existing sources, so a reasonable course of action would be to conduct a simple random sample to estimate the strata shares and variances and then to use that information to design a second stratified sample. We call any strategy in which one sample is used to design a second a *double sample*. The concept obviously extends to more than two levels and is often termed *multistage sampling* or *sequential sampling*.

The literature on sequential sampling falls into two broad classes. In

what might be termed the *classical sample design analysis*, estimators of relevant population characteristics made from early sampling stages and used to design later stages are treated in the sample design as fixed, not random. This greatly simplifies the analysis but is obviously an approximation that may be more or less valid depending on the size of the early sample. In *Bayesian sequential sampling* we explicitly treat the population estimates from each sampling stage as random variables and solve for the next stage sample designs which are optimal based on an expected value criterion. This unfortunately complicates the analysis enormously, and analytic solutions are only available for limited classes of distributions. Readers interested in the subject are referred to De Groot (1970), Berger (1985) or Wetherill and Glazebrook (1986).

Double sampling gives the analyst an additional dimension of control in the sample design. This is particularly valuable when very little is known a priori about the population of interest. It also, however, imposes additional costs per observation because of start-up costs associated with each of the samples. In addition the analyst must be careful to ensure that structural changes (e.g., induced by an energy crisis or an economic recession) in the characteristics being measured have not occurred in the interval between the two samples or, if changes have occurred, that they are appropriately accounted for. For example, income can be adjusted by using published data on inflation or real wages. However, many structural changes may occur and be totally unknown to the analyst.

Systematic Sampling

Although, in practice, most sampling strategies are either combinations or extensions of the ones just discussed, some relatively minor variants do exist. One of these is *systematic sampling*, in which elements are drawn from the sampling frame by deterministic rather than random rules. For example, interviewers might be told to interview every tenth house on a block or to interview all employees with social security numbers ending with a one. As long as the order of observations within the sampling frame is random, there is no real difference between a systematic sample and a simple random one. The case where they do differ is when for some reason there is nonzero correlation ρ between pairs of sampling units in the same systematic sample. In such a case one can show that for the sample average as an estimator of the mean Y,

$$\mathrm{Var}[\widehat{Y}] = \frac{\sigma^2}{N_s}[1 + (N_s - 1)\rho], \tag{2.32}$$

where σ^2 is the variance of an individual observation. In most cases of practical interest, ρ can be assumed to be zero. One particular exception is where samples are being drawn over time, and observations that are equal time intervals apart are likely to be correlated.

2.3.3 Sample size calculations

In discussing sampling strategies above, we referenced how the sampling strategy could be used to make a given sample more efficient by intelligently allocating observations among strata based on heterogeneity and cost of observation. The "efficiency" in this case is measured in terms of minimizing sampling error. For sample size calculations, we turn this problem around, asking how many observations are needed to obtain a desired sampling error within a certain level of confidence. We discuss the sample size calculation for four different estimation cases: population means, proportions, regression models, and discrete choice analysis.

Sample size calculation for estimating a population mean

We begin by focusing on the objective of estimating a population mean μ and population variance σ^2 of some characteristic X.

Given a sample size N_s, the sample will have N_s observations of the variable X, which we denote as $x_1, x_2, ..., x_n, ..., x_{N_s}$. The sample average:

$$\overline{X} = \frac{1}{N_s} \sum_{n=1}^{N_s} x_n \tag{2.33}$$

is an estimator of μ. The sample variance:

$$s^2 = \frac{1}{N_s - 1} \sum_{n=1}^{N_s} (x_n - \overline{X})^2 \tag{2.34}$$

is an estimator of σ^2. The sample standard deviation is $s = \sqrt{s^2}$.

The variability of the estimated mean $\overline{X}$ from sample to sample depends on both the variability of X in the population (σ^2) and the sample size (N_s). The sampling distribution represents the variability of a sample statistic among samples drawn according to the same sampling strategy. By the central limit theorem, the sampling distribution of $\overline{X}$ is normally distributed with mean equal to the population mean μ and variance σ^2/N_s, written as

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{N_s}\right). \tag{2.35}$$

This result is independent of the form of the distribution of X in the population. The variable is transformed into a standard normal distribution as follows

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{N_s}} \sim N(0,1). \tag{2.36}$$

From this we can calculate confidence intervals. Denote the level of significance as α and the level of confidence as $1-\alpha$. Denote $Z_{(\alpha/2)}$ as the critical value for a given level of significance, meaning that $1-\alpha$ percent of the standard normal distribution lie between $\pm Z_{(\alpha/2)}$. The $1-\alpha$ confidence interval for the standardized variable is then

$$-Z_{(\alpha/2)} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{N_s}} \leq Z_{(\alpha/2)}. \tag{2.37}$$

And we can rearrange terms to derive the confidence interval for the population characteristic of interest μ

$$\overline{X} - Z_{(\alpha/2)}\frac{\sigma}{\sqrt{N_s}} \leq \mu \leq \overline{X} + Z_{(\alpha/2)}\frac{\sigma}{\sqrt{N_s}}. \tag{2.38}$$

Recall that $Z_{(\alpha/2)} = 1.96$ for a 95% confidence interval and $Z_{(\alpha/2)} = 1.65$ for a 90% confidence interval.

A sample size determination involves determining the sample size N_s necessary to achieve an allowable error, denote as d, within a given level of confidence. This is directly derived from the confidence interval equation above: $d = Z_{(\alpha/2)}\sigma/\sqrt{N_s}$ and solving for N_s leads to the necessary sample size

$$N_s = \frac{{Z_{(\alpha/2)}}^2\sigma^2}{d^2} \tag{2.39}$$

This sample size will result in an estimate of $\overline{X}$ that is within d units of the true value of μ with $(1-\alpha)100\%$ confidence. Note that in order to calculate the necessary sample size, one needs to know the population parameter σ^2. While this is not known, educated guesses can be made based on prior information from other samples or populations. Because of this, sometimes the coefficient of variation σ/μ is used, because it is unit free and easier to develop an educated guess. The sample size calculation that uses coefficient of variation specifies d as the percent allowable error and is as follows

$$N_s = \frac{{Z_{(\alpha/2)}}^2(\sigma/\mu)^2}{d^2} \tag{2.40}$$

Sample size calculation for estimating a proportion

The above analysis for estimating population means can be applied to a situation closer to discrete choice analysis, which is that of estimating a proportion of a population with a specific characteristic. For example, in Chapter 1, we were interested in estimating the share of mobile phone owners who have smartphones. This share is simply a population mean of the bernoulli instances of whether the mobile phone is a smartphone or not. Therefore, we can apply (2.39). In this case the population variance σ^2 is equal to $\pi(1-\pi)$, a characteristic of the bernoulli distribution. Knowing the population variance requires knowing the parameter of interest π. However, we can make the most conservative guess by assuming $\pi = 0.5$, which results in the maximum possible variance. Plugging into (2.39), the sample size required to be within 5 percentage points with 95% confidence is equal to $1.96^2(0.5)(1-0.5)/(0.05)^2 = 384$. If you are willing to stretch the range to 10 percentage points, then the necessary sample size drops to 96. Alternatively, if you are willing to decrease the confidence level to 90%, then the sample size drops to 274.

Sample size calculation for a regression model

Sample size calculation for estimating a discrete choice model

2.3.4 Errors other than sampling errors

Sampling error was emphasized above in the analysis of estimators and calculations of sample size. This is the error in an estimate of a population characteristic that is based on a sample rather than a census. With probability samples, the size of this error can be estimated, sample size calculations can be performed, and the sample can be adjusted to minimize the error. There are other errors that arise due to sampling, which are more difficult to precisely account for. One potential source of error is non-response bias, which is the error due to the inability to elicit information from some respondents in a sample; often caused by refusals. Another is response bias, which is the error due to systematic distortion of survey responses for reasons such as social desirability, prestige seeking, and post-purchase justification. The treatment for these is to first attempt to minimize the error via careful survey design. And then use external data sources to measure the degree of the error. For example, for non-response bias, key demographics in the sample can be compared with accurate distributions of demographics available from other sources such as a census. Once the degree of non-response is known, one can correct the non-response bias by reweighing the sample ob-

servations using a post-stratification method. Iterative proportional fitting (IPF), which estimates a joint distribution from known marginal distributions, and Gibbs sampling, that draws from the joint distribution based on known marginals, are useful technique for determining the appropriate individual weights. These procedure are described in Chapter 10.

2.3.5 Lessons from sampling theory for discrete choice analysis

While the different sampling strategies we have discussed were originally developed for estimating simple population characteristics, the sampling problem for discrete choice analysis is conceptually no different. Instead of seeking to estimate population means or variances, we are trying to infer the vector of parameters of the choice model or the distribution of the characteristics of the population (e.g., the shares). The fact that we are attempting to infer the parameters of a nonlinear choice model significantly complicates the sample design problem. However, many of the basic lessons are transferable. Here we discuss these general lessons, and in Chapter 11 we go into further detail as to how the discrete choice estimation is impacted under different sampling strategies.

Data collected for discrete choice analysis are generally collected using one of the sampling strategies described above.

sample more where you know least (higher variance) or where exogenous versus endogenous stratification. As we return to later in this chapter as well as later in the book, the fact that we are attempting to infer the parameters of nonlinear choice models significantly complicates the sample design problem.

One of the most important results is that the optimum sample for estimating the parameters of a discrete choice model will depend on the values of the unknown parameters. This result is in direct contrast to the results for optimal sample design in the typical linear regression model.

To elaborate on this distinction, consider first the standard regression case. For simplicity we consider as an example the simple linear model. Let

$$y_n = \alpha + \beta x_n + \xi_n, \tag{2.41}$$

where α, β and x are scalars and ξ is an independent disturbance with mean 0, variance σ^2. The least squares estimator of β is

$$\widehat{\beta} = \frac{\sum_{n=1}^{N_s}(x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^{N_s}(x_n - \bar{x})^2}, \tag{2.42}$$

where $\bar{x}$ and $\bar{y}$ denote the sample averages for x and y, respectively. The variance of $\widehat{\beta}$ for a sample stratified on the basis of the independent variable x is given by

$$\mathrm{Var}(\widehat{\beta}) = \frac{\sigma^2}{\sum_{n=1}^{N_s}(x_n - \bar{x})^2}. \tag{2.43}$$

In this case the variance of $\widehat{\beta}$ is not affected by the value of the parameter β.

This property leads directly to the standard classical result that in choosing a stratification one should oversample "extreme" values of x, that is, choose an equal number of very small and very large value of x_n thereby maximizing $\sum_{n=1}^{N_s}(x_n - \bar{x})^2$.

Now consider a simple binary logit model, where

$$P(i|x_n) = \frac{1}{1 + e^{-\beta x_n}}, \tag{2.44}$$

and β and x are scalars and x_n is defined as $x_{in} - x_{jn}$.

If the sample is an exogenous one and maximum likelihood estimation is used, then by the Cramér-Rao bound the asymptotic variance of $\widehat{\beta}$ is

$$\mathrm{Var}[\widehat{\beta}] = \frac{-1}{E[\partial^2 \mathcal{L}/\partial \beta^2]} = \frac{1}{N_s}\left(E\left[\frac{x_n^2 e^{\beta x_n}}{(1 + e^{\beta x_n})^2}\right]\right)^{-1}. \tag{2.45}$$

Now note that if we wish to minimize the asymptotic variance of $\widehat{\beta}$ for a given sample size N_s, we should maximize the term inside the expected value. If we were to adopt an analog to the standard linear model, we might try to do this by making x_{in} and x_{jn} as different as possible, thus moving x_n as close to $\pm\infty$ as practical. However, as long as $\beta \neq 0$,

$$\lim_{x \to \pm\infty} \frac{x_n^2 e^{\beta x_n}}{(1 + e^{\beta x_n})^2} = 0, \tag{2.46}$$

which is its minimal, not maximal, value. Thus not only is the sampling rule analogous to the one for the regression model not optimal when $\beta \neq 0$, it is the worst possible rule.

Still more troubling is that for the case where $\beta = 0$, this result changes entirely:

$$\frac{x_n^2 e^{\beta x_n}}{(1 + e^{\beta x_n})^2} = \frac{x_n^2}{4} \tag{2.47}$$

which is now maximized for large values of x_n^2. *Thus the same exogenous stratified sampling rule that is the worst one for some values of the parameters is the best one for another value.*

This result generalizes as follows: There is no single sample design for discrete choice analysis that is unambiguously optimal for all values of the parameters. Rather, whether a sample design is good or bad (in the classical sense) depends on the unknown parameter values.

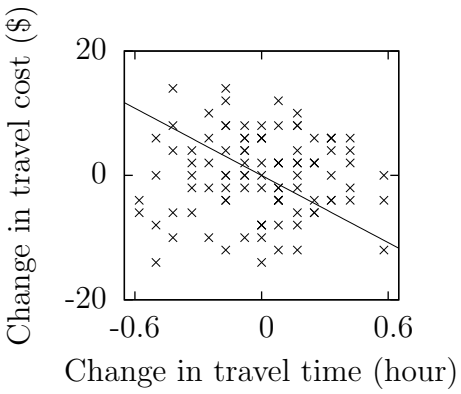
2.4 Summary

In this chapter we have reviewed issues related to the data used for discrete choice analysis, first describing the components and then discussing advantages and disadvantages of the two primary types of discrete choice data: revealed preferences (actual market behavior whether reported by the decision-maker or otherwise observed) and stated preferences (responses to hypothetical choice environments). We provided an overview of one of the key issues of stated preference surveys, which is the issue of experimental design, that is how to populate the attribute values in hypothetical choices presented to people in the survey.

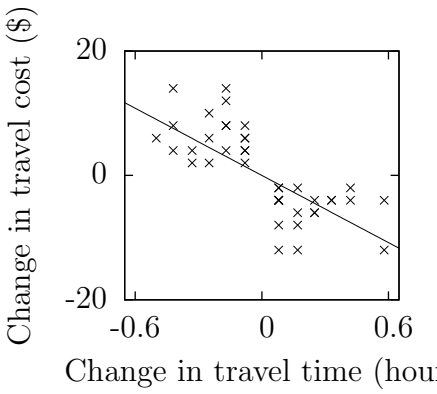
We described the stages of the sampling process, including defining a population, identifying a sampling frame, selecting a sampling strategy, determining the sample size, and collecting the data. We discussed issues of how to sample respondents from a population and why different sampling strategies may be employed. We demonstrated how the sampling strategy impacts the properties of estimators. Careful design can increase the efficiency of population estimates (for example by oversampling strata with more heterogeneity) and can decrease the cost (for example, by oversampling strata with lower costs). However, poor sampling strategies can make one worse off than using a simple random sample. We also discussed the issue of sample size. For much of the sampling discussion, we discussed the theory in terms of calculating basic population characteristics from a population. Therefore, we ended with a discussion of how the basic theory applies to discrete choice analysis. As is often the case in discrete choice analysis, the straightforward analytical solutions used for descriptive statistics and regression are not applicable. Nonetheless, the basic strategies for the linear cases provide insights into the discrete choice context.

We focussed in this chapter on technical issues specific to surveys used for choice modeling. Typically a survey will include questions beyond the choice question(s), including queries regarding socio-demographics, attitudes, and situational constraints. The broader issue of designing such surveys is considered outside of the scope of this book. We refer the reader to a number of other texts that provide practical guidance on conducting survey research: Converse and Presser (1986), Fowler (1995), Rea and Parker (2005), Fowler

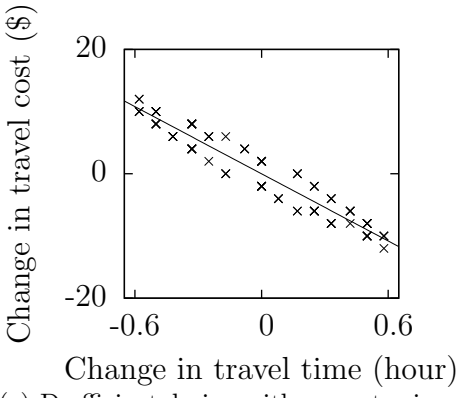
(2008), and Babbie (2009).



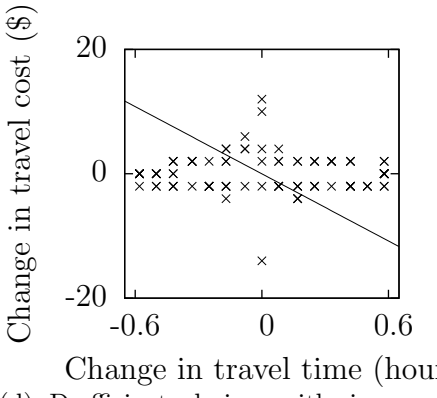
(a) Orthogonal design



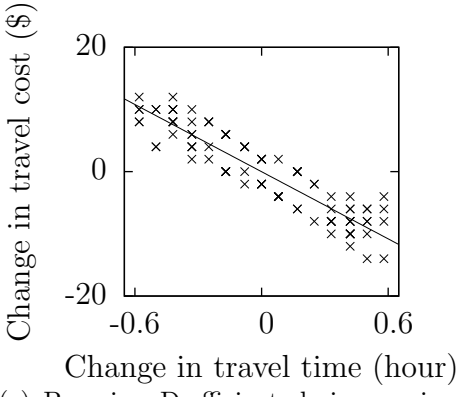
(b) Cleaned orthogonal design: choice sets with dominant alternative removed



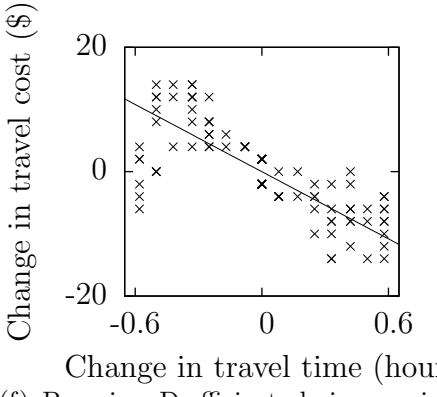
(c) D-efficient design with correct prior: $\beta_c = -1$, $\beta_t = -18$, VOT = \$18/hour



(d) D-efficient design with incorrect prior: $\beta_c = -1$, $\beta_t = 0$, VOT = \$0/hour



(e) Bayesian D-efficient design: prior VOT ~ N(\$18/hour, \$6/hour)



(f) Bayesian D-efficient design: prior VOT ~ N(\$18/hour, \$12/hour)

Figure 2.4: Experimental designs generated using different methods. Points indicate trade-offs of choice sets. Line is the *assumed* true value of the slope to be estimated (value of time "VOT" = \$18/hour).

Figure to be included

Figure 2.5:

		Travel time by car (min)			Household automobile ownership			Annual household income ($)	
Chosen mode (i)		$\leq$ 15	]15-30]	$>$ 30	0	1	2+	$\leq$ 15K	$>$ 15K
	Drive alone								
	Carpooling								
	Transit								

Figure 2.6: Illustration of dimensions for general stratified sampling

		Travel time by car (min)		
Choice i		$\leq$ 15	]15-30]	$>$ 30
	Drive alone			
	Carpooling			
	Transit			

Figure 2.7: Example of strata definition for exogenous sampling

		Travel time by car (min)		
Choice i		$\leq$ 15	]15-30]	$>$ 30
	Drive alone			
	Carpooling			
	Transit			

Figure 2.8: Example of strata definition for pure choice-based sampling

Chapter 3

Choice theories

"*A man has free choice to the extent that he is rational.*", Thomas of Aquin.

Contents

3.1 Objectives

The objectives of this chapter are to describe the principles of individual choice theories that are useful in the formulation of empirical discrete choice models. Our emphasis is on making operational quantitative models of discrete choice behavior. To this end, in this chapter we describe the choice process and introduce terminology, provide relevant background information from microeconomic theory and the behavioral sciences, and introduce the random utility model.

3.2 Introduction

The discrete choice models presented in this book are employed for a variety of reasons. At a most fundamental level, they are used to better understand behavior. In this regard they can be used to explore behavioral processes and to test empirically behavioral theories and confirm or refute behavioral hypotheses. They are also used to tease out market segmentation and individual consumer preferences in order to inform target marketing, recommendation systems, or behavior change programs. And they are used for product design and policy analysis to understand the impact that various proposed designs or scenarios have on demand.

While the question of interest is often related to aggregate quantities, such as the market demand for a commodity or service, such aggregate behavior is the result of individual decisions. Therefore the modeling of individual behavior is either explicitly or implicitly at the core of all models of behavior. While there are approaches that directly model aggregate demand, the methods in this book focus on explicitly modeling behavior at the individual level and then, if of interest, the individual results are aggregated to estimate behavior at the market level.

In creating models of individual behavior, we are concerned here with a theory of behavior that is (1) *descriptive*, in the sense that it postulates how human beings behave and does not prescribe how they ought to behave, (2) *abstract*, in the sense that it can be formalized in terms that are not specific to particular circumstances, and (3) *operational*, in the sense that it results in models with parameters and variables that can be measured or estimated, and can be used for forecasting.

Unfortunately there does not exist a single, universally accepted choice theory that satisfies these requirements. Alternative theories differ in the level of detail in which they abstract the choice processes that produce observed behavior. The level of description of choice theory in this book is

coarse relative to recent developments in behavioral sciences. Our emphasis is instead on the quantitative and operational methods that can be used to model choice behavior.

We begin this chapter with relatively simple and mechanical descriptions of the choice process in order to provide a clear frame of reference for discrete choice methods. We first present a general framework for choice theories and describe some common assumptions used for these theories. We focus on microeconomic consumer and discrete choice theories, which suppose the existence of a single objective function called utility. With this theoretical background we then proceed to present in detail the theories and properties of probabilistic choice models that are the basis for the empirical discrete choice models developed in the following chapters. We end the chapter with a broader discussion of work in behavioral science, emphasizing that the statistical methods described in this book are readily adaptable to reflect such nuanced behavior. Indeed, much of the current research in discrete choice analysis involves enriching choice models to be more behaviorally realistic and more consistent with findings in the behavioral sciences.

3.3 A Framework for Choice Theories

Discrete choice models are used to model a decision-maker's choice among a set of mutually exclusive and collectively exhaustive alternatives. In this section, the basic elements of the choice problem are presented and terminology is introduced.

At a basic level, a choice can be viewed as an outcome of a sequential decision-making process consisting of defining the choice problem, generating alternatives, evaluating alternatives, making a choice and executing it. An example of a choice problem would be that of a commuter deciding on a mode of travel to work. Her environment and the supply of transportation services determine the alternative modes available for the trip, although she may not be aware of all the possibilities. Suppose that the commuter's alternatives are car, bus, and walk. In the next step of the decision process the commuter evaluates or collects information about the attributes of each available alternative. Assume that there are three relevant attributes: travel time, travel cost, and comfort. The information in Table 3.1 then schematically reflects the information available to the commuter. This information is then processed by the commuter to arrive at a choice of travel mode. To do this, the commuter applies a decision rule — a specific sequence of calculations, for example the commuter could select the fastest mode that costs less than one dollar, irrespective of comfort. The final step in this decision-making

	Attributes		
Alternatives	Travel time (t)	Travel cost (c)	Comfort (o)
Car	t_1	c_1	o_1
Bus	t_2	c_2	o_2
Walk	t_3	c_3	o_3

Table 3.1: A mode choice example

process is obviously the trip to work itself, using the chosen mode.

To develop a choice theory to reflect such a decision-making process requires a collection of procedures that defines the following elements:

1. who (or what) is the decision maker,
2. what are the characteristics of the decision maker,
3. what are the alternatives available for the choice,
4. what are the attributes of the alternatives, and
5. what is the decision rule that the decision maker uses to make a choice.

These elements are described in detail next. It is worth noting that not all observed choice behavior is an outcome of such an explicit decision-making process. An individual can, for example, follow a habit, assume some form of conventional behavior, follow intuition, or imitate someone else who is considered to be an expert or a leader or a friend. Such forms of behavior can be represented as a choice process, for example one in which the decision maker considers only one alternative.

The Decision Maker

The unit of decision making can be an individual person or a group of persons, such as a family or a household. It can also be an organization such as a firm or a government agency. By considering a group of persons or an organization as a single decision maker, it is possible to abstract partially the complex interactions within, say, a household or a firm. We ignore all interactions, negotiations that may take place to come up with the choice. We consider is as an atomic entity making a choice. Thus, though we refer to the decision maker as an individual, it is taken to represent an "actor" in a more general sense. We use the notation n to represent the decision maker throughout this book.

Characteristics of the Decision Maker

Individuals face different choice situations and have widely different tastes. Therefore, though we are ultimately interested in predicting aggregate demand, we must explicitly treat the differences in decision-making processes among individuals. To illustrate this, consider the car travel cost variable in the mode choice example. It depends on the type of car used and the local price of gasoline. Moreover the extent to which the commuter is willing to pay the higher travel cost of a car may depend on this individual's income. Therefore the characteristics of the decision maker are used to capture some of this heterogeneity, and decision-maker characteristics such as income, sex, age, or firm size become an important part of the problem.

Further differences may arise among group decision processes because of the variations of within-group interactions that affect the outcomes. For example, in selecting an automobile, some household decisions may be the result of the preferences of a single, dominant household member, whereas other household choices may result from complex intrahousehold bargaining processes. While basic analysis may treat within-group interaction as a black box, other approaches may explicitly capture the within-group interaction.

The Alternatives

Any choice is, by definition, made from a nonempty set of alternatives. The environment of the decision maker determines what we shall call the *universal set* of alternatives. Any single decision maker considers a subset of this universal set, termed a *choice set*.

It is useful to distinguish between two general types of choice sets. In the first type the choice set is continuous. This is most natural in the case of "commodity bundles" that form the basis for much of microeconomic demand analysis. For example, the choice set might be defined as the set of all economically feasible amounts of milk (q_1), bread (q_2), and butter (q_3) purchased by a household. In this case it is natural to think of the choice set as depicted in Figure 3.1 (the non-negative area under the budget plane), where p_1, p_2 and p_3 represent the (unit) prices of milk, bread, and butter, respectively, and I is the household's available income.

The second type of choice set—and the one we shall focus on in this book—is where the alternatives are naturally discontinuous, i.e. discrete. Further, the choice of interest in this book is that when the consumer selects a single alternative from a set of mutually exclusive alternatives. This choice set is the finite and countable set of alternatives from which a decision maker chooses one and only one alternative. This includes the alternatives that are

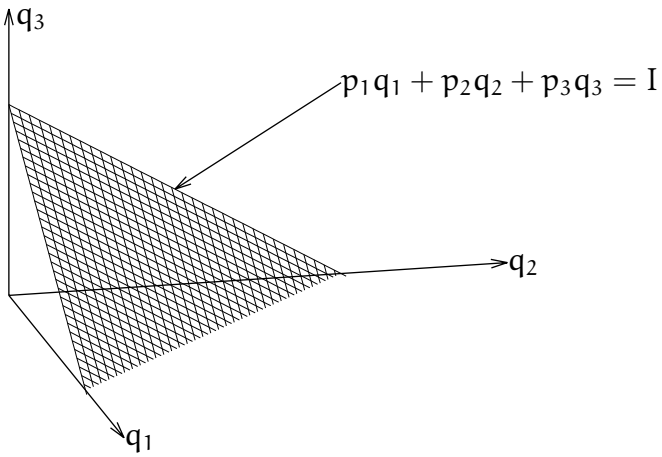


Figure 3.1: A continuous choice set

both feasible to the decision maker and known during the decision process. The feasibility of an alternative is defined by a variety of constraints such as physical availability (e.g., the availability of a bus service between the commuter's home and place of work), monetary resources (e.g., a taxi fare may be unaffordable to a low-income worker), time availability (e.g., the walk mode may be infeasible for a long-distance commuter), informational constraints (e.g., lack of knowledge about the bus service), and so on. Ben-Akiva and Boccara (1995) and Morikawa (1996) contain further discussion of the role of environmental and personal constraints on the composition of the choice set. As an example suppose we are interested in a household's choice of one of a set of three televisions, denoted A, B, and C. In this case, assuming the household can afford any of the three models, the choice set is simply the set of "points" defined as $\{A, B, C\}$ as shown in Figure 3.2. Perhaps the household not only chooses what type of television but how many televisions. Take, for example, a maximum number of 2 televisions per household and the same three television types, then the choice set expands to ten alternatives: $none, A, B, C, AA, BB, CC, AB, AC, BC$. The distinction between these two types of choice sets is expanded on in Section 3.6.1.

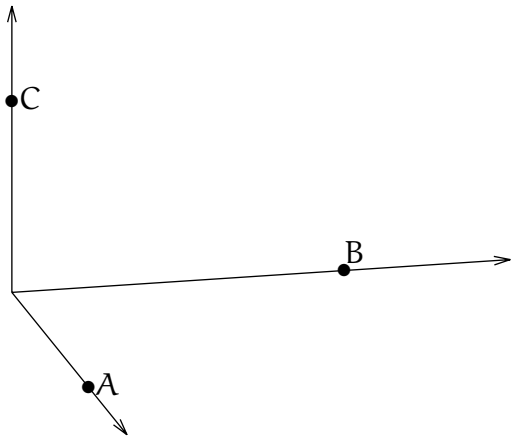


Figure 3.2: A discrete choice set

Alternative Attributes

The attractiveness of an alternative is evaluated in terms of a vector of attribute values. In microeconomic consumer theory an alternative is just a vector of quantities of the commodities, this vector of "attributes" simply reduces to the quantities (e.g., gallons of milk and pounds of butter). However, here we are concerned with situations where alternatives are described by their attributes and where decision makers may have different choice sets, evaluate different attributes, and assign diverse values for the same attribute of the same alternative. In these cases it is much more natural to work directly with a general characterization of each alternative by its attributes rather than just the quantities associated with it. Examples of such attributes include price, time, color and size. Attributes may be measured on a continuous scale (for example, time or price), or be categorical. Furthermore, categorical can be ordinal (1, 2, 3 bathrooms) or nominal (red, green, blue). Some attributes may be more straightforward to measure, such as price or number of rooms, and others may be more complex, such as comfort or quality.

While the values of some attributes are known with certainty, other values

may not be known with certainty. For example, travel time by automobile in congested urban areas may be an uncertain attribute for mode choice because of its great variability from day to day. This aspect of auto travel would be represented as an attribute of the alternative. We could consider, for example, the variance of travel time as an additional attribute. Another example is a consumer who, before buying a new television set, evaluates the expected maintenance costs of alternative brands. The purchase price of a specific new television set may be known with certainty, but its performance over time can only be evaluated in terms of expectations or ranges of possible values.

The Decision Rule

A choice from a choice set containing two or more alternatives requires a decision rule. It describes the internal mechanisms used by the decision maker to process the information available and arrive at a unique choice. From the utility maximizing assumption in microeconomics, to the heuristic decision rules in psychology (e.g., Slovic et al., 1977, and Svenson, 1979), to the latest in behavioral economics (e.g., Kahneman, 2011), there exist a wide variety of decision rules that have been proposed. Here we provide a high level discussion of this literature, describing only a handful in any detail. Further, we emphasize the one most closely associated with discrete choice analysis, *utility maximization*.

The decision rules can be broadly divided into those based on a heuristics to those based on an optimality criterion. Heuristic decision rules are those where the decision maker uses shortcuts to simplify the choice process. These are motivated by the idea that humans have cognitive limitations and cannot (or do not) have full information on the choice problem nor can they process the information to obtain an optimal result. Three examples are described here: *dominance*, *satisficing*, and *lexicographic*.

Dominance: An alternative is dominant with respect to another if it is better for at least one attribute and no worse for all other attributes. The dominance rule states that decision-makers choose this dominant alternative. However, in most real-world decisions there are many relevant attributes, and it is rare to find an alternative that is dominant over all attributes. For example, in the commuter mode choice it is unlikely that there is an alternative that is the fastest, cheapest and most comfortable. A variation of this rule is to introduce a range of indifference for each attribute, so that the alternative is inferior only if the difference in the attribute values exceeds a threshold.

Satisficing: Satisficing is also based on examining attribute levels, but

with the idea that for every attribute there is an acceptability threshold or a level of aspiration. These are based on the decision maker's expectations of the attainable, derived from his current information and previous experiences. Alternatives are eliminated if they do not meet the acceptability threshold for at least one attribute. As with dominance, this rule by itself does not necessarily lead to a choice. In combination with other rules, it can be more decisive. For example, a combination of satisficing and dominance would lead to the choice of the alternative that meets the criterion for at least one attribute and is no worse than the other alternatives for all the other attributes. The commuter may, for example, set upper limits on the travel time and travel cost to be met by car and bus, and car is then chosen because of a higher level of comfort.

Lexicographic: In a lexicographic rule, the decision-maker orders the attributes by their level of "importance" and then chooses the alternative that is the most attractive for the most important attribute. In the case where the use of the most important attribute fails to eliminate all but one alternative, the decision maker goes on with the second most important attribute and continues until the process reaches a unique choice. Alternatively, the process can, at each stage, eliminate the most inferior alternative.

Often decision rules are combined to make a composite decision rule, such as the combination of satisficing and dominance described above. Another example is *elimination by aspects* (Tversky, 1972), which is a combination of lexicographic and satisfaction rules when elimination of alternatives is first done based on the most important attribute, then the second most important attribute, and so forth, until a single alternative remains.

The other main class of decision rules are those based on some form of optimization. The idea is that decision-makers are optimizers who do have sufficiently complete information and are able to process this information according to an objective optimization criterion to result at an optimal choice. Or, at least their decision-process can be relatively well mimicked by such an optimization routine. These are often called *compensatory* decision rules, whereas the heuristic process above are *non-compensatory*. In a compensatory decision process, there is a single objective function that expresses the attractiveness of an alternative in terms of its attributes. As such, a decision-maker can make trade-offs (either explicitly or implicitly) between all the relevant attributes, that is a lower rating on one attribute can be *compensated* by higher ratings on another attribute.

The objective function that is being optimized can take many forms, the most common of which is based on the principle of *utility maximization*.

Utility maximization: This class of decision rules assumes the attractiveness of an alternative can be expressed by a vector of attributes values that

are reducible to a scalar, which is referred to as *utility*. This defines a single objective function expressing the attractiveness of an alternative in terms of its attributes. Utility is then the measure that the decision-maker attempts to maximize through his choice. For the commuter mode choice example this implies that the information in Table 3.1 is reduced to three utility values, U_1, U_2 and U_3. The commuter selects the mode with the highest utility — that is, the mode with the best combination of travel time, travel cost, and comfort. Thus the more costly mode may be chosen if it compensates sufficiently by offering better service (travel time and comfort).

A utility function can be constructed in many different ways. One important distinction is between ordinal and cardinal utilities. An ordinal utility is a mathematical expression of a preference *ranking* of alternatives. An ordinal utility is unique only up to an order-preserving transformation. The comparisons of the numerical assignments to the utilities of alternatives have no meaning except for the relationships of greater than, less than, or equal to. A cardinal utility, on the other hand, implies a uniqueness of its numerical assignment and is therefore more restrictive than an ordinal utility. The latter is used most often in theories of decision making under uncertainty in which decision makers are assumed to maximize a measure of expected utility. Taking expectation involves multiplication and addition that, to be meaningful, require the use of cardinal utility. In the following developments based on the concept of a utility function, we assume an ordinal utility unless it is otherwise stated.

Utility is a broad term representing the attractiveness of an alternative. Depending on the application, this objective function can be defined more specifically. For example, the optimization may be in terms of cost to be minimized or profit to be maximized. Alternatively, rather than maximizing utility, the objective function can be formulated in terms of maximizing gains or expected utility (as referenced above) or as minimizing losses or regret. The key to the optimization decision rules is that there is a clearly specified objective function that is to be maximized or minimized to result in the selection of the optimal alternative.

There are any number of other potential decision rules, for example decisions can be driven by habit or imitation. Further, decision rules can be multi-stage processes and can combine both heuristics and optimization. For example, in the first stage the decision maker may determine what alternatives he/she is to consider (choice set generation) using a heuristic such as satisficing and then the second stage he/she would choose from amongst the alternatives using utility maximization.

Review of the choice modeling framework

The basic elements of the choice modeling process is to define the decision maker, characteristics of the decision maker, the alternatives, the attributes of the alternatives, and the decision rule. Indeed, when discussing data in Chapter 2, we described the example data sets along these categories and Table 2.2 and Table 2.3 provide examples of these dimensions for different choice problems. What wasn't discussed in Chapter 2 was the decision rule. This aspect is emphasized, starting next with a discussion of rationality.

3.4 Rational Behavior

We introduce the concept of rational behavior here, because it is often closely tied to discrete choice analysis. However, it is important to note that nothing in the discrete choice analysis framework requires the assumption of rationality.

The common use of the term "rational behavior" is based on the beliefs of an observer about what the outcome of a decision should be. Obviously different observers may have varying beliefs and may assume different objective functions. Thus, as colloquially used, the notion of rationality is not a useful concept in describing individual behavior.

In the scientific literature the concept is used to describe the decision process itself. In general, it means a consistent and calculated decision process in which the individual follows his or her own objectives, whatever they may be. It stands in contrast to impulsiveness, in which individuals respond to choice situations in different ways depending on their variable psychological state at the time a decision is made. It also assumes individuals make decisions without error and/or biases that would lead to the individual making a decision not in his or her own interests.

The classical concept of perfect rationality assumes an omniscient individual who can gather and store large quantities of information, perform very complex computations, and make consistently optimal decisions based on those computations that are in her best interest. We further define the rationality assumption later when discussing microeconomic consumer theory, and we revisit the rationality assumption at the end of this chapter when discussing broader behavioral theories.

3.5 Microeconomic Consumer Theory

While microeconomic consumer theory is not necessary for the behavioral models described in this book, it does provide a helpful means of interpreting the framework, deriving and constructing choice models, and developing policy analysis metrics such as measures of welfare change. This section provides a brief outline of the key concepts of this theory that are useful for the subsequent sections. For detailed expositions, the reader is referred to many available microeconomic textbooks at different levels of presentation; see, for example, Nicholson and Snyder (2007), Pindyck and Rubinfeld (2008), or Varian (2009). We begin with the classic case of continuous goods in this section, extend the theory to discrete goods in the next section, and introduce probabilistic choice after that.

Microeconomic consumer theory provides a basic approach to the mathematical theories of individual preferences. The objective of the theory is to provide the means to transform assumptions about desires into a demand function expressing the action of a consumer under given circumstances.

Economic consumer theory is concerned with an individual consumer choosing a consumption bundle

$$Q = \begin{pmatrix} q_1 \\ \vdots \\ q_L \end{pmatrix} \tag{3.1}$$

where $q_1, \ldots, q_L$ are the quantities of each of the commodities and services (such as food, shelter, clothing, education, and leisure), $\ell = 1, 2, \ldots, L$ and L indexes each commodity or service. In consumer theory these quantities are generally assumed to be nonnegative, continuous variables. The mathematical analysis employed by this theory to produce its most important results are dependent on this assumption. In the next section we extend the theory to the case of discrete goods, and present different means of analysis.

The choice of consumption bundle is subject to a budget constraint that defines the consumption possibilities, or the choice set. For a fixed income I and fixed prices (in terms of cost per unit) $p_1, p_2, \ldots, p_L$, the budget constraint is

$$\sum_{\ell=1}^{L} p_\ell q_\ell \leq I. \tag{3.2}$$

That is, one can only consume what one can afford. Using vector notation and denoting column vector $p^T = (p_1, \ldots, p_L)$, the budget constraint is written as

$$p^T Q \leq I. \tag{3.3}$$

In microeconomic consumer theory there is no explicit treatment of attributes in addition to the quantities $q_1, \ldots, q_L$ that define a bundle Q.

Microeconomic consumer theory assumes that consumers are rational decision makers. Consumers choose their consumption bundle using consistent and calculated decisions that follow their objectives, whatever they may be. The theory is based on the concept of *preferences*: when consumers are faced with a set of possible consumption bundles, they assign preferences to each of the various bundles and then choose the most preferred bundle from the set of affordable bundles. Mathematically, such preference relationships are denoted by the operators $\succ$, $\sim$, and $\succsim$: $Q_a \succ Q_b$ denotes Q_a is preferred to Q_b, $Q_a \sim Q_b$ denotes indifference between Q_a and Q_b, and $Q_a \succsim Q_b$ denotes Q_a is at least as preferred as Q_b. The rationality assumption means that these preferences have certain properties:

Property 1 completeness: The consumer is assumed to have preferences over alternative consumption bundles. There is no indecision and any two bundles can be compared, i.e., either bundle Q_a is preferred to bundle Q_b, or bundle Q_b is preferred to bundle Q_a, or they are equally preferred. Mathematically:

$$Q_a \succ Q_b \text{ or } Q_a \prec Q_b \text{ or } Q_a \sim Q_b. \tag{3.4}$$

Property 2 transitivity: Rational behavior is defined in the sense of a transitive preference ordering of alternatives, which means the rankings are internally consistent. If Q_a is preferred to Q_b and Q_b is preferred to Q_c, then Q_a is preferred to Q_c. Mathematically,

$$\text{if } Q_a \succsim Q_b \text{ and } Q_b \succsim Q_c \text{ then } Q_a \succsim Q_c. \tag{3.5}$$

Property 3 continuity: This property is required to obtain smooth mathematical functions necessary for the calculus-based derivations. Continuity states that if Q_a is preferred to Q_b and Q_c is arbitrarily "close" to Q_a, then Q_c is preferred to Q_b.

Under these assumptions, there exists an ordinal utility function that expresses mathematically the consumer's preferences:

$$\widetilde{U} = \widetilde{U}(q_1, \ldots, q_L; \theta), \tag{3.6}$$

or, in vector notation

$$\widetilde{U} = \widetilde{U}(Q; \theta), \tag{3.7}$$

where θ is a column vector of parameters that represent the tastes of the individual. The utility is unique up to an order-preserving transformation such as a shift by a constant or a rescale by a positive quantity. The utility function associates a real number with each possible bundle, such that it

summarizes the preference orderings of the consumer. Thus, $\widetilde{U}(Q_a;\theta) \geq \widetilde{U}(Q_b;\theta)$ is equivalent to $Q_a \succsim Q_b$. Any equation or number that gets the preference ordering of the bundles correct is acceptable. An ordinal utility function merely ranks alternatives. This means that all we can say about a bundle with 1000 utils versus a bundle with 100 utils is that the 1000 util bundle is preferred to the 100 util bundle. It would be incorrect to say that the 1000 util bundle is 10 times preferred to the 100 util bundle. (However, note that once the scale of the utility is set, for example when a discrete choice model is estimated, the result is a cardinal utility and the differences in the utility become meaningful.)

Consumer behavior can then be expressed as an optimization problem in which the consumer selects the consumption bundle Q such that his utility is maximized subject to his budget constraint, or

$$\max_{Q} \widetilde{U}(Q;\theta) \tag{3.8}$$

$$\text{s.t. } p^T Q \leq I.$$

This optimization is solved to obtain the demand functions

$$q_\ell^* = f(I,p;\theta) \text{ for all } \ell = 1,\ldots,L, \tag{3.9}$$

which is the amount demanded of each good ℓ as a function of the income of the consumer I, the prices of all of the goods p, and the tastes of the consumer θ. The asterisk denotes that this equation defines the level of *optimal* consumption resulting from the optimization of (3.8). For the sake of notational simplicity, we omit the asterisk in the equations that follow.

The demand function (3.9) can then be substituted back into the utility equation (3.7) to derive the *indirect utility function*:

$$U(I,p;\theta), \tag{3.10}$$

defined as the maximum utility that is achievable for a given set of prices and income. The indirect utility function is a function of prices and income, and not a function of quantities as in the direct utility function of Equation (3.7). The indirect utility function is indirect because the utility is gained indirectly via income, which is used to purchase the goods that are consumed. The indirect utility function is useful to understand how realized utility is impacted by changes in either prices or income. The indirect utility function is also what is used in discrete choice analysis, and is referred to simply as utility throughout most of this text. This is the reason why we use the more complex notion $\widetilde{U}$ for (direct) utility and retain the simpler U for the indirect utility.

Above we mentioned that the microeconomic framework is useful for discrete choice because we can employ the tools developed in microeconomics when doing discrete choice analysis. For example, in discrete choice we work directly with indirect utility functions (the rationale for this is described later). However, there is a useful relationship that allows the demand functions to be derived from the indirect utility function, which is known as Roy's identity:

$$q_\ell = -\frac{\partial U(I, p; \theta)/\partial p_\ell}{\partial U(I, p; \theta)/\partial I} \tag{3.11}$$

This relationship has been deployed in theoretical derivations of discrete choice models such as in McFadden (1981) and Dubin and McFadden (1984).

There are also constructs from microeconomics that prove extremely useful when using discrete choice models in policy analysis. One is the concept of elasticity. This is the percent change in demand resulting form a 1% change in an attribute. It summarizes the quantitative impacts of interest and is attractive because it is unit free. In microeconomic consumer theory, the only attribute of a good is price, and therefore the only elasticity of relevance is the price elasticity. This can be the goods *own* price elasticity, referencing a change in its own price:

$$E_{p_\ell}^{q_\ell} = \frac{\%\text{ change in } q_\ell}{\%\text{ change in } p_\ell} = \frac{\Delta q_\ell/q_\ell}{\Delta p_\ell/p_\ell} = \frac{p_\ell}{q_\ell}\frac{\Delta q_\ell}{\Delta p_\ell}. \tag{3.12}$$

We can calculate the asymptotic version of (3.12) (as $\Delta p_\ell \to 0$) from the demand function (3.9) as follows:

$$E_{p_\ell}^{q_\ell} = \frac{p_\ell}{q_\ell(I, p; \theta)}\frac{\partial q_\ell(I, p; \theta)}{\partial p_\ell} = \frac{\partial \ln q_\ell(I, p; \theta)}{\partial \ln p_\ell}. \tag{3.13}$$

Or, it can be a *cross* price elasticity, referencing a change in the price of another good (m) calculated as:

$$E_{p_m}^{q_\ell} = \frac{p_m}{q_\ell(I, p; \theta)}\frac{\partial q_\ell(I, p; \theta)}{\partial p_m}. \tag{3.14}$$

When we move to the attribute space where goods are described by prices and attributes, we can calculate demand elasticities of other attributes.

The other important concept is consumer surplus, or welfare. Consumer surplus is the difference between what a consumer is willing to pay for a good and what they actually pay for a good. It is equal to the area under the demand curve and above the market price as shown in Figure 3.3. The change in consumer surplus between, for example, different policy environments or

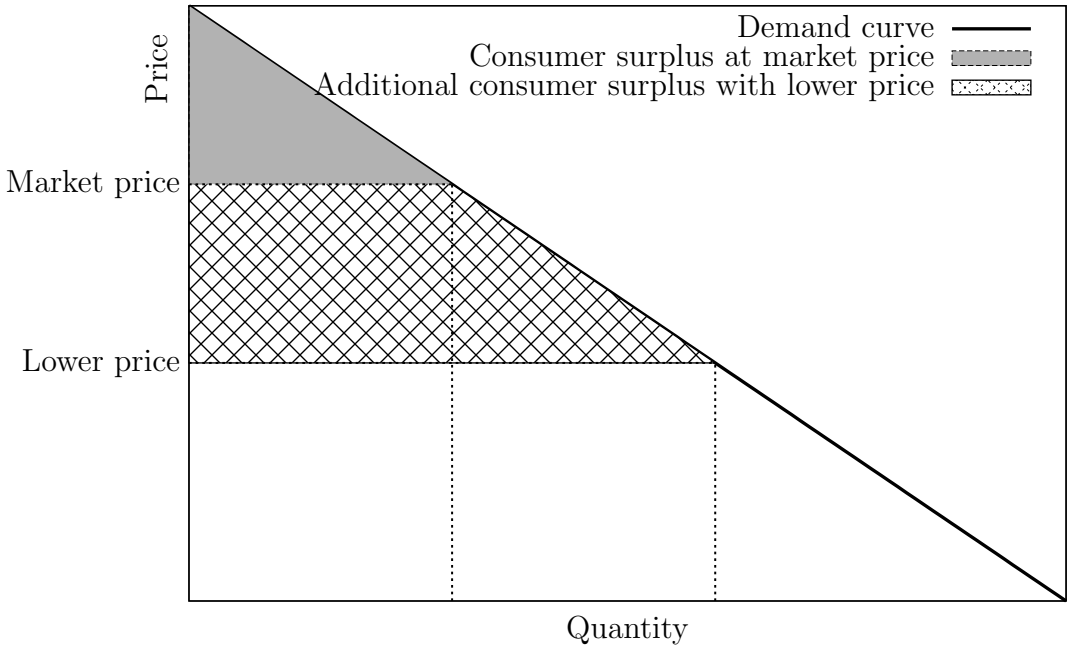


Figure 3.3: Illustration of the consumer surplus

government investment scenarios is used to evaluate public policy decisions. For example, the impact on consumers of changing emissions regulations or increasing investments in the public transit system. Figure 3.3 displays the change of consumer surplus when prices are lowered. The change in consumer surplus is the fixed amount of consumer surplus gained from those goods consumed at the market place (the hatched rectangle) plus the consumer surplus from the additional quantity of the good that is consumed (the hatched triangle). In Chapter 10 we discuss how to calculate both elasticities and consumer surplus from a choice model.

3.5.1 Example with Two Commodities

In this section, we illustrate microeconomic consumer theory using a simple example with two commodities. We also introduce additional concepts that are easier to understand in this simple commodity space.

The utility function can take on any number of mathematical forms. Here we use a Cobb Douglas form, which for our two commodity case is as follows:

$$\widetilde{U}(q_1, q_2; \theta) = \theta_0 q_1^{\theta_1} q_2^{\theta_2} \tag{3.15}$$

where $\theta = (\theta_0, \theta_1, \theta_2)^\mathsf{T}$ is a column vector containing the three positive parameters and representing the tastes of the consumer.

A three-dimensional plot of this utility function is shown in Figure 3.4. The value of the utility starts at zero at the origin where no amount of either q_1 or q_2 are consumed (the leftmost point on the figure). It remains zero as long as at least one of the commodities is not consumed. This is a property of the Cobb Douglas, not utility functions in general, and represents the idea that satisfaction is gained only if at least some of both commodities are consumed, for example at least *some* food and *some* housing need to be consumed to have positive utility. Utility then increases as one moves away from the origin and axis, with increasing consumption of both goods.

A two-dimensional plot of this utility function is projected on the base in Figure 3.4 and plotted directly in 2D in Figure 3.5. In this 2D plot, the utility axis is coming out of the page from the origin. Each location in the quadrant represents a particular consumption bundle made up of a specific quantity of q_1 and q_2. The values of the utility resulting from the consumption bundles are shown by plotting utility isoquants or *indifference curves*; any combination of quantities q_1 and q_2 that fall on the same indifference curve results in the same value of utility. This graph shows many indifference curves, the closer to the origin, the lower the value of the utility and vice-versa. In reality there are infinitely many indifference curves. Any consumption bundle along a particular indifference curve are preferred to all those bundles that lay on indifference curves closer to the origin. Two points A and B are plotted. Since they are on the same indifference curve, the consumer is indifferent between the consumption bundles A and B, that is he is equally satisfied with A which has a relatively more amount of q_1 and relatively less amount of q_2 as he is satisfied with B which has less q_1 but more q_2.

That one can move along an indifference curve and retain the same level of utility represents the notion of trade-offs, which is that a loss (or gain) of consumption of one good can be compensated by a gain (or loss) of another good. Mathematically, this concept of tradeoffs is represented by the *marginal rate of substitution* or MRS, which is calculated as the slope of the indifference curve $\partial q_2/\partial q_1$. This is calculated from the utility equation by taking the ratio of the *marginal utilities* $\partial\widetilde{U}(q_1, q_2; \theta)/\partial q_i$. For the Cobb Douglass utility function, the marginal rate of substitution of q_1 for q_2 is:

$$\text{MRS} = \frac{\partial\widetilde{U}(q_1, q_2; \theta)/\partial q_1}{\partial\widetilde{U}(q_1, q_2; \theta)/\partial q_2} = \frac{\theta_1 q_2}{\theta_2 q_1}. \tag{3.16}$$

This is equal to the amount q_2 must be increased if q_1 is decreased by one unit in order for the consumer's utility to remain unchanged. For the Cobb

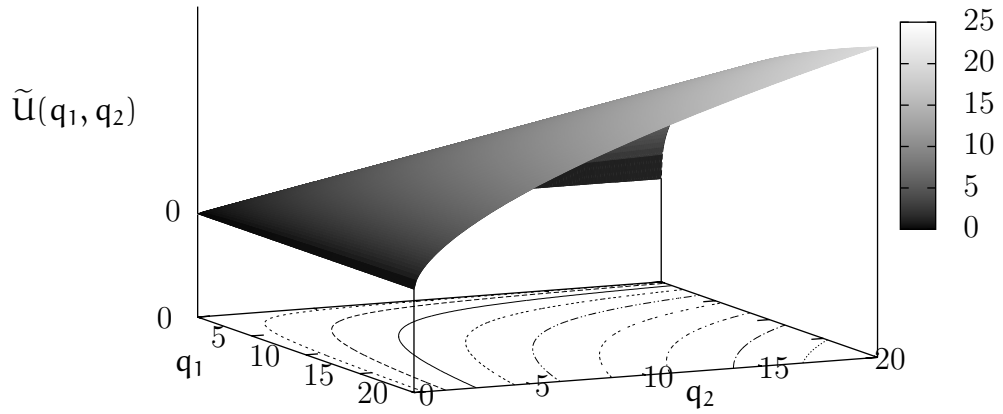


Figure 3.4: Plot of utility for 2 commodity case (3D)

Douglas, the MRS varies only in terms of the parameters and the existing ratio of q_2 to q_1 and not in the absolute quantities of each good. The higher the ratio of q_2 to q_1, the more valuable each unit of q_1 and the more q_2 is needed to compensate for a loss of a unit of q_1. The more q_1 is preferred to q_2, represented by the ratio of θ_1 to θ_2, the more q_2 is needed to compensated a loss in q_1.

Now that we introduced the utility function and some of its properties, we move on to consumer behavior. Given all the possible values of q_1 and q_2, which specific consumption bundle does the consumer choose? The behavioral assumption is that the consumer wants to maximize his utility. Graphically, he want to get as high up the mountain in Figure 3.4 as he can. What stops him from continuing to climb farther away from the origin and consuming more goods? That these goods have prices and that the consumer has a given income to spend on the goods. Therefore, following (3.8), the consumer choice problem is formulated as the following optimization problem:

$$\max_{q_1, q_2} \widetilde{U} = \theta_0 q_1^{\theta_1} q_2^{\theta_2} \tag{3.17}$$

$$\text{s.t. } p_1 q_1 + p_2 q_2 = I.$$

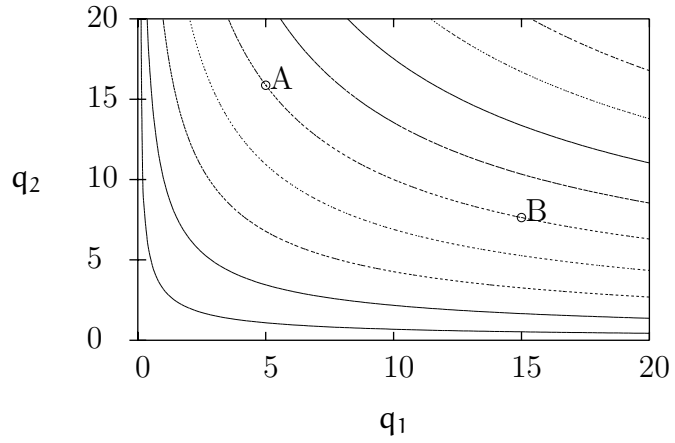


Figure 3.5: Plot of utility for 2 commodity case (2D indifference curves)

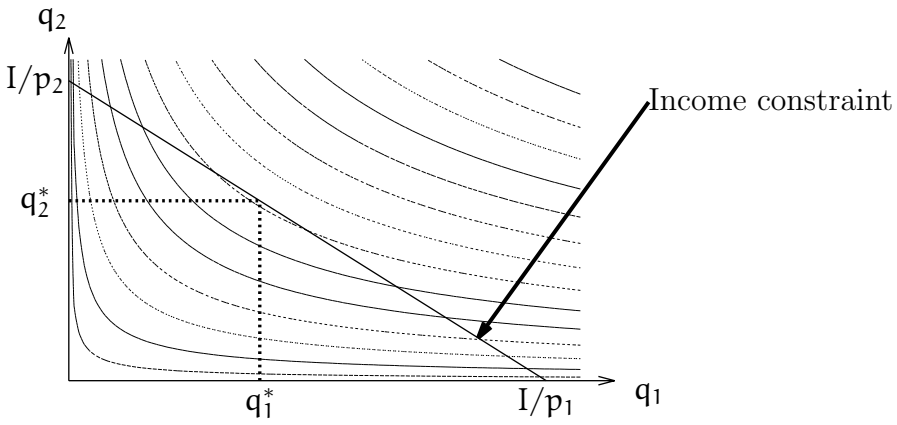


Figure 3.6: Geometry of the consumer choice problem

This is simply stating that the consumer selects the consumption bundle (a quantity of q_1 and a quantity of q_2) that maximizes his utility $\widetilde{U}$ and is within his available budget I. The outcome is the amount of q_1 and q_2 that he consumes.

The solution to this optimization problem is shown graphically in Figure 3.6. The utility function is plotted as before using indifference curves. Added to this figure is the income constraint, shown as a line connecting $(0, I/p_2)$ (the point where all the budget is spent on q_2 and none on q_1) with $(I/p_1, 0)$ (the point where all the available budget is spent on q_1 and none on q_2). The feasible set of consumption bundles is the triangular area below the budget curve. However, as the consumer can alway increase his utility by spending more money, he spends all of his available money and the optimal consumption bundle lay directly on the budget line. More specifically, it is at the point along the budget line that touches the highest achievable utility indifference curve. This point is shown on the figure at q_1^* and q_2^*, the asterisks denoting it is the point of optimal consumption. It can also be seen graphically that this is the point where the slope of the indifference curve (the MRS) is equal to the slope of the budget line (in this case $-p_1/p_2$).

To solve the optimization problem mathematically, the Lagrangian func-

tion is formed:

$$L(q_1, q_2, \lambda; \theta) = \theta_0 q_1^{\theta_1} q_2^{\theta_2} - \lambda(p_1 q_1 + p_2 q_2 - I), \tag{3.18}$$

where λ is the Lagrange multiplier. The Lagrangian procedure is simply a trick to turn a constrained optimization problem (3.17) into an unconstrained optimization problem (3.18). In this way, the straightforward procedure from beginning Calculus can be used: the optimal is at the point at which the first derivatives are equal to zero (and the second derivatives are negative to ensure a maximum rather than a minimum). In this case the objective function has three unknowns: q_1, q_2 and the Lagrange multiplier λ (a byproduct of the construction of the Lagrange function). The Lagrange function is differentiated with respect to each of the three unknowns to obtain the first-order conditions:

$$\begin{array}{lcl} \partial L/\partial q_1 = \theta_0\theta_1 q_1^{\theta_1-1} q_2^{\theta_2} - \lambda p_1 & = & 0, \\ \partial L/\partial q_2 = \theta_0\theta_2 q_1^{\theta_1} q_2^{\theta_2-1} - \lambda p_2 & = & 0, \\ \partial L/\partial \lambda = p_1 q_1 + p_2 q_2 - I & = & 0. \end{array} \tag{3.19}$$

These conditions can be solved for the demand functions, expressing the quantities that are consumed of each good for given prices and income:

$$q_1 = \frac{\theta_1}{\theta_1 + \theta_2}\frac{I}{p_1}, \tag{3.20}$$

$$q_2 = \frac{\theta_2}{\theta_1 + \theta_2}\frac{I}{p_2}.$$

The Cobb Douglass has the property that the demand for a good is only dependent on its own price and independent of the price of any other good, which is a fairly restrictive assumption.

The equations can also be solved for the third unknown, the Lagrange multiplier λ:

$$\lambda = \theta_0 (\theta_1 + \theta_2) \left(\frac{\theta_1}{\theta_1 + \theta_2}\right)^{\theta_1} \left(\frac{\theta_2}{\theta_1 + \theta_2}\right)^{\theta_2} \frac{I^{(\theta_1+\theta_2-1)}}{p_1^{\theta_1} p_2^{\theta_2}}. \tag{3.21}$$

The parameter λ is not just a nuisance parameter but has a useful interpretation. Its value is the marginal utility of income, that is the increase in utility that results if income is increased by one unit. Equivalently, λ is equal to the marginal utility of good ℓ ($\partial\widetilde{U}/\partial q_\ell$) divided by the marginal cost of good ℓ (equal to p_ℓ in this example) for all goods, or

$$\lambda = \frac{\partial\widetilde{U}/\partial q_\ell}{p_\ell} \text{ for all goods } \ell. \tag{3.22}$$

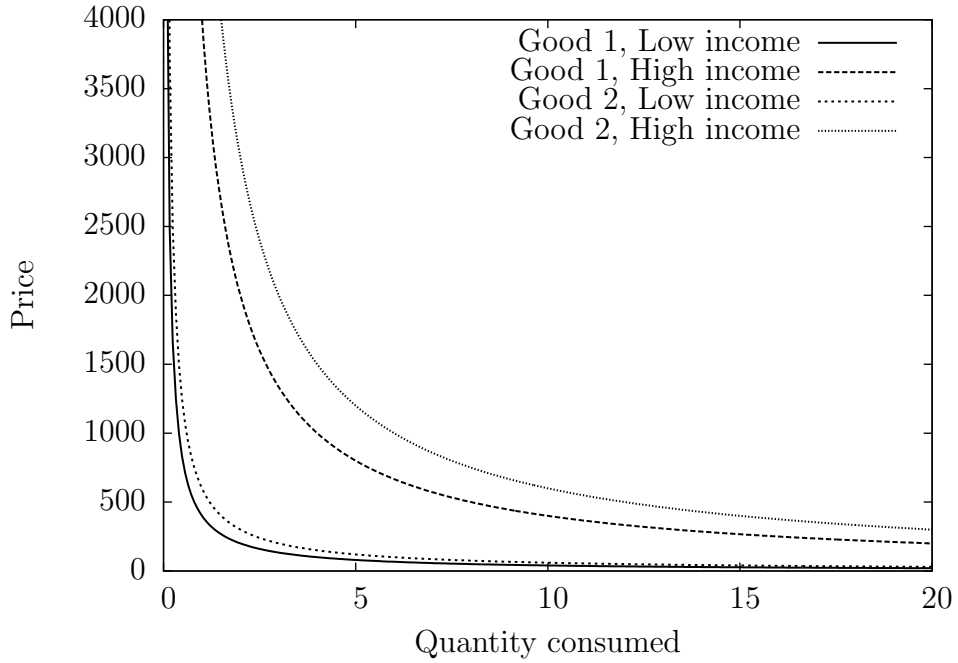


Figure 3.7: Demand functions for the two good case

This is often described as an optimality condition. Conceptually, at optimal consumption each good should yield the same marginal utility per monetary unit spent. At optimality, if given one extra unit of income to spend, the consumer is indifferent as to which good to purchase more. If the consumer is not indifferent, then he was not at optimality and should adjust his consumption bundle towards the preferred good. The optimality conditions can also be rearranged to state that the marginal rate of substitution of good i for good j is equal to the ratio of the marginal costs of good i relative to good j. For this case of a linear budget constraint, this optimality condition is written as

$$\frac{\partial \widetilde{U}/\partial q_i}{\partial \widetilde{U}/\partial q_j} = \frac{p_i}{p_j} \text{ for all goods } i \text{ and } j. \tag{3.23}$$

Conceptually, this means that at optimality, the psychic trade off between goods (that is, the relative utility gain) should equal the market trade off (that is, the ratio of prices). These optimality conditions are true for any utility equation, not just the simple example here.

As demand is the focus of this book, we return to the demand equations. They are plotted in Figure 3.7 (with $\theta_1 = 0.4$, $\theta_2 = 0.6$, low income is

1000 and high income is 10,000) using the standard axes of microeconomics with price on the vertical axis and quantity demanded on the horizontal axis (sideways!). They reflect the expected behavior that as the price of a good decreases the demand for the good increases, and as income increases the demand for the good increases. In this case, good 2 is preferred to good 1 ($\theta_2 > \theta_1$).

The demand functions provide an expression for the "optimal" consumption bundle that can now be substituted in the utility function to obtain the maximum utility that is achievable under the given prices and income. As described above, the result is called the *indirect utility function*. For this example the indirect utility is:

$$U(p_1, p_2, I; \theta) = \theta_0 \left(\frac{\theta_1}{\theta_1 + \theta_2}\right)^{\theta_1} \left(\frac{\theta_2}{\theta_1 + \theta_2}\right)^{\theta_2} \frac{I^{(\theta_1+\theta_2)}}{p_1^{\theta_1} p_2^{\theta_2}} \tag{3.24}$$

Note that Roy's identity (3.11) can be applied to this indirect utility equation to recover the demand functions (3.20). Further, taking the derivative of this indirect utility function with respect to income gives us the marginal utility of income, or the amount that utility increases with a unit increase of income. Recall that this same construct was discussed when interpreting the Lagrange multiplier λ, and indeed the partial derivative of (3.24) with respect to income is identical to the Lagrange multiplier (3.21).

While very brief, this section aimed to cover the basic elements of microeconomic consumer theory that are important for discrete choice analysis. In the next subsection we cover some extensions to the theory that are useful or operationalizing the theory for discrete choice analysis. The two most important extensions are highlighted in the sections that follow where we extend the microeconomic framework to discrete choice and discuss choice probability theory.

3.5.2 Extensions of Microeconomic Theory

Economic consumer theory is developed in a rather theoretical framework, in which the consumer is faced with a large number of commodities, no details on the nature of the alternatives, and an unreasonably complex optimization problem. As a result, the demand functions that are derived have limited empirical usefulness because they require information on the prices of all commodities. Practical applications require further restrictions and descriptions of the consumption problem. A number of people have worked to expand the framework in order to develop an operational framework and also to better map the framework to an underlying behavioral process. Some of the more

important developments for the operational discrete choice models of this text are described here.

In principle, all commodities and services to which the consumer devotes some fraction of his or her budget are related because of the budget constraint. However, it is possible to impose some plausible structure on the utility function that restricts this dependence. Strotz (1957) proposed the concept of a "utility tree." The commodities are arranged into groups or branches, and the total utility function is composed from separable branch utilities as follows:

$$\widetilde{U} = \widetilde{U}[\widetilde{U}^1(q_1^1, \ldots, q_{L_1}^1), \widetilde{U}^2(q_1^2, \ldots, q_{L_2}^2), \ldots, \widetilde{U}^B(q_1^B, \ldots, q_{L_B}^B)], \qquad (3.25)$$

where $\widetilde{U}^b(q_1^b, \ldots, q_{L_b}^b)$ is the utility of the commodities in branch $b = 1, \ldots, B$, $L = L_1 + L_2 + \cdots + L_B$, and $B =$ the number of branches.

The function $\widetilde{U}[\,]$ determines the level of separability between branches. Strong separability implies an additive function. The branches correspond to "natural" groups of commodities such as housing, food, clothing, vacation, and transportation. Behaviorally, a utility tree can be interpreted as a representation of a sequential decision process. The consumer follows a two-stage procedure. The first stage consists of the decision on the allocation of the available resources, namely income to the different branches. The consumer can be assumed to base this allocation on some composite prices, such as an average cost of food or an index of vacation related prices. In the second stage the consumer decides on the within-branch allocations. Given that income is allocated to a branch, this is a classical consumer theory problem applied to a subset of commodities. The resulting demand functions include only prices of *related* commodities. This is the justification to examine, for example, housing consumption in isolation from other consumer decisions.

Another interpretation of a separable utility function was proposed by Muth (1966). It views the commodities and services purchased by the consumer as inputs into a home production process, the output of which is a bundle of "non-market" goods that represent basic needs of the consumer such as shelter, nourishment, and relaxation. That is, the non-market goods (basic needs) are created by consuming the market goods (traditional products and services). The inputs to the home production process are the market goods themselves as well as the income needed to purchase the goods. The production function expresses the quantity of the non-market good as a function of the input market commodities. The output of the production process is a bundle of the non-market goods, and these are then the arguments of the traditional utility function. The utility function as defined over these non-market goods can be interpreted in the same manner as the Strotz's branches of a utility tree.

Another interpretation was suggested by Lancaster (1966). He proposed that it is the attributes of the goods that determine the utility they provide. Therefore, utility can be expressed as a function of the attributes of the commodities, $U = f(A)$ and $A = f(Q)$ where A are the attributes and Q is the vector of quantities. For example, housing can be described by its square meters, quality, style, amenities, and location. The production process in Muth's interpretation is replaced by technical relationships between commodities and attributes. The consumer derives utility from the attributes rather than the goods themselves. The preferences for commodities are indirect in the sense that they arise because the commodities are needed to "produce" attributes. Such a framework is central in discrete choice analysis, where the discrete alternatives are almost always represented by their attributes.

Becker (1965) extended the traditional consumer theory formulation by emphasizing that in addition to the income constraint there is also a time constraint. The utility function is defined in terms of human activities. Performing an activity requires the purchase of market commodities and services *and* the spending of time. This again defines a partial utility function (or a technological relationship expressing the utility or quantity) of an activity as a function of commodities and time. This function can also include other attributes of an activity such as comfort or safety.

3.6 Microeconomic Theory of Discrete Goods

Microeconomic consumer theory as discussed above is concerned with continuous (i.e., infinitely divisible) products, which is a necessary condition to employ the calculus used to derive many of the key results. There are many consumption problems that are discrete in nature such as which health or retirement plan to join, whether or not to get married, or what occupation to have. Therefore, the microeconomic framework needs to be expanded to address the case of discrete goods. In this section we introduce discrete goods into the microeconomic framework and also discuss how this analytical framework leads to the operational utility equations that are used throughout this text.

Following McFadden (1981), the consumption bundle in consumer choice that was described above can be expanded to include two classes of goods: those that are continuous (i.e., infinitely divisible) and those that are discrete. The consumer choice is then to choose both the quantities of continuous goods $Q = (q_1, \ldots, q_L)$ (as before) as well as the selection of a discrete choice $i = 1, \ldots, j, \ldots, J$ from among the set of available discrete alternatives

$\mathcal{C}$.

Just as the continuous goods are said to include the full range of (continuous) goods consumed by an individual (food, leisure, education, clothing, housing, etc.), the discrete alternative represents a combination of all of the discrete choices that are made by the consumer (a combination of the choice of labor force participation, occupation, marital status, education level, computer brand purchased, travel mode to work, etc.). As is nearly always done in the application of microeconomics, we later place more bounds on the consumption problem to make it tractable, but for now we retain this general presentation.

The discrete alternatives are mutually exclusive and collectively exhaustive and consist of all of the combination of goods in the discrete good space. Each discrete alternative i represents a particular combination of the available discrete goods, and the consumer selects only one of these alternatives to consume. For example, given the list of discrete goods listed above, a particular choice i may be the combination to work, not get married, obtain a bachelors degree, buy a macintosh, and bike to work. While in this general framework, each alternative i is a bundle of discrete goods, in the derivation that follows we refer to each alternative i as a discrete good (singular). The consumption of the discrete good is represented by a J-dimensional column vector $y = (y_1, \ldots, y_i, \ldots, y_J)^T$, where y_i is equal to 1 for the single discrete alternative that is consumed and is equal to 0 for all of the other discrete alternatives (all of which are not consumed). So the utility function is

$$\widetilde{U} = \widetilde{U}(Q, y; \theta). \tag{3.26}$$

This is the continuous and discrete analogy of (3.7). Note that, due to the discrete nature of the y variables, the optimization problem has no optimality condition, so that the demand function cannot be derived directly like in the continuous case.

The y vector consists of zeros and ones and is not very meaningful; therefore, along with the y vector, we also include *attributes* of the alternatives. Denote $\tilde{z}_i$ as a K-dimensional column vector of attributes of discrete alternative i, where K is the number of different attributes used to describe the alternatives. $\tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_i, \ldots, \tilde{z}_J)^T$ is then a $J \times K$ matrix of attributes. The attribute vector of the consumed alternative is then equal to $\tilde{z}^T y$, and the utility function becomes $\widetilde{U}(Q, y, \tilde{z}^T y; \theta)$.

The budget constraint is as before (3.3) except that the cost of the discrete alternative must be included in the budget. Denote c_i as the cost of discrete alternative i and $c = (c_1, \ldots, c_i, \ldots, c_J)^T$ as the J-dimensional column vector costs. The cost of the consumed alternative is then equal to $c^T y$. The budget

equation is then $p^T Q + c^T y \leq I$. Putting the utility equation together with the budget equation, the consumer optimization problem is then written as

$$\max_{Q,y} \widetilde{U}(Q, y, \tilde{z}^T y; \theta) \tag{3.27}$$

subject to

$$p^T Q + c^T y \leq I.$$

and

$$\sum_j y_j = 1 \quad y_j \in \{0, 1\}, \forall j.$$

This is the continuous and discrete analogy of (3.8).

The question becomes how to solve this optimization problem now that there are both continuous arguments Q and discrete arguments y. McFadden (1981) describes a *two step procedure*, which is now the theoretical basis of discrete choice analysis that is performed today.

Step 1 is to solve the continuous consumption problem *conditional* on the selection of a *particular* discrete alternative $i \in \mathcal{C}$. This leads to the following conditional demand functions

$$q_{\ell|y} = f(I - c^T y, p, \tilde{z}^T y; \theta). \tag{3.28}$$

Or, stated individually for each discrete good i, the conditional demand function for each alternative i is written as

$$q_{\ell|i} = f(I - c_i, p, \tilde{z}_i; \theta) \text{ for all } i \in \mathcal{C}. \tag{3.29}$$

These demand functions are the conditional (on i) versions of the continuous demand functions (3.9). The conditional demand function is a function of income, prices (of both the continuous and discrete goods), and the attributes of the discrete goods consumed. $I - c_i$ is termed *income remaining*, which is the income that is left to spend on the continuous goods after the discrete good is purchased. As the demand function is conditional on consuming i, the cost of i must be removed from the budget available to spend on the continuous goods. If the cost c_i exceeds the income, then alternative i is not available.

Inserting these conditional demand functions (3.28) back into the utility (3.27) leads to the *conditional indirect utility functions*:

$$U_y = U(I - c^T y, p, \tilde{z}^T y; \theta) \tag{3.30}$$

Or, stated individually for each discrete good i, the conditional indirect utility function for good i is

$$U_i = U(I - c_i, p, \tilde{z}_i; \theta) \text{ for all } i \in \mathcal{C}. \tag{3.31}$$

These are analogous to the pure continuous indirect utility functions of (3.10), but conditional on consuming choice i.

In *Step 2*, the consumption of the discrete goods is determined by maximizing these conditional indirect utilities to find the discrete consumption bundle i that maximizes the consumer's utility. Mathematically, the optimization problem is written as:

$$\max_{y} U(I - c^T y, p, \tilde{z}^T y; \theta) \tag{3.32}$$

In this case there is no income constraint, because it is already incorporated in the conditional indirect utilities via the income remaining term.

Just as the outcome of the optimization over continuous goods is the demand for the continuous goods (3.9), the outcome of this optimization over the discrete goods (3.32) is the demand for discrete goods (i.e., which discrete alternative is chosen.) It is a function of all prices p and c, income I, the attributes of the alternatives $\tilde{z}$, and the tastes of the consumer θ. The demand function for alternative i is written as

$$y_i = \begin{cases} 1 & \text{if } U(I - c_i, p, \tilde{z}_i; \theta) \geq U(I - c_j, p, \tilde{z}_j; \theta) \text{ for all } j \in \mathcal{C}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.33}$$

The demand y_i for each bundle i is either equal to 1 (if it is consumed) or 0 (if it is not consumed). Along with the demand for the good, the value of the realized utility and level of attributes are determined. This is the demand for a discrete good, analogous to the demand equations for continuous goods (3.9).

The rest of this section focuses on how this methodology is applied in practice, and in particular how the conditional indirect utility function of equation (3.31) reduces to a form closer to what is used in practice.

While including income remaining $I - c^T y$ in the conditional indirect utility function is most consistent with the microeconomic derivation, if income remaining enters the conditional indirect utility linearly then income I cancels out. Recall that when we work with ordinal utility functions, adding the same constant to all of the utility functions (such as I in this case) does not mathematically alter the problem. In most cases this is the specification that is seen in practice, and therefore only the costs of the alternatives are included in the utility equations. With this specification, the utility function reduces to:

$$U_i = U(c_i, p, \tilde{z}_i; \theta) \text{ for all } i \in \mathcal{C}. \tag{3.34}$$

Note that if a transform such as $\ln(I - c^T y)$ is used, income does not cancel out.

Another notational convention is that cost is included within the vector of the attributes of the alternatives, rather than distinguishing it as a different type of variable. In the optimization problem, the distinction was necessary because the costs of the alternatives enter the budget constraint and the attributes of the alternatives enter the (direct) utility function. However, such distinction is not necessary in the conditional indirect utility function. Therefore, we denote $z_i = (c_i, \tilde{z}_i)$ and write the conditional indirect utility as follows:

$$U_i = U(z_i, p; \theta) \text{ for all } i \in \mathcal{C}. \tag{3.35}$$

The vector of prices of the continuous goods p can be combined and represented by a price index. If this price index is something that varies by discrete alternative, then the price index would become a part of z_i. If the price index can be assumed not to vary by discrete alternative, then the same value enters each indirect utility equation and therefore cancels out of the equations. In either case, the explicit notation of p can be dropped from the equation, and the conditional indirect utility function is written as:

$$U_i = U(z_i; \theta) \text{ for all } i \in \mathcal{C}. \tag{3.36}$$

For example, consider a cost of living index. If the choice of interest is the city in which to live, the cost of living would impact the alternatives and would be included as part of z_i. If the choice is which mode to take to work, the cost of living would not be impacted and would cancel out. Further, and as implied by this example, typically the separability argument is invoked and only a narrow subset of discrete alternatives are considered rather than the entire space of discrete goods. For example, the choice of interest could be the type of mobile phone a person chooses and the choice set $\mathcal{C}$ would consist of the various mobile phones on the market. For each mobile phone in the choice set, a conditional indirect utility function as in (3.36) would be written, and all other conditional indirect utility functions would be disregarded. The datasets described in Chapter 2 provide other examples of choice sets that are modeled using discrete choice analysis.

All of the discussion thus far has dealt with a single consumer whose tastes are implicitly contained in the form of the utility function and its parameter values θ. In empirical applications, when observations of different consumers are used to estimate unknown parameters of demand functions, it is necessary to specify how tastes, and consequently utility functions, vary among consumers. Therefore, we introduce a subscript n to denote the utility and tastes of a given member of the population:

$$U_{in} = U(z_{in}; \theta_n), \tag{3.37}$$

where n signifies a particular decision maker. However, rather than specify individual-specific parameters, we typically estimate population level parameters that may vary based on sociodemographic characteristics. With this approach, both attributes of the alternatives z_{in} and characteristics of the decision-maker S_n are arguments to the utility function as follows:

$$U_{in} = U(z_{in}, S_n; \theta), \tag{3.38}$$

where S_n are the characteristics of the decision maker n. z_{in} now has a subscript n as well because the attributes of the alternative are specific to the choice context of person n. For example, in the commute mode choice decision, the travel times and travel costs of the alternative vary based on the home and work location of the commuter.

There is still one major aspect to operationalizing the analytical framework that is yet to be introduced, which is that of stochasticity. Thus far, we have assumed deterministic consumer behavior. We first work through in the next section a simple example of the deterministic case, and then introduce the random utility model in the section that follows.

3.6.1 Example with Two Alternatives

We now expand on the ideas laid out above of microeconomic consumer theory and extensions necessary for discrete choice analysis. We work through a simple case of 2 alternatives to review how the basic theory is applied in discrete choice analysis and to also introduce the concept of taste parameters and estimation.

Consider again the commuter mode choice problem depicted in Table 3.1. However, we simplify it even more for the case of this example. First, assume there are only two modes available, and that they are named *travel alternative 1* and *travel alternative 2* or *alternative 1* and *alternative 2* for short. Further, let's say there are only two attributes of relevance: travel time t and travel cost c. In this case, the interesting choice behavior arises when one has a choice of a cheaper but slower mode versus a faster but more expensive mode, and therefore in making the choice it is necessary to trade off time and cost.

Recall that the assumption of a separable utility allows us to exclude all other consumption in this example and focus solely on the decision of mode choice to work. Therefore, we start by writing the direct utility function (omiting index n to simplify the notations):

$$\widetilde{U} = \widetilde{U}(y_1, y_2), \tag{3.39}$$

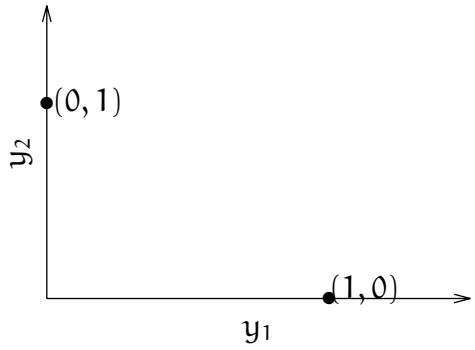


Figure 3.8: Choice set for the travel choice with two alternatives

where we impose the restrictions that

$$y_1 = \begin{cases} 1 & \text{if travel alternative 1 is chosen,} \\ 0 & \text{otherwise;} \end{cases}$$
$$y_2 = \begin{cases} 1 & \text{if travel alternative 2 is chosen,} \\ 0 & \text{otherwise;} \end{cases}$$

and that only one alternative is chosen:

$$y_1 + y_2 = 1. \tag{3.40}$$

Under these restrictions the utility function can attain only two possible values corresponding to the two points on Figure 3.8: $\widetilde{U}(1,0)$ and $\widetilde{U}(0,1)$. These are so-called "corner" solutions and are therefore points where the usual first-order conditions (the derivatives with respect to the quantities y_1, y_2 as in (3.19)) for an optimum do not hold.

Therefore, we apply the techniques described for discrete goods and work with the conditional indirect utility function of each good as in (3.37). The attribute in this case is travel time and a utility is defined for each of the modes in terms of time and cost as follows:

$$\begin{aligned} U_1 &= U(t_1, c_1; \theta), \\ U_2 &= U(t_2, c_2; \theta). \end{aligned} \tag{3.41}$$

The function $U()$, which maps the attributes values to a utility scale, is an ordinal utility function. It expresses mathematically the consumer's preferences and associates a real number (or, more precisely, a cardinal realization) with each possible alternative such that it summarizes the preference orderings (or rankings) of the consumer.

Utility maximization theorizes that travel alternative 1, for example, is chosen if

$$U_1 > U_2. \tag{3.42}$$

Indifference between two alternatives occurs when there is a tie or the difference between two utilities is smaller than some perceptual threshold level. For example, $U_1 = U_2$ means the commuter is indifferent between travel alternatives 1 and 2, and the choice is therefore indeterminate. We later resolve this potential problem by assuming that the probability of a tie is zero. When ties are ignored, travel alternative 1 is chosen if

$$U_1 \geq U_2. \tag{3.43}$$

To operationalize this model, it is necessary to determine the form of the utility function. This is one of the difficult assumptions to make, and a good portion of the book is concerned with this issue. For convenience, an additive utility function that is "linear in parameters" is most often assumed. In this text, we use β to denote such coefficients in the utility, and these are a specific kind (those that are linear in parameters) of the more general θ taste parameters used in the discussion thus far. For this example, such a utility function would take the form:

$$\begin{aligned} U_1 &= -\beta_t t_1 - \beta_c c_1, \\ U_2 &= -\beta_t t_2 - \beta_c c_2, \end{aligned} \tag{3.44}$$

where β_t and β_c (both > 0) are parameters that express the tastes of the commuter, that is, the commuter's sensitivity to time and cost. Note the signs, indicating that as either time or cost of an alternative increases, the utility of that alternative decreases.

Such a model can be used to predict how a commuter responds to changes in travel time or cost of travel alternatives. However, application requires that we assign numerical values to the parameters β_t and β_c. How do we assign numerical values? It is possible to infer the value of the parameters (or, more generally, information about the individual utility function) by observing choices made by an individual. In our example, we would need information on travel mode choices made by a commuter and the travel times and costs that exist in his or her choice environment. For most of the book, we discuss population parameters inferred from samples of decision-makers, however here we stick closer to the microeconomic framework of examining an individual's utility.

We describe two different ways of making this inference. The first is to design a sequential experiment that gradually hones in on the values of the

parameters. The second is the case where we have a set of choices from the individual to work with. In both cases, assume that we are using data from an experiment in which we present to an individual a hypothetical situation that presents two travel alternatives, each with different travel times and travel costs, and then the individual is asked which alternative he or she would choose. Of course this inference can also be done with real world choices, but in the case of mode choice to work, we likely would not observe enough variability to infer the parameters at an individual level.

The objective is to find the values of the parameters such that the chosen mode always has the highest utility. Since the utilities are ordinal, we can divide them by the positive constant β_c without mathematically modifying the preferences:

$$\begin{aligned} U_1 &= -\tfrac{\beta_t}{\beta_c} t_1 - c_1, \\ U_2 &= -\tfrac{\beta_t}{\beta_c} t_2 - c_2, \end{aligned} \tag{3.45}$$

While this changes the scale of the utility, it does not alter the rankings captured by the utilities. Thus, we are able to reduce the dimension of the search to finding a single parameter β, which is the ratio of our original parameters. The utilities are then:

$$\begin{aligned} U_1 &= -\beta t_1 - c_1, \\ U_2 &= -\beta t_2 - c_2, \end{aligned} \tag{3.46}$$

where $\beta = \beta_t/\beta_c$.

The choice situation is only interesting when neither alternative dominates the other, that is there is not an alternative that is both faster and cheaper. Further, without loss of generality, we assume that the first alternative is always slower and less expensive than the second alternative (and vice versa), that is

$$t_1 - t_2 > 0 \text{ and } c_2 - c_1 > 0. \tag{3.47}$$

The choice situation then comes down to whether or not the traveler is willing to pay the extra money $c_2 - c_1$ to save the extra time $t_1 - t_2$. If so, she chooses alternative 2, if not she chooses alternative 1. In making this choice, she reveals to us something about how much she values her time. This trade-off between money and time is called the value of time.

Now that we have the problem defined, on to the sequential experiment approach. First we offer a hypothetical set of alternatives to the individual. Say he or she chooses alternative 1 (the slower and less expensive alternative) over alternative 2. This implies that $U_1 \geq U_2$ or

$$-\beta t_1 - c_1 \geq -\beta t_2 - c_2. \tag{3.48}$$

Rearranging terms leads to the following lower bound for the parameter β :

$$\beta \leq \frac{c_2 - c_1}{t_1 - t_2}. \tag{3.49}$$

Since the traveler chose the slower and less expensive alternative, we know she is not willing to pay $c_2 - c_1$ to save time $t_1 - t_2$. This trade-off is the ratio on the right of the equation. The taste parameter β is shown mathematically to be below this trade-off. To further refine our knowledge on the taste parameter, we need to observe more choices. For example, say we now present the individual with the same alternative 1 but an improved alternative 2 that has a decreased travel time $t_2{}^*$ (with the same cost as before), and now the individual selects alternative 2. With the improved travel time of alternative 1, she is now willing to pay $c_2 - c_1$ to save time $t_1 - t_2$. This implies $U_1 \geq U_2{}^*$ and, following the same procedure above, we now have information on both the upper bound (from the first choice) and the lower bound (from this choice) on the parameter β:

$$\frac{c_2 - c_1}{t_1 - t_2{}^*} \leq \beta \leq \frac{c_2 - c_1}{t_1 - t_2} \tag{3.50}$$

More observations (changing travel times and travel costs and observing her choice) produce other inequalities that further narrow down the range of possible values of β. The parameter β is thus the value of time of the traveler, and by observing choices we develop an estimate of this value of time. Note that the value of time is a marginal rate of substitution (MRS) as introduced above, but here it is a trade-off in the attribute space (*time* versus *cost*), and the discussion above concerned trade-offs in the quantity space (mixture of q_1 versus q_2 in a consumption bundle). Nonetheless, the calculation is analogous to (3.16) but calculated from the conditional indirect utilities $U = -\beta_t t - \beta_c c$ and the MRS of time for money is:

$$MRS = \frac{\partial U(t,c)/\partial t}{\partial U(t,c)/\partial c} = \frac{\beta_t}{\beta_c} = \beta, \tag{3.51}$$

Further note that the units are consistent with the notion of value of time. As the utility equation has units of *utils*, $\beta_t t$ and $\beta_c c$ also has units of *utils*, and therefore β_t has units of *utils/time* and β_c has units of *utils/money*. Therefore, β has units of *money/time*, which is expected for a value of time. A value of time of $\$20/\mathrm{hour}$ indicates that one is willing to pay \$20 to save an hour of time.

For this analysis, we have implicitly assumed that transitivity of preferences (defined above) holds for the individual. If this weren't the case, we

could obtain inconsistent inequalities as we attempt to narrow the range of the parameter. We consider this possibility in the graphical example that is considered next.

The other approach to infer information about the utility from an individual's choices is to examine a set of observed choices all at one time. The experiment setup is similar to that above, except we ask many different choice questions and do the analysis on the entire group of choices rather than narrowing the estimation band. The information contained in choices made by an individual is plotted graphically in Figure 3.9. Each point on this plot represents a choice made by the individual, and the location of the point represents the hypothetical travel times and travel costs presented for the choice. We saw above that the problem can be reduced to the impact of the difference in travel time $(t_1 - t_2)$ and difference in travel cost $(c_1 - c_2)$, and so these choices are plotted in this difference space. The symbol of the point signifies whether she chose alternative 1 or alternative 2 for the specific value of time and cost differences presented.

In looking at this figure, there are two quadrants that are not very interesting: alternative 2 is dominant in the upper right quadrant because it has a faster time and a lower cost than alternative 1, and alternative 1 is dominant in the lower left quadrant because it is fastest and cheapest. The interesting parts of the figure are the other two quadrants, when the individual must make a trade-off between travel time and cost. Does she choose the faster but more expensive alternative? Or the cheaper but slower alternative? The answer depends on how much cheaper/expensive and how much faster/slower the alternatives are *and* on her value of time, i.e. the parameter β.

For points in the bottom right quadrant that are very close to the horizontal axis $t_1 - t_2$, we would expect the traveler to choose alternative 2 as it is hardly more expensive than alternative 1 but it is significantly faster. For points in the bottom right quadrant that are very close to the vertical axis $c1 - c2$ we would expect the traveler to choose alternative 1 as it is hardly slower than alternative 2 and significantly cheaper. The empirical question is where is the transition point in this quadrant between choosing alternative 2 and choosing alternative 1. This transition point happens when the traveler is indifferent between the alternatives, that is when $U_1 = U_2$ or $-\beta t_1 - c_1 = -\beta t_2 - c_2$. Rearranging terms, this equality becomes $(c_1 - c_2) = -\beta\,(t_1 - t_2)$, which is a line through the origin of the figure with slope $-\beta$. Note that this results in an equation analogous to (3.49), but with an equal sign to reflect the point of indifference. The parameter β can then be determined by pivoting the line through the origin until all the points where alternative 2 were chosen are above the line and all the points where alternative 1 were chosen are below the line. The fitted line has slope of

$-\beta_t/\beta_c$, or the negative of the value of time. If such a β can be found, it is an indication that preference transitivity holds. If, however, the fitted line cannot definitively divide alternative 1 choices from alternative 2 choices, then the transitivity assumption is violated.

The fact that the fitted line does not perfectly divide the chosen alternatives reflects that the model is not exact. We address this issue in the next section by introducing stochasticity into the framework.

3.6.2 Generalization to Many Alternatives and Many People

Above we worked through a simple example of two alternatives and two attributes for a single individual. In more general terms we consider a universal set of alternatives, denoted $\mathcal{C}$ and a sample of individuals $n = 1, \ldots, N_s$. The constraints faced by an individual decision maker n determine his or her choice set $\mathcal{C}_n \subseteq \mathcal{C}$. The simple example had 2 travel alternatives in the choice set, both of which were feasible for the traveler. As in consumer theory the individual is assumed to have consistent and transitive preferences over the alternatives that determine a unique preference ranking. Thus a real-valued conditional indirect utility index associated with every alternative can be defined,

$$U_{in},\ i \in \mathcal{C}_n \tag{3.52}$$

such that alternative $i \in \mathcal{C}_n$ is chosen if and only if

$$U_{in} > U_{jn},\ \text{all } j \neq i, j \in \mathcal{C}_n. \tag{3.53}$$

We define the utility function in terms of attributes (including cost)

$$U_{in} = U(z_{in}; \theta),\ \text{for all } i \in \mathcal{C}_n \tag{3.54}$$

where z_{in} is a vector of the attribute values for alternative i as viewed by decision maker n and θ a vector of parameters representing tastes. The simple example had attributes of time and cost and two parameters β_c and β_t. Income and time budgets and other external restrictions determine the choice set $\mathcal{C}_n$. While above we examined the choices of a single decision-maker, generally in empirical applications we consider parameters that are applicable to a population. Because tastes vary across people within the population, we introduce into the utilities a vector of socioeconomic characteristics that explains the variability of tastes across the portion of the population to which our model of choice behavior applies. Thus we write

$$U_{in} = U(z_{in}, S_n; \theta), \tag{3.55}$$

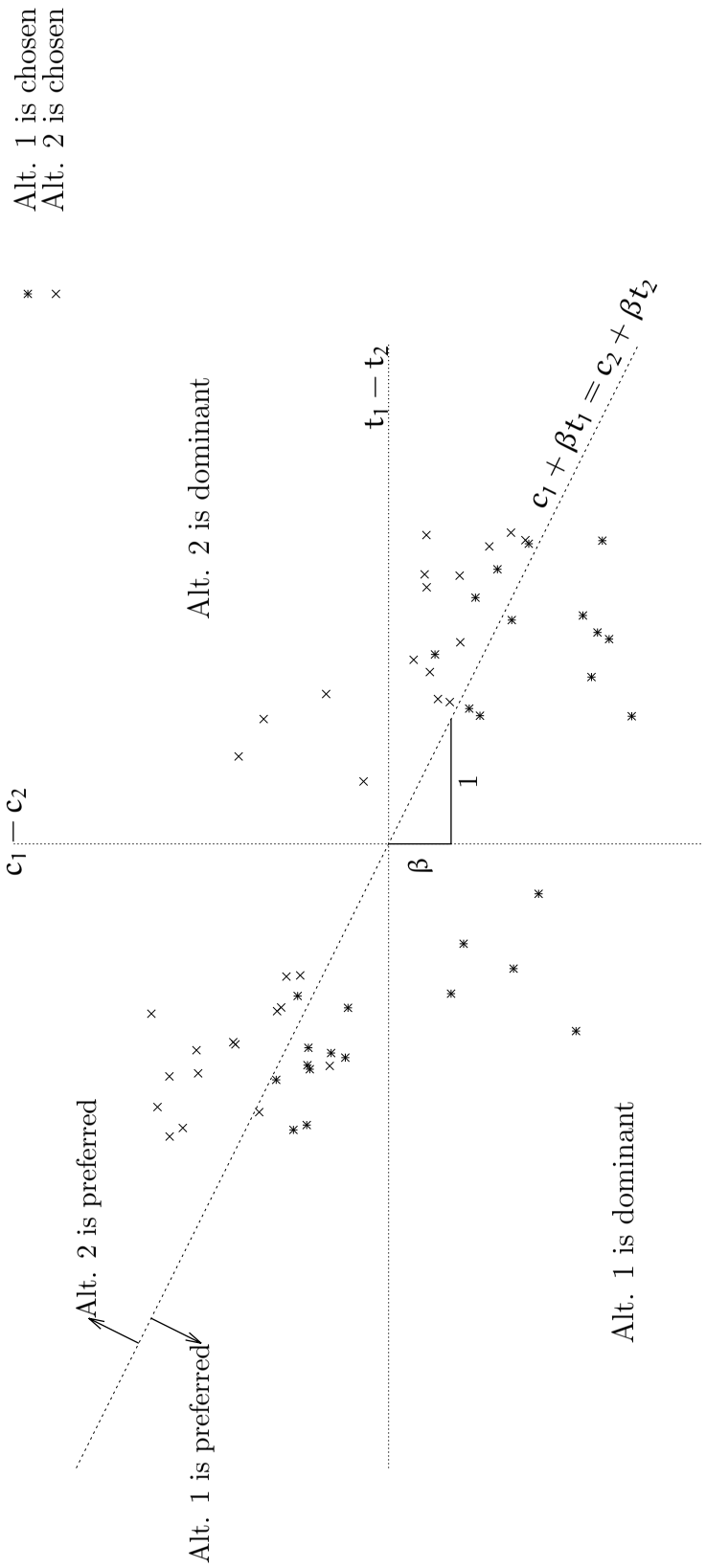


Figure 3.9: Graphic example of parameter estimation

where S_n is a vector of characteristics of the decision maker n such as income, age, and education. In the simple example we may introduce income of the decision maker I_n in the utility equation as follows:

$$U_{in} = -\beta_t t_{in} - (\beta_c / I_n)\, c_{in} \tag{3.56}$$

to reflect decreasing sensitivity to cost as income increases. Note that here we are using income as a proxy to capture taste heterogeneity in the cost parameter, rather than as an explicit element of the budget constraint.

Despite this generalization to more alternatives, more attributes, potentially non-linear utility functions, and the introduction of socio-demographics in the utility equation, conceptually the process of structuring the problem (conditional utilities that are a function of attributes) and inferring the parameters of the utility equation is just as described above for the simple two alternative travel example. We collect empirical data on the choices that people make along with the choice environment (choice set $\mathcal{C}_n$ and attributes of the alternatives z_{in}) and the characteristics of the decision-makers S_n. Assuming a utility maximizing decision process, the observed choices provide information on the taste parameters in the utility function, i.e. on the compensatory trade-offs between the attributes of alternatives as reflected by the MRS between attributes. Of course with the generalization we can no longer graphically demonstrate the process, and in Chapter 4 we discuss a procedure called maximum likelihood estimation to estimate the parameters from the data.

3.7 Probabilistic Choice Theory

Thus far, the theoretical discussion has assumed that consumers behave deterministically. Decision-makers are assumed to be all knowing with perfect discriminatory power, able to process information, choose the best choice, and repeat this identical choice under identical circumstances. This requires that certain properties of preferences hold, including completeness, transitivity, and continuity. However, there are copious examples both in laboratory experiments and in the field in which it appears that decision-makers do not behave as such. As Tversky (1969) points out, "when faced with repeated choices between x and y, people often choose x in some instances and y in others." Inspired by the need to explain experimental observations of inconsistent preferences, probabilistic choice theory was developed. In probabilistic choice theory, rather than assuming there is a deterministic process that can be used to establish the choice outcome, it is recognized that the best that can be done is to determine the probability of different choice outcomes

given a particular choice situation and decision-maker. Probabilistic choice models originated in psychology with Thurstone (1927), and were further developed by Luce (1959), Marschak (1960), and McFadden (1974).

There are two distinctly different ways of thinking about probabilistic choice, which diverge in terms of the assumptions made regarding the *source* of the stochasticity. Luce and Suppes (1965) illuminate this dichotomy and Anderson et al. (1992) provide a more recent discussion. One viewpoint is that human behavior is inherently probabilistic. Given repeated and identical choice situations, a decision-maker can and will make different choices in different instances. This approach is called the *constant utility approach*, because it is assumed the utilities of the alternatives are fixed and known. However, the consumer does not maximize her utility when making a choice, but behaves with choice probabilities defined by a probability distribution function over the alternatives. This probability distribution function includes the utilities as parameters. The probability distribution is such that the consumer has the highest probability of choosing the alternative that maximizes her utility, but also has non-zero probability of choosing alternatives that are suboptimal.

The other viewpoint is that the source of the stochasticity is due to errors made by the analyst in developing the model. Here the assumption is that while humans *are* deterministic and rational utility maximizers, analysts are unable to understand and model fully all of the relevant factors that affect human behavior. The individual is assumed to be all knowing and rational and select the alternative with the highest utility. However, the utilities are not known to the analyst with certainty and are therefore treated by the analyst as random variables. This is called the *random utility* approach. The value of the random utility approach is that it provides a link with behavioral theory from microeconomics and therefore a link to the concepts and methods that are useful for both developing model specifications and using the models for analysis.

Anderson et al. (1992) provide an extensive discussion on these two ways of thinking. They emphasize that in the random utility paradigm, the utility is stochastic and the decision rule is deterministic; whereas, in the constant utility paradigm, the utility is constant and the decision rule is stochastic. They liken random utility theory to Einstein's position that "God does not play dice," that is, everything can be explained with physical principles although we may not be aware of the explanation now. And constant utility is more aligned with Heisenberg's uncertainty principle in that there is inherent stochasticity in the actions of nature that are impossible to overcome. An example provided is that of a person visiting a wine store on two separate but seemingly identical occasions, and buying a Cabernet on one visit and

a Merlot on the other. The conjecture of a random utility theorist would be that even though the two purchase instances may appear to be identical, there is something unobservable that led to the cabernet having a higher utility on one trip and the Merlot the higher utility on a different trip. That is, in both instances the consumer maximized her utility, but the analyst may not be able to fully observe or measure the reason. This viewpoint is more consistent with consumer theory in which the utility maximization principle is paramount. The conjecture of a constant utility theorist would be that the utility of the two bottles of wine do not change between the visits nor does anything about the choice situation, it is just that a given individual may in some instances choose Cabernet and in other instances Merlot. The outcome probabilities represent the share of the time that Merlot is chosen by the individual versus the share of the time Cabernet is chosen over the repeated, identical choice situations. If the utility of Merlot is higher than the utility of Cabernet for an individual, then the individual more frequently chooses the Merlot. The key is that the utilities are fixed.

While both viewpoints result in probability equations that explain choice (and often analogous probability equations), one way in which the dichotomy manifests itself is in the way the probability equations are derived and discussed. In a random utility framework, the focus is on consumer theory and utility maximizing behavior and the starting point is the formulation of the distribution of the random utilities (that is, their errors). Assumptions on the distribution of the errors are combined with the assumption of utility maximizing behavior in order to derive the choice probability. In this realm, concepts regarding the error distribution such as the correlation (or lack thereof) between utility errors are discussed. Such assumptions may lead to mathematical properties of the probability equation that are of interest and analyzed, but the beginning point is the error distribution of the utilities.

In a constant utility framework, the focus is more on the mathematical properties of the choice probabilities themselves, which are a function of (or may considered being a function of) the utilities of alternatives. The utility function is constant in this framework and so there are no random errors and therefore the concept of the distribution of errors does not exist. Simple examples of important mathematical properties of the choice probabilities are that they lay between 0 and 1; sum to one across all alternatives; and as the utility of an alternative increases, the probability of that alternative increases. In this framework, emphasis is placed on the substitution patterns among alternatives and the related mathematical properties that deal with substitution among alternatives such as cross elasticities and odds-ratios.

Identical choice models can be derived from either framework. For example, Luce (1959) used a constant utility framework to derive the logit model

(see Chapter 5) by starting from assumptions about choice probabilities (the independence of irrelevant alternative property), whereas McFadden (1974) derived the logit model from the random utility framework by making an assumption on the distribution of the errors of the utilities (the Extreme Value distribution). Similarly, the nested logit model that is introduced in Chapter 7 was originally derived using a constant utility (or mathematical property) approach in Ben-Akiva (1973). The logsum term that connects the different levels of the nests was selected from among various contenders because it had the correct mathematical properties. McFadden (1978) later derived the same nested logit formulation from the perspective of random utility applying the Multivariate Extreme Value distribution to the random errors of the utilities. This is discussed in detail in Chapter 7.

The viewpoints cross interdisciplinary borders. The constant utility framework is more common in psychology, (see Luce and Suppes, 1965 for a review of the early work in this area) but also espoused early on by the economist Quandt (1956). The random utility framework is more common in economics, but was used by psychologist Thurstone (1927) (and later formalized as random utility by Marschak, 1960) in his development of what we now call the binomial probit model (see Chapter 4).

While this discussion is useful in terms of providing a backdrop and framework for thinking about choice behavior, neither the constant utility nor the random utility framework is essential to the methods presented in this book. At its most abstract level, the methods can be viewed purely as statistical tool used to fit to data without any behavioral underpinning at all. Nonetheless, because of the close associations of both random utility theory and constant utility theory to discrete choice analysis, we elaborate further on the mathematics of both of these veins below.

3.7.1 The Random Utility Model

As described above, the random utility model is more in line with consumer theory. The observed inconsistencies in choice behavior are taken to be a result of observational deficiencies on the part of the analyst. The individual is always assumed to select the alternative with the highest utility, which can be written as:

$$\mathsf{P}(\mathfrak{i}|\mathcal{C}_\mathfrak{n}) = \Pr(\mathsf{U}_{\mathfrak{i}\mathfrak{n}} \geq \mathsf{U}_{\mathfrak{j}\mathfrak{n}}, \text{ all } \mathfrak{j} \in \mathcal{C}_\mathfrak{n}), \tag{3.57}$$

where $\mathsf{U}_{\mathfrak{i}\mathfrak{n}}$ is the utility of alternative $\mathfrak{i}$ for person $\mathfrak{n}$. However, the utilities are not known to the analyst with certainty and are therefore treated by the analyst as random variables. In the random utility approach we derive choice probabilities by assuming a joint probability distribution for the set

of random utilities U_{in}.

We express the random utility of an alternative as a sum of observable or systematic components (denoted as V_{in}) and unobservable components (denoted as ε_{in}) of the total utilities as follows:

$$U_{in} = V_{in} + \varepsilon_{in}. \tag{3.58}$$

As described above, we use Lancaster's approach to describe the utility by attributes and we capture taste heterogeneity by including characteristics. Therefore, both V_{in} and ε_{in} may be a function of attributes z_{in} and characteristics S_n. As discussed above, the random component is necessary to capture deficiencies in the specification.

In formalizing the structure of random utility models, Manski (1977) cites four distinct sources of analyst error that lead to randomness: (i) unobserved attributes, (ii) unobserved taste variation, (iii) measurement errors and imperfect information, and (iv) instrumental or proxy variables. Analyzing these cases more formally, *unobserved attributes* is the situation where the known vector of attributes affecting the decision z_{in} is missing one or more attributes $\breve{z}_{in}$. So the utility is then $U_{in} = U(z_{in}, S_n, \breve{z}_{in}; \theta)$. As $\breve{z}_{in}$ is unobserved and therefore a random variable, the utility is also a random variable. *Unobserved taste variation* is a similar situation except that the unobserved argument is an individual characteristic $\breve{S}_n$, a random variable, and the utility is then $U_{in} = U(z_{in}, S_n, \breve{S}_n; \theta)$, also a random variable. *Measurement errors* is the situation where the precise value of an attribute $\breve{z}_{in}$ is not observable and an imperfect measure z_{in} is used instead, where $\breve{z}_{in} = z_{in} + \varepsilon_{in}$. The imperfect measurement is used in the utility function which is then $U_{in} = U(z_{in} + \varepsilon_{in}, S_n; \theta)$, and the utility is a random variable because ε_{in} is a random variable. *Instrumental variables* are used as proxies when an element of the utility is not observable. Rather than excluding the variable from the utility function (as with unobserved attributes or unobserved taste variation), the problem variable is replaced with a function of instrumental variables (i.e., variables related to the actual attributes). For example a function $g(z_{in})$ is used to replace the unobserved $\breve{z}_{in}$. As the relationship is imperfect, the utility function then becomes $U_{in} = U(g(z_{in}) + \varepsilon_{in}, S_n; \theta)$, which is a random variable due to ε_{in}.

Regardless of the source of the random error, the idea is that ε_{in} exists as depicted in (3.58) due to specification errors of the analyst. Substituting (3.58) into (3.57) leads to

$$P(i|\mathcal{C}_n) = \Pr(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}, \text{ all } j \in \mathcal{C}_n). \tag{3.59}$$

Rearranging terms leads to the following expression:

$$P(i|\mathcal{C}_n) = \Pr(\varepsilon_{jn} - \varepsilon_{in} \leq V_{in} - V_{jn}, \text{ all } j \in \mathcal{C}_n). \tag{3.60}$$

Note this is a cumulative distribution function of the joint distribution of error differences $\varepsilon_{jn} - \varepsilon_{in}$. A fully specified model requires specification of the V_{in} and assumptions on the joint distributions of the random vector of disturbances $\varepsilon_n = (\varepsilon_{1n}, \ldots, \varepsilon_{J_n n})$ or on the joint distribution of the vector of differences. The derivation of the choice model from distributional assumptions on the error term is not straightforward in general. In the following chapters, we discuss concrete models which are particularly useful for applications. In general, if the cumulative distribution function of the error term is available in a closed form, the choice model is easier to derive. Indeed, it can be shown (see Appendix 3.A) that, if ε_n is a multivariate random variable with CDF $F_{\varepsilon_n}(\varepsilon_1, \ldots, \varepsilon_{J_n})$ and pdf

$$f_{\varepsilon_n}(\varepsilon_1, \ldots, \varepsilon_{J_n}) = \frac{\partial^{J_n} F}{\partial \varepsilon_1 \cdots \partial \varepsilon_{J_n}}(\varepsilon_1, \ldots, \varepsilon_{J_n}), \tag{3.61}$$

then

$$P_n(i|\mathcal{C}_n) = \int_{\varepsilon=-\infty}^{+\infty} \frac{\partial F_{\varepsilon_{1n}, \varepsilon_{2n}, \ldots, \varepsilon_{J_n}}}{\partial \varepsilon_i}(\ldots, V_{in} - V_{(i-1)n} + \varepsilon, \varepsilon, V_{in} - V_{(i+1)n} + \varepsilon, \ldots) d\varepsilon, \tag{3.62}$$

and the choice model is obtained by solving a uni-dimensional integral, which can be done analytically for simple models, and numerically for more complex ones. If the CDF is not available in a closed form, the choice model is expressed as a multi-dimensional integral, which can be cumbersome to evaluate when the number of alternatives is more than a handful, as discussed in Appendix 3.B.

Different assumptions on the error distribution leads to various choice models. The output of a random utility model is the probability of an individual selecting each alternative.

For example, the assumption that the errors are independently and identically distributed Extreme Value (McFadden, 1974) leads to the logit probability model:

$$P(i|\mathcal{C}_n) = \frac{e^{\mu V_{in}}}{\sum_{j \in \mathcal{C}_n} e^{\mu V_{jn}}}, \text{ for all } i \in \mathcal{C}_n \tag{3.63}$$

where μ is a scale parameter (a parameter of the error distribution). Note that the reason why this model is sometimes referred to as *conditional* logit is due to the derivation from consumer theory that makes use of the *conditional* indirect utility equation (3.31). In the random utility framework, the focus is on the random utilities and their variances and covariances. The error distribution assumption for logit is fairly simplistic, with constant variance ($\text{Var}(\varepsilon_{jn}) = \pi^2/6\mu^2$, all $j \in \mathcal{C}_n$) and covariances equal to zero

($\text{Cov}(\varepsilon_{in}, \varepsilon_{jn}) = 0$, all $i \neq j$). This utility is an ordinal utility in that it only ranks alternatives, and can be transformed by any order preserving transformation and still result in the same probability equation. For example, multiplying the U_{in} in (3.57) by apositive constant and adding a constant results in the identical probability as shown in (3.63). Chapter 5 covers the logit model in detail (the simpler 2 alternative logit is covered in Chapter 4) and alternative distributional assumptions, specification of the systematic utilities, and the resulting probabilistic choice models are considered in great detail in the following chapters.

From the resulting choice probability equation (either logit or other) we can examine its properties discussed in the next section. We can also aggregate individual probabilities to produce demand functions as well as other demand analysis such as elasticities and consumer surplus (see Chapter 10).

3.7.2 Properties of Probability Models

While the random utility approach begins with a focus on the specification of the random utility as described in the section above, the constant utility approach begins with a focus on the properties of the probability equations and what criterion the model needs to satisfy to be a proper probability. In this section we review some of these properties. Recall that in the constant utility approach, the utilities of the alternatives are fixed. Instead of selecting the alternative with the highest utility, the decision-maker is assumed to behave with choice probabilities defined by a probably distribution function over the alternatives that includes the utilities as parameters.

Denote the probability of decision maker n choosing alternative i as

$$P_n(i), \tag{3.64}$$

or as

$$P(i|\mathcal{C}_n), \tag{3.65}$$

when an explicit notation for the choice set is also useful. Obviously we require that the probabilities are between 0 and 1:

$$0 \leq P(i|\mathcal{C}_n) \leq 1, \tag{3.66}$$

where the equality signs hold in the limiting case of a deterministic choice. Further that the probabilities summed over the available choices equals 1:

$$\sum_{i \in \mathcal{C}_n} P(i|\mathcal{C}_n) = 1. \tag{3.67}$$

The usual theorems of probability theory are assumed to hold. In particular, since one and only one alternative is chosen, the following must hold:

$$\mathsf{P}(\mathfrak{i} \text{ and } \mathfrak{j}|\mathcal{C}_n) = 0, \mathfrak{i} \neq \mathfrak{j} \in \mathsf{C}_n, \text{ and} \tag{3.68}$$

$$\mathsf{P}(\mathfrak{i} \text{ or } \mathfrak{j}|\mathcal{C}_n) = \mathsf{P}(\mathfrak{i}|\mathcal{C}_n) + \mathsf{P}(\mathfrak{j}|\mathcal{C}_n), \mathfrak{i} \neq \mathfrak{j} \in \mathsf{C}_n. \tag{3.69}$$

Generally, for any subset of the choice set $\widetilde{\mathcal{C}}_n \subseteq \mathcal{C}_n$, we can write

$$\mathsf{P}(\widetilde{\mathcal{C}}_n|\mathcal{C}_n) = \sum_{\mathfrak{i} \in \widetilde{\mathcal{C}}_n} \mathsf{P}(\mathfrak{i}|\mathcal{C}_n), \tag{3.70}$$

expressing the probability that the choice lies in the subset $\widetilde{\mathcal{C}}_n$.

Further, using the standard definition of conditional probability, it is also possible to calculate the following conditional probability:

$$\mathsf{P}(\mathfrak{i}|\widetilde{\mathcal{C}}_n \subseteq \mathcal{C}_n) = \frac{\mathsf{P}(\mathfrak{i}|\mathcal{C}_n)}{\mathsf{P}(\widetilde{\mathcal{C}}_n|\mathcal{C}_n)}, \tag{3.71}$$

for any alternative $\mathfrak{i} \in \widetilde{\mathcal{C}}_n \subseteq \mathcal{C}_n$, provided that $\mathfrak{p}(\mathfrak{j}|\mathcal{C}_n) > 0$ for at least one alternative $\mathfrak{j} \in \widetilde{\mathcal{C}}_n$.

And there are other general rules that should apply to the probability function for it to behave as one would expect. For example if the utility of alternative $\mathfrak{i}$ improves and there are no other changes to the choice set, then the probability of $\mathfrak{i}$ should not get worse and the probability of the other alternatives should not get better. That is:

$$\begin{array}{rcl} \partial \mathsf{P}(\mathfrak{i}|\mathcal{C}_n)/\partial \mathsf{U}_{\mathfrak{i}\mathfrak{n}} & \geq & 0, \\ \partial \mathsf{P}(\mathfrak{j}|\mathcal{C}_n)/\partial \mathsf{U}_{\mathfrak{i}\mathfrak{n}} & \leq & 0 \text{ for all } \mathfrak{i} \neq \mathfrak{j} \in \mathcal{C}_n. \end{array} \tag{3.72}$$

If this is not the case, then the choice model can produce nonsensical results.

The above conditions should hold for *any* choice model in order to follow basic rules of probability theory. However, there are other conditions that apply only to some models. One such important and well known condition is now called *independence from irrelevant alternatives* or IIA. The IIA condition can be stated in many different equivalent ways. One statement of IIA is that the relative choice probabilities between any two alternatives is independent of the other available alternatives, which is written as follows:

$$\frac{\mathsf{P}(\mathfrak{i}|\mathcal{C}_n)}{\mathsf{P}(\mathfrak{j}|\mathcal{C}_n)} = \frac{\mathsf{P}(\mathfrak{i}|\widetilde{\mathcal{C}}_n)}{\mathsf{P}(\mathfrak{j}|\widetilde{\mathcal{C}}_n)}, \mathfrak{i}, \mathfrak{j} \in \widetilde{\mathcal{C}}_n \subseteq \mathcal{C}_n \tag{3.73}$$

That is, when calculating the ratio of probabilities between $\mathfrak{i}$ and $\mathfrak{j}$, the outcome is independent of what other alternatives (beyond $\mathfrak{i}$ and $\mathfrak{j}$) are included

in the choice set. In (3.63) we presented the classic logit model, highlighting its derivation from random utility theory as in McFadden (1974). Luce (1959) derived the logit model using a constant utility approach, starting with the IIA assumption. He referred to it as the *choice axiom* and motivated it behaviorally as a probabilistic version of the concept of transitivity. Luce stated that a set of choice probabilities satisfies the choice axiom if for all i, $\widetilde{C}_n$, and C_n such that $i \in \widetilde{C}_n \subseteq C_n$, the following holds

$$P(i|\widetilde{C}_n \subseteq C_n) = P(\widetilde{C}_n|C_n), \tag{3.74}$$

whenever the conditional probability exists. In other words, if some alternatives are removed from a choice set, the conditional choice probabilities from the reduced choice set are unchanged. The choice probabilities from a subset of alternatives is dependent only on the alternatives included in this subset and is independent of any other alternatives that may exist. This is simply IIA stated a different way. Luce (1959) proved that if the choice axiom holds and a utility measure is directly proportional to the choice probability, then there exists a *strict utility* model as follows:

$$P(i|C_n) = \frac{U_{in}}{\sum_{j \in C_n} U_{jn}}, \text{ all } i \in C_n \tag{3.75}$$

where the utilities are now restricted to be positive and must be defined on a ratio scale that is unique up to multiplication by a positive constant. (See Anderson et al. (1992) for a presentation of Luce's proof.) Note that this equation also applies to any subset $\widetilde{C}_n \subseteq C_n$. This is essentially the same as (3.63) with Luce's utility function defined to be $\exp(V_{in})$.

Luce, in the constant utility framework, describes the IIA property behaviorally as a probabilistic version of the concept of transitivity and from there derives the choice model. Whereas in the random utility framework, the IIA property is a result of assumptions on the distribution of the errors and in particular is a result of assumptions where the errors are independently and identically distributed.

Luce saw IIA as a positive property, because it greatly simplified data collection for behavioral experiments by allowing choice probabilities in multiple alternative settings to be estimated from choice experiments that only address two alternatives at a time. However, Debreu (1960) pointed out that the validity of Luce's choice axiom in terms of correctly representing behavior depends on the structure of the choice set. He pointed out that a model with the IIA property performs poorly when there are some alternatives that are very similar to others. Take the commuter mode choice as an example where

there are two options: car and bus and the choice probabilities are

$$\begin{aligned} P_n(\text{car}) &= 1/2 \\ P_n(\text{bus}) &= 1/2 \end{aligned} \tag{3.76}$$

Now suppose that another bus service is introduced that is equal in all attributes to the existing bus service except that its buses are painted differently. We now have red and blue buses as two of the available alternatives. Under the choice axiom, the ratio of choice probabilities is constant, and therefore the new choice probabilities are

$$\begin{aligned} P_n(\text{car}) &= 1/3 \\ P_n(\text{red bus}) &= 1/3 \\ P_n(\text{blue bus}) &= 1/3. \end{aligned} \tag{3.77}$$

This is unrealistic because the commuter in reality is most likely to treat the two bus modes as a single alternative and behave with the following choice probabilities

$$\begin{aligned} P_n(\text{car}) &= 1/2 \\ P_n(\text{red bus}) &= 1/4 \\ P_n(\text{blue bus}) &= 1/4. \end{aligned} \tag{3.78}$$

Thus the choice axiom is invalid in this case. The model predicts that the new bus alternative (say the blue bus) would draw equally from those who drive and those who were on the red bus. However, behaviorally we would expect that the new blue bus would draw more heavily (or in this extreme case, solely) from those who were riding the red bus. The foregoing example reflects an extreme case where two alternatives in a choice set are for all practical purposes identical and should really be considered a single alternative. In general, the IIA condition (or choice axiom) is questionable when alternatives are perceived to be similar. For example, a more realistic mode choice scenario would be the addition of a subway alternative to a pre-existing situation of only car and bus alternatives. In this case, we would expect the subway alternative to draw from both the auto and the bus riders, but likely more from the bus riders.

The IIA assumption produces a lot of bad publicity for basic choice models such as logit. However, the condition is both a blessing and a curse. It is a blessing because there are very useful theoretical results that only are applicable when IIA holds. For example, if IIA holds then it is valid to estimate a choice model using a sample of alternatives (McFadden, 1978). This is extremely useful for large choice sets and also useful when data are not available on all of the alternatives. Also, even in cases where alternatives are

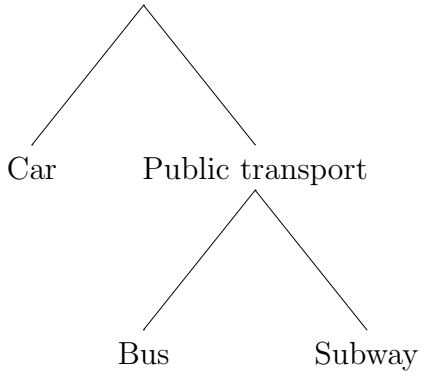


Figure 3.10: An example of a choice hierarchy

very similar, it is possible for IIA to hold if the systematic utility is well specified. In Chapter 6, statistical procedures are presented for testing whether or not IIA holds for a particular dataset and model. Further, the extreme results of the red bus/blue bus example shown above are less severe when considering population shares from models that include socio-demographics to capture those travelers who prefer auto and those who prefer transit (see Section 5.3.1). Finally, while IIA is a property of the simple logit model, there are many ways to relax the IIA assumption and many variations of discrete choice models that are covered in this text aim at doing just that.

As a simple example of how to modify a model that violates the IIA assumption, it is possible to introduce an assumption of a *choice hierarchy*. Take a case related to the red bus/ blue bus example, but more realistic where the two public transport alternatives are bus and subway. Figure 3.10 provides an example of what the choice hierarchy may look like. In this case we may assume that the choice axiom applies at each level of the hierarchy but it is not valid for the full set of alternatives. In other words, the choice axiom may be valid for a choice within a subset (e.g., bus versus subway), for a choice between subsets (e.g., bus and subway versus car), but it does not apply for the choice from the whole set of alternatives (e.g., car versus bus versus subway).

The problem for the analyst is to know whether a decision maker does indeed decompose a certain decision into a hierarchical choice process and, if so, what is the specific structure that he or she uses. In some cases it may appear as if there are "natural" partitions of the alternatives, as shown in the example in the figure. However, for more complex decisions there are many possible choice hierarchies, and different structures can yield different choice probabilities. Reasonable choice hierarchies may be postulated and

empirically tested against each other.

There are many ways to estimate models of such choice hierarchies (and thereby relax the IIA assumption), and many variations of discrete choice models covered in this text aim at doing just that. For example, Chapter 7 describes the nested logit model, which is one way of modeling the choice hierarchy as shown in Figure 3.10.

3.7.3 Expected Maximum Utility

As discussed above, the largest utility within the choice set represents the benefit obtained by the decision-maker $\mathfrak{n}$ from the chosen alternative. It may be useful to compare several choice situations, and identify the corresponding benefit of each of them. As the utility is a random variable, we are interested in the average of the "benefit", that is the average of the maximum utility. This quantity is called the *expected maximum utility* (EMU). It plays an important role in discrete choice analysis, in particular for policy implication, as discussed in details in Chapter 10. It also plays a role in the characterization of choice models, as discussed in Chapter 9.

The *expected maximum utility* is defined by

$$\mathrm{E}[\max_{\mathfrak{i}\in\mathcal{C}_\mathfrak{n}} \mathsf{U}_{\mathfrak{i}\mathfrak{n}}]. \tag{3.79}$$

It is a scalar summary of the expected "worth" of a set of alternatives. In the transportation context, it is sometimes called *measure of accessibility.*

In a more general context the systematic component of the maximum utility of all alternatives in a choice set is a measure of the individual's expected utility associated with a choice situation. This measure is individual specific, reflecting the differences in how various individuals evaluate their alternatives.

The value of this measure is not well defined unless some benchmark level of utility is established. For example, we have shown that the ordinal nature of utility is such that choice probabilities are unaffected by the addition of a constant to the utility of each alternative; the EMU, however, would be shifted upward by that constant. To be used properly, we must avoid comparing these measures using different models, and be careful to define the specification of the model we are using before interpreting those measures.

Two properties that EMU would ideally have are as follows:

1. *Monotonicity with respect to choice set size.* This property implies that any addition to a person's choice set leaves the individual no worse off than before the addition, namely

$$\mathrm{E}[\max_{\mathfrak{i}\in\mathcal{C}_\mathfrak{n}} \mathsf{U}_{\mathfrak{i}\mathfrak{n}}] \leq \mathrm{E}[\max_{\mathfrak{i}\in\mathcal{C}'_\mathfrak{n}} \mathsf{U}_{\mathfrak{i}\mathfrak{n}}], \tag{3.80}$$

where $\mathcal{C}_n \subseteq \mathcal{C}'_n$.

2. *Monotonicity with respect to the systematic utilities.* This property implies that the EMU does not decrease if the systematic utility of any of the alternatives in $\mathcal{C}_n$ increases. Mathematically this can be expressed as

$$\frac{\partial}{\partial V_{jn}} \mathrm{E}\left[\max_{i \in \mathcal{C}_n} U_{in}\right] \geq 0, \quad \text{for all } j \in \mathcal{C}_n. \tag{3.81}$$

The first property holds for all choice models in which the expected value in equation (3.80) exists for both $\mathcal{C}_n$ and $\mathcal{C}'_n$ (see Ben-Akiva and Lerman, 1979 and Williams, 1977 for proofs of this proposition.) The second property holds for *translationally invariant*[1] choice models.

Such models have interesting properties. Indeed, one can show that if the error terms do not depend on V_{in}, we have

$$\frac{\partial \, \mathrm{E}[\max_{j \in \mathcal{C}_n} U_{jn}]}{\partial V_{in}} = P_n(i), \ \forall i \in \mathcal{C}_n. \tag{3.82}$$

In words, equation (3.82) states that the derivative of the EMU with respect to the systematic component of the utility of any alternative is equal to that alternative's choice probability. To see that, we note that the derivative of the maximum utility $\max_{j \in \mathcal{C}_n} U_{jn} = \max_{j \in \mathcal{C}_n} V_{jn} + \varepsilon_{jn}$ with respect to the systematic utility V_{in} is 1 if alternative i corresponds to the maximum utility, and 0 otherwise, that is

$$\frac{\partial \max_{j \in \mathcal{C}_n} U_{jn}}{\partial V_{in}} = \frac{\partial \max_{j \in \mathcal{C}_n} V_{jn} + \varepsilon_{jn}}{\partial V_{in}} = \begin{cases} 1 & \text{if } U_{in} \geq U_{jn} \ \forall j \in \mathcal{C}_n, \\ 0 & \text{otherwise.} \end{cases} \tag{3.83}$$

This is true if no ε_{jn} depends on V_{in}. Equivalently, the derivative is the choice indicator, that is 1 if alternative i is chosen, and 0 otherwise. The expected value of the choice indicator is the choice probability $P_n(i)$, that is

$$\mathrm{E}\left[\frac{\partial \max_{j \in \mathcal{C}_n} U_{jn}}{\partial V_{in}}\right] = \mathrm{E}\left[\begin{cases} 1 & \text{if } U_{in} \geq U_{jn} \ \forall j \in \mathcal{C}_n, \\ 0 & \text{otherwise.} \end{cases}\right] = P_n(i). \tag{3.84}$$

The expectation operator being linear, we have that

$$\mathrm{E}\left[\frac{\partial \max_{j \in \mathcal{C}_n} U_{jn}}{\partial V_{in}}\right] = \frac{\partial \, \mathrm{E}[\max_{j \in \mathcal{C}_n} U_{jn}]}{\partial V_{in}}, \tag{3.85}$$

[1]Translationally invariant choice models are such that a shift in the systematic component of the utility just translates the joint distribution of the utilities without altering its basic functional form. Most choice models used in practice are translationally invariant, or can be transformed into an equivalent translationally invariant model, as discussed in Chapter 9.

and we obtain (3.82). For this reason, the expected maximum utility is called a *Choice Probability Generating Function*[2] (CPGF).

An obvious consequence of the foregoing result is that

$$\sum_{j \in \mathcal{C}_n} \frac{\mathrm{E}[\max_{i \in \mathcal{C}_n} U_{in}]}{\partial V_{jn}} = 1. \tag{3.86}$$

That is, if we increase the utility of every alternative by some amount ΔV, then the value of the EMU increases by that same amount. Another consequence of the result in equation (3.82) is that derivatives of any two choice probabilities with respect to the systematic utilities of the other alternatives are equal:

$$\frac{\partial P_n(j)}{\partial V_{in}} = \frac{\partial^2 \mathrm{E}[\max_{i \in \mathcal{C}_n} U_{in}]}{\partial V_{in} \partial V_{jn}} = \frac{\partial P_n(i)}{\partial V_{jn}}. \tag{3.87}$$

3.8 Beyond Rationality

Above we described decision-makers as being rational. That they have full information on all available alternatives and are able to accurately calculate and compare the value of options before choosing to follow the course of action that is best for them. However, human beings have cognitive limitations: our abilities as problem solvers are constrained by our limited information gathering and processing capabilities. Behavioral science researchers have a long history of raising serious questions about the rationality assumption. Their research has often succeeded in pointing out seemingly inconsistent and non-sensible choices. Other behavioral researchers aren't driven by refuting rationality, but are interested in understanding and influencing behavior, such as in psychology, marketing and the health science and in this quest have developed various theories of the behavioral process that are in many ways far from the economics-based paradigm emphasized above. In this section we provide a brief introduction to this vast literature, starting first with the literature more closely linked with the rationality discussion and following with a discussion of behavioral theories from other domains.

Amongst the earliest theories to diverge from rationality was proposed by Simon (1957). He recognized cognitive limitations that would preclude rational behavior and suggested the notion of *bounded* rationality as a more realistic representation. He postulated that individuals when faced with a complicated choice situation employ heuristics and rules of thumb to reduce

[2]To be precise, the choice probability generating function is the *exponential* of the expected maximum utility, as discussed in the next section.

the complexity of the original problem, and are subsequently rational within the resulting simplified framework. Thus bounded rationality regards individual decision-makers as "satisficers" - they seek a satisfactory solution rather than an optimal one.

Such heuristics and rules of thumb can manifest themselves in several different ways and can result in often surprising departures from perfect rationality. The psychologists Daniel Kahneman and Amos Tversky led early work in this area. Kahneman was awarded the 2002 Nobel Prize in Economics for their work (Kahneman, 2003), six years after the death of Tversky. Their experiments repeatedly demonstrated how the use of heuristics for making probability judgments under uncertainty could result in systematic errors, or cognitive biases. Examples of cognitive biases include *loss aversion*, which implies that the disutility of giving up an object is greater than the utility associated with acquiring it, or that individuals attach a higher value to a commodity that they already own than to an identical commodity that they do not; *anchoring*, which refers to situations where individuals disproportionately weigh one specific piece of information, and readjust or reinterpret other information to conform to the anchored information; and *framing*, which is when the same information when presented differently to the same individual can lead her to surmise a different conclusion. Numerous other cognitive biases have been identified, and their implications for decision-making theory comprise an ongoing field of research.

The work of Kahneman and Tversky forms the foundation of the field of behavioral economics, which is a blend of psychology and microeconomics. Behavioral economics has sought to bring greater psychological realism to neoclassical economic models of how individuals make decisions, and has attempted to expand our understanding of what it means to be rational. Kahneman and Tversky (2000) contains some of the seminal papers in the field and Ariely (2010) and Thaler and Sunstein (2009) provide particularly accessible introductions to more recent developments. They discuss issues such as lack of self-control and our all too frequent tendency to indulge in short-term behavior that is at odds with our long-term interests. The need for instant gratification can result in erratic and potentially deleterious actions, such as procrastinating, smoking, eating unhealthily, or not saving enough for retirement. On the positive side of human behavior, individuals often exhibit preferences that deviate from pure self-interest and are often rooted in considerations of fairness and the welfare of others. For example, experiments have repeatedly found that individuals when free to allocate money between themselves and others do not keep the money all to themselves. Evidence also shows that individuals are reciprocal: they are nicer to people who are nice to them, and unduly cruel to people who are cruel to them, above and

beyond what they should be if self-interest were their only concern. For example, seeking revenge even when there is no material gain to be had from its attainment would be an instance where reciprocity conflicts with the definition of perfect rationality.

Psychologists have long stressed the importance of the cognitive processes on choice behavior (Payne et al., 1992, and Prelec, 1991). Far from the concept of innate, stable preferences that are the basis of traditional discrete choice models, they emphasize the importance of experience and circumstances and a whole host of amorphous concepts, such as context, knowledge, point of view, degree of complexity, familiarity, risk of the choice at hand, and the use of non-utility maximizing decision protocols such as problem-solving, reason-based, and rule-driven processes. Explicit frameworks for modeling the behavioral choice process have been developed in a number of fields. The Theory of Planned Behavior (TPB, Ajzen, 1991) from psychology emphasizes the influences of attitudes, social norms, behavioral control and intention. The TPB framework has been extended in the Model of Goal-Directed Behavior (MGB, Perugini and Bagozzi, 2001) to include emotions, habits, and behavioral desire to the framework. Psychometricians, in their quest to understand behavioral constructs and causal relationships, have pioneered the use of psychometric data, for example, answers to direct survey questions regarding attitudes, perceptions, motivations, affect, etc. A general approach to synthesizing models with latent variables and psychometric-type measurement models has been advanced by a number of researchers including Keesling (1972), Joreskog (1973), Wiley (1973), and Bentler (1980).

In marketing, the emphasis of the behavioral process is on the dynamics of information acquisition and the influence of marketing triggers in product evaluation. For example, the Consumer Process Model (CDP, Blackwell et al., 2005) is a seven stage model that consists of need recognition, search, pre-purchase evaluation, purchase, consumption, post-consumption evaluation, and divestment. The health sciences have their own model consisting of behavioral stages, but their model is motivated by the desire to modify destructive behaviors such as drug use and poor nutrition. The Transtheoretical Model of Behavior Change (TTM, Prochaska and Velicer, 1997), also known as the stages of change, emphasizes that health behavior change involves a staged process which starts at pre contemplation, and moves through contemplation, preparation, action, maintenance and then either relapse or termination. Also focusing on behavior change, but more from a computer science orientation is the rapidly growing persuasive technology field. The Fogg Behavioral Model (FBM, Fogg, 2009) espouses that behavior change occurs when motivation, ability, and trigger occur at the same time and further elaborates on how to motivate, simplify (ability), and trigger people to

achieve the target behavior. And this is all just to name a few of the behavioral frameworks that have been developed. The angles by which one can study and structure human behavior are endless.

As implied by the discussion above, there is a large gap between behavioral theory by psychologists and behavioral researchers and discrete choice models. This gap arises from the difference in driving forces behind the two disciplines. While discrete choice modelers aim for operational, quantitative models of choice behavior, the profession would do well to study behavioral research and incorporate, where relevant, these ideas. Indeed, much of the motivation for the advances in discrete choice analysis that are presented in this text, are the result of efforts to improve the behavioral realism of quantitative models of choice.

McFadden (1999) provides a summary of behavioral science research from a discrete choice modeler's view. He argues that "most cognitive anomalies operate through errors in perception that arise from the way information is stored, retrieved, and processed" and that "empirical study of economic behavior would benefit from closer attention to how perceptions are formed and how they influence decision-making." We wholeheartedly agree. Since a prime objective is to model behavior, it is important to recognize the potential role of other fields such as psychology and behavioral sciences in this discipline. In developing discrete choice models, it is important not to lose sight of the underlying behavioral processes that are driving the behavior. Focusing too much on the statistical formulation confines the development process, whereas advanced discrete choice approaches are immensely capable of capturing complex behavioral dynamics.

3.9 Summary

This chapter has presented some of the basic elements of a decision problem. In it we characterized a choice as the outcome of five steps:

1. definition of the choice problem,
2. generation of alternatives,
3. evaluation of attributes of alternatives,
4. choice, and
5. implementation.

The decision maker for any choice problem can be treated as a single individual, or more abstractly as a family group, a firm, or a public agency. Different decision rules, including dominance, satisfaction, lexicographic rules, and the optimization of a scalar objective function were considered. Although all of these are potentially useful, the remainder of this book focuses on the last of these decision rules, which is generally referred to as utility maximization.

A brief overview of microeconomic consumer theory was provided. This view of demand is appropriate to situations where the feasible choices are continuous variables such as the quantities of various homogeneous consumption commodities. However, the types of problems considered in subsequent chapters are better described as a selection of one of a finite set of discrete bundles of attributes. For such problems discrete choice theory is a more appropriate basis for demand analysis. We presented the extension of the economic framework to include discrete as well as continuous goods, and introduced probabilistic choice theory (that specifies the probability with which an individual selects any feasible alternative) as a potentially powerful framework for analyzing discrete choice situations.

In this book, emphasis is often placed on utility maximization and rationality. However, it is important to note that utility theory is not necessary for the statistical models described in this book. What utility theory does is provide a helpful means of interpreting the framework, deriving choice models, and deriving measures of welfare change. For these reasons it has been used extensively in the development of predictive models of human behavior. However, the resulting statistical models can, and have been, modified to reflect other decision making protocols.

Indeed, two distinct interpretations of probabilistic choice theory were reviewed, one tied closely to microeconomic theory and one not. The *constant utility approach*, rooted primarily in mathematical psychology, hypothesizes that the utilities of alternatives are constant and that the choice probabilities for an individual are functions parameterized by those utilities. The alternate perspective, the *random utility approach*, is more consistent with consumer theory. In this view utilities are treated by the analyst as random due to observational deficiencies resulting from (1) unobserved attributes, (2) unobserved taste variations, (3) measurement errors, and (4) use of instrumental (or proxy) variables. Decision makers are assumed always to choose the utility-maximizing alternatives; the choice probabilities are interpreted as the analyst's statement of the probability that for any decision maker, the utility of an alternative exceeds the utilities of all other feasible alternatives.

The discussion of microeconomics, its extensions, and random utility theory led to the operational framework for discrete choice analysis that is de-

scribed in this book. That is, a framework that works for discrete choices among related commodities, and makes use of a stochastic utility broken into systematic and random components and is a function of attributes of the alternatives and characteristics of the decision maker. The general form of the utility function U that is used throughout this book is:

$$U_{in} = U(z_{in}, S_n, \varepsilon_{in}; \theta) \text{ for all } i \in \mathcal{C}_n, \tag{3.88}$$

where i is the alternative, n the individual, $\mathcal{C}_n$ the choice set, z_{in} the attributes of the alternatives, S_n the characteristics of the decision-maker and context, ε_{in} the random error, and θ the unknown parameters (estimated from data) that represent the tastes. From here on out, we refer to this as simply *utility*, rather than conditional indirect utility.

The subsequent chapters build on the basic concept of random utilities. In particular, we begin with this relatively abstract concept and make plausible further assumptions to derive a rich class of operational models of individual choice. The presentation in earlier chapters focuses on more straightforward models that are more in line with the idea of a rational decision maker. However, later chapters introduce techniques that enable richer models of behavior to be specified and estimated, bridging the gap between the statistical models of behavior and the behavioral science literature emphasized in the last section of this chapter.

Chapter appendix

We provide two derivations of the random utility model. In Section 3.A, it is derived from distributional assumptions on the error terms, while in Section 3.B, the model is derived from the utility differences.

3.A Derivation of the Random Utility Model

We derive the random utility model defined by (3.57) and (3.58). We assume that ε_n is a multivariate random variable with CDF $F_{\varepsilon_n}(\varepsilon_1, \dots, \varepsilon_{J_n})$ and pdf

$$f_{\varepsilon_n}(\varepsilon_1, \dots, \varepsilon_{J_n}) = \frac{\partial^{J_n} F}{\partial \varepsilon_1 \cdots \partial \varepsilon_{J_n}}(\varepsilon_1, \dots, \varepsilon_{J_n}). \tag{3.89}$$

In order to simplify the notations, we derive the model for the probability that the first alternative is chosen, without any loss of generality. Then, the model (3.57)–(3.58) writes

$$P_n(1|\mathcal{C}_n) = \Pr(V_{2n} + \varepsilon_{2n} \leq V_{1n} + \varepsilon_{1n}, \dots, V_{Jn} + \varepsilon_{Jn} \leq V_{1n} + \varepsilon_{1n}), \tag{3.90}$$

or

$$P_n(1|\mathcal{C}_n) = \Pr(\varepsilon_{2n} - \varepsilon_{1n} \leq V_{1n} - V_{2n}, \dots, \varepsilon_{Jn} - \varepsilon_{1n} \leq V_{1n} - V_{Jn}). \tag{3.91}$$

We apply a change of variables defined as

$$\xi_{1n} = \varepsilon_{1n}, \ \xi_{in} = \varepsilon_{in} - \varepsilon_{1n}, \ i = 2, \dots, J_n, \tag{3.92}$$

that is

$$\begin{pmatrix} \xi_{1n} \\ \xi_{2n} \\ \vdots \\ \xi_{(J_n-1)n} \\ \xi_{J_n n} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 & 1 & \cdots & 0 & 0 \\ & & \vdots & & \\ -1 & 0 & \cdots & 1 & 0 \\ -1 & 0 & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{1n} \\ \varepsilon_{2n} \\ \vdots \\ \varepsilon_{(J_n-1)n} \\ \varepsilon_{J_n n} \end{pmatrix}. \tag{3.93}$$

Consequently,

$$P_n(1|\mathcal{C}_n) = \Pr(\xi_{2n} \leq V_{1n} - V_{2n}, \dots, \xi_{J_n n} \leq V_{1n} - V_{J_n n}). \tag{3.94}$$

Note that the above model involves only $J_n - 1$ inequalities, as ξ_{1n} can take any arbitrary value. This is an expression of the fact that only utility differences matter for the choice. The choice probability is given by the CDF of

the random vector $(\xi_{2n}, \ldots, \xi_{J_n n})$ evaluated at $(V_{1n} - V_{2n}, \ldots, V_{1n} - V_{J_n n})$. As the determinant of the change of variable matrix in (3.93) is 1, the pdf of $(\xi_{1n}, \ldots, \xi_{J_n n})$ is equal to the pdf of $(\varepsilon_{1n}, \ldots, \varepsilon_{J_n n})$. Consequently,

$$\begin{aligned} P_n(1|\mathcal{C}_n) &= F_{\xi_{1n},\xi_{2n},\ldots,\xi_{J_n}}(+\infty, V_{1n} - V_{2n}, \ldots, V_{1n} - V_{J_n n}) \\ &= \int_{\xi_1=-\infty}^{+\infty} \int_{\xi_2=-\infty}^{V_{1n}-V_{2n}} \cdots \int_{\xi_{J_n}=-\infty}^{V_{1n}-V_{J_n n}} f_{\xi_{1n},\xi_{2n},\ldots,\xi_{J_n}}(\xi_1, \xi_2, \ldots, \xi_{J_n}) d\xi, \\ &= \int_{\varepsilon_1=-\infty}^{+\infty} \int_{\varepsilon_2=-\infty}^{V_{1n}-V_{2n}+\varepsilon_1} \cdots \int_{\varepsilon_{J_n}=-\infty}^{V_{1n}-V_{J_n n}+\varepsilon_1} f_{\varepsilon_{1n},\varepsilon_{2n},\ldots,\varepsilon_{J_n}}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{J_n}) d\varepsilon, \end{aligned} \tag{3.95}$$

using the definition of the marginal distribution of $(\xi_{1n}, \xi_{2n}, \ldots, \xi_{J_n})$, and the change of variables (3.93). Integrating $J_n - 1$ times the density function of ε_n, we obtain

$$P_n(1|\mathcal{C}_n) = \int_{\varepsilon_1=-\infty}^{+\infty} \frac{\partial F_{\varepsilon_{1n},\varepsilon_{2n},\ldots,\varepsilon_{J_n}}}{\partial \varepsilon_1}(\varepsilon_1, V_{1n} - V_{2n} + \varepsilon_1, \ldots, V_{1n} - V_{J_n n} + \varepsilon_1) d\varepsilon_1. \tag{3.96}$$

Generalizing this result for any alternative i, we obtain (3.62).

As an example, we consider the following CDF for a model with three alternatives:

$$F_{\varepsilon_1,\varepsilon_2,\varepsilon_3}(\varepsilon_1, \varepsilon_2, \varepsilon_3) = \exp\left(-(e^{-\mu\varepsilon_1} + e^{-\mu\varepsilon_2} + e^{-\mu\varepsilon_3})\right). \tag{3.97}$$

Note that it is the product of three i.i.d. univariate random variables with an extreme value distribution. We therefore derive[3] a logit model with three alternatives from (3.62):

$$P(1) = \int_{\varepsilon=-\infty}^{+\infty} \frac{\partial F}{\partial \varepsilon_1}(\varepsilon, V_1 - V_2 + \varepsilon, V_1 - V_3 + \varepsilon) d\varepsilon.$$

Using

$$\frac{\partial F}{\partial \varepsilon_1}(\varepsilon_1, \varepsilon_2, \varepsilon_3) = \exp\left(-(e^{-\mu\varepsilon_1} + e^{-\mu\varepsilon_2} + e^{-\mu\varepsilon_3})\right) \mu e^{-\mu\varepsilon_1},$$

[3]We drop index n to simplify the notations.

and substituting $t = -\exp(-\mu\varepsilon_1)$ and $dt = \mu\exp(-\mu\varepsilon)d\varepsilon$, we obtain

$$\begin{aligned} P(1) &= \int_{t=-\infty}^{0} \exp\left(t\left(1 + e^{-\mu(V_1-V_2)} + e^{-\mu(V_1-V_3)}\right)\right) dt \\ &= \left(1 + e^{-\mu(V_1-V_2)} + e^{-\mu(V_1-V_3)}\right)^{-1} \\ &= \frac{e^{\mu V_1}}{e^{\mu V_1} + e^{\mu V_2} + e^{\mu V_3}}, \end{aligned}$$

which is indeed the choice probability of the logit model.

3.B Derivation of RUM from utility differences

We consider the vector U_n of utility functions defined by (3.58), and we derive the random utility model from the utility differences. Indeed, as discussed in Chapter 3, only the differences of utilities matter for the choice probability. To introduce the idea, we consider an example with 3 alternatives, that is $J_n = 3$, and we compute the probability that alternative 2, say, is chosen by individual n. We have

$$\begin{aligned} U_{1n} &= V_{1n} + \varepsilon_{1n} \\ U_{2n} &= V_{2n} + \varepsilon_{2n} \\ U_{3n} &= V_{3n} + \varepsilon_{3n}, \end{aligned} \tag{3.98}$$

and

$$P_n(2|\mathcal{C}_n = \{1,2,3\}) = \Pr(U_{2n} \geq U_{in},\ i = 1,3) = \Pr(U_{in} - U_{2n} \leq 0,\ i = 1,3), \tag{3.99}$$

where

$$\begin{aligned} U_{1n} - U_{2n} &= V_{1n} - V_{2n} + \varepsilon_{1n} - \varepsilon_{2n}, \\ U_{3n} - U_{2n} &= V_{3n} - V_{2n} + \varepsilon_{3n} - \varepsilon_{2n}. \end{aligned} \tag{3.100}$$

In order to use vector notations, we introduce the $(J_n - 1) \times J_n$ matrix Δ_i that transforms the utilities to differences with respect to alternative i, that is

$$\begin{pmatrix} U_{1n} - U_{in} \\ U_{2n} - U_{in} \\ \vdots \\ U_{J_n n} - U_{in} \end{pmatrix} = \Delta_i U_n. \tag{3.101}$$

In the general case, the matrix Δ_{in} is a $J_n - 1 \times J_n$ matrix, obtained by inserting one additional column into the identity matrix of size $J_n - 1 \times J_n - 1$. The

added column is in position i, and all its entries are -1. The structure of Δ_{in} is summarized as follows:

	1	2	$\cdots$	$i-1$	i	$i+1$	$\cdots$	J_n-1	J_n
1	1	0		0	-1	0		0	0
2	0	1		0	-1	0		0	0
	$\vdots$		$\ddots$		$\vdots$				$\vdots$
$i-1$	0	0	$\cdots$	1	-1	0	$\cdots$	0	0
$i+1$	0	0		0	-1	1		0	0
	$\vdots$				$\vdots$		$\ddots$		$\vdots$
J_n-2	0	0		0	-1	0		1	0
J_n-1	0	0	$\cdots$	0	-1	0	$\cdots$	0	1

(3.102)

For our example, we have

$$\Delta_{2n} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \tag{3.103}$$

such that the left-hand side of (3.100) writes

$$\Delta_{2n}U = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} U_{1n} \\ U_{2n} \\ U_{3n} \end{pmatrix} = \begin{pmatrix} U_{1n} - U_{2n} \\ U_{3n} - U_{2n} \end{pmatrix}. \tag{3.104}$$

The choice model (3.99) can be generalized to J_n alternatives as

$$\begin{aligned} P_n(i|\mathcal{C}_n) &= \Pr(\Delta_{in}U_n \leq 0) \\ &= \Pr(\Delta_{in}V_n + \Delta_{in}\varepsilon_n \leq 0) \\ &= \Pr(\Delta_{in}\varepsilon_n \leq -\Delta_{in}V_n) \end{aligned} \tag{3.105}$$

where the comparison of two vectors must be understood componentwise, that is

$$\Delta_{in}U_n \leq 0 \text{ if and only if } (\Delta_{in}U_n)_k \leq 0, \ \forall k. \tag{3.106}$$

If a distributional assumption is made on $\Delta_{in}\varepsilon_n$, the choice model is obtained directly from the corresponding CDF evaluated at $-\Delta_{in}V_n$:

$$P_n(i|\mathcal{C}_n) = \Pr(\Delta_{in}\varepsilon_n \leq -\Delta_{in}V_n) = F_{\Delta_{in}\varepsilon_n}(-\Delta_{in}V_n), \tag{3.107}$$

where $F_{\Delta_{in}\varepsilon_n}(\cdot) : \mathbb{R}^{J_n-1} \to \mathbb{R}$ is the CDF of the assumed distribution. It is defined as

$$F_{\Delta_{in}\varepsilon_n}(v) = \int_{-\infty}^{v_{J_n-1}} \cdots \int_{-\infty}^{v_1} f_{\Delta_{in}\varepsilon_n}(\xi) d\xi_1 \ldots d\xi_{J_n-1}, \tag{3.108}$$

where $f_{\Delta_{in}\varepsilon_n}(\cdot) : \mathbb{R}^{J_n-1} \to \mathbb{R}$ is the pdf of the assumed distribution. Therefore, the model is also given by

$$P_n(i|\mathcal{C}_n) = \int_{-\infty}^{-(\Delta_{in}V_n)_{J_n-1}} \cdots \int_{-\infty}^{-(\Delta_{in}V_n)_1} f_{\Delta_{in}\varepsilon_n}(\xi)\, d\xi_1 \dots d\xi_{J_n-1}. \tag{3.109}$$

For example, assume that $\Delta_{in}\varepsilon_n$ is normally distributed. It is a direct consequence of the direct assumption that the random vector ε_n follows a multivariate normal distribution

$$\varepsilon_n \sim N(0, \Sigma), \tag{3.110}$$

where Σ is a variance-covariance matrix. Indeed, a linear transform of a multivariate normal random variable is also normally distributed, and the random vector $\Delta_{in}U_n$ also follows a multivariate normal distribution, that is

$$\Delta_{in}U_n \sim N(\Delta_{in}V_n, \Delta_{in}\Sigma\Delta_{in}^T), \text{ and } \Delta_{in}\varepsilon_n \sim N(0, \Delta_{in}\Sigma\Delta_{in}^T). \tag{3.111}$$

The pdf is

$$f_{\Delta_{in}\varepsilon_n}(\xi) = (2\pi)^{-\frac{J_n}{2}} |\Delta_{in}\Sigma\Delta_{in}^T|^{-\frac{1}{2}} e^{-\frac{1}{2}\xi^T(\Delta_{in}\Sigma\Delta_{in}^T)^{-1}\xi}, \tag{3.112}$$

where the notation $|\Delta_{in}\Sigma\Delta_{in}^T|$ denotes the determinant of the matrix $\Delta_{in}\Sigma\Delta_{in}^T$.

The model (3.109) where the pdf is defined by (3.112) is called the *probit* model. The multifold integral is difficult to handle. Various algorithms have been proposed in the literature (Bolduc, 1999, Natarajan et al., 2000). They are however limited to choice models with a few alternatives (less than 10).

Part II

Basic methods

Chapter 4

Binary choice

Contents

In the preceding chapter we developed the concept of the individual decision maker who, faced with a set of feasible discrete alternatives, selects the one that yields greatest utility. We noted that for a variety of reasons the utility of any alternative is, from the perspective of the analyst, best viewed as a random variable. This leads directly to the notion of the random utility model in which the probability of any alternative i being selected by person n from choice set C_n is given by

$$P(i|C_n) = \Pr(U_{in} \geq U_{jn}, \forall j \in C_n). \tag{4.1}$$

We ignore situations where $U_{in} = U_{jn}$ for any i and j in the choice set because if U_{in} and U_{jn} are continuous random variables then the probability $\Pr(U_{in} = U_{jn})$ that they are equal is zero.

In this chapter we pursue this basic idea further by considering the special case where the choice set C_n contains exactly two alternatives. Such situations lead to what are termed binary choice models. For convenience we denote the choice set C_n as $\{i, j\}$, where, for example, alternative i might be the option of driving to work and alternative j would be taking the train to work. The probability of person n choosing i is

$$P_n(i) = \Pr(U_{in} \geq U_{jn}), \tag{4.2}$$

and the probability of person n choosing alternative j is

$$P_n(j) = 1 - P_n(i).$$

Our goal in this chapter is to develop the basic theory of random utility models into a class of operational binary choice models.

A detailed discussion of binary models serves a number of purposes. First, the simplicity of binary choice situations makes it possible to develop a range of practical models, which is more tedious in more complicated choice situations. Second, there are many basic conceptual problems that are easiest to illustrate in the context of binary choice. Many of the solutions we present in this chapter can be directly applied to situations with more than two alternatives.

In section 4.1 we first consider how the general theory developed in chapter 3 can be made operational. Then in section 4.2 we derive the most widely

used binary choice model forms: probit and logit. Section 4.3 presents two examples of binary choice models estimated using real datasets. In Section 4.4 we consider the problem of estimating the parameters of a binary choice model by maximum likelihood. This is followed by a simple example and then a more extended example of maximum likelihood estimation in section 4.5. Finally, Section 4.6 summarizes the key points in the chapter.

Additional and more advanced materials on binary choice models are included in the appendices of this chapter.

4.1 Making Random Utility Theory Operational

An operational model allows one to calculate the choice probability $\mathsf{P}_n(\mathtt{i})$ as a function of explanatory variables that can be observed. Once the model is available, these variables can be manipulated to reflect "what if" scenarios, so that the model can predict the associated impact on individuals' choices and market shares. The ways to apply the models for policy, marketing and planning analysis is described in Chapter 10. In this section, we introduce the three basic steps to obtain this operational model:

1. The separation of total utility into systematic and random components.

2. The specification of the systematic component, that is the selection of the explanatory variables and the functional form describing how they influence the systematic part of the utility.

3. The specification of the random component, that is assumptions about its distribution.

Each of these is considered in turn in this section for the case of binary choice. Note that the utility associated by an indivual with an alternative is a latent quantity, meaning that it cannot be directly observed. This specific feature of the model implies various normalization issues, namely to fix the origin and the units of the utility, that are also discussed. Normalization is particularly important when the parameters have to be estimated from data, as not all parameters can be identified.

4.1.1 Systematic component and disturbances

Recalling that U_{in} and U_{jn} are random variables, we begin by dividing each of the utilities into two additive parts as follows:

$$\begin{aligned} U_{in} &= V_{in} + \varepsilon_{in}, \\ U_{jn} &= V_{jn} + \varepsilon_{jn}. \end{aligned} \tag{4.3}$$

V_{in} and V_{jn} are called the systematic (or representative) components of the utility of i and j; ε_{in} and ε_{jn} are the random parts and are called the disturbances (or random components).

It is important to stress that V_{in} and V_{jn} are functions and are assumed at this point in the text to be deterministic (i.e., nonrandom). The terms ε_{in} and ε_{jn} may also be functions, but they are random from the observational perspective of the analyst.

Substituting (4.3) into (4.2), we obtain

$$\begin{aligned} P_n(i) = \Pr(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}) &= \Pr(\varepsilon_{in} - \varepsilon_{jn} \geq V_{jn} - V_{in}) \\ &= \Pr(\varepsilon_{jn} - \varepsilon_{in} \leq V_{in} - V_{jn}). \end{aligned} \tag{4.4}$$

We can see that the absolute levels of V's and ε's do not matter; all that matters is the relative values of the differences. This is important to keep in mind, as it has implications on the specification of the model.

4.1.2 Specification of the Systematic Component

We have already noted that the specification of the absolute levels of utilities is irrelevant; only their difference matters. We could therefore develop choice models by specifying only the differences, ignoring the individual components. However, in later chapters in which we analyze choice situations with more than two alternatives, we have the utility of an alternative appearing in multiple utility differences, and we find it more convenient to write each utility function separately, keeping in mind that only their difference matters in terms of the choice probabilities.

The first issue in specifying V_{in} and V_{jn} is to ask, What types of variables can enter these functions? As discussed in chapter 3, for any individual n any alternative i can be represented by a vector of attributes z_{in}. In a choice of travel mode, z_{in} might include travel time and cost. It is also useful to characterize the decision maker n by another vector of characteristics, which we shall denote by S_n. These are often variables such as income, household size, age, occupation, and gender. The vector S_n may also contain variables describing the choice context for individual n, such as weather

conditions, or the day of the week. All variables in S_n have the same values for each alternative, and vary across individuals. The problem of specifying the functions V_{in} and V_{jn} consists of defining combinations of z_{in}, z_{jn}, and S_n that reflect reasonable hypotheses about the effects of such variables.

It is generally convenient to define a new vector of variables, which includes both z_{in} and S_n. Notationally, we write the vectors $x_{in} = h(z_{in}, S_n)$ and $x_{jn} = h(z_{jn}, S_n)$, where h is a function. The function h can be as simple as a pure attribute model, with $x_{in} = z_{in}$, but can also involve interactions of z_{in} with elements of S_n such as price divided by income or the log of income minus price. More examples of such transforms are discussed later in this chapter.

Now we can write the systematic components of the utilities of i and j in (4.3) as

$$V_{in} = V(x_{in}) \quad \text{and} \quad V_{jn} = V(x_{jn}). \tag{4.5}$$

The second question is then, What is a reasonable functional form for V? Here we are generally concerned with two sometimes contradictory criteria for selecting a functional form. First, we would like the function to reflect any theory we have about how the various elements in x influence utility; second, we would like to use functions that have convenient computational properties that make it easy to estimate their unknown parameters. In most cases of interest, researchers faced with these conflicting criteria have chosen to use functions that are *linear in the unknown parameters*.

If we denote $\beta^T = (\beta_1, \beta_2, \ldots, \beta_K)$ as the (row) vector of K unknown parameters, we have

$$\begin{array}{rcl} V_{in}(x_{in}, \beta) = \beta^T x_{in} = \beta_1 x_{in1} + \beta_2 x_{in2} + \cdots + \beta_K x_{inK}, \\ V_{jn}(x_{jn}, \beta) = \beta^T x_{jn} = \beta_1 x_{jn1} + \beta_2 x_{jn2} + \cdots + \beta_K x_{jnK}. \end{array} \tag{4.6}$$

When such a linear formulation is adopted, parameters $\beta_1, \ldots, \beta_K$ are called *coefficients*. The reader should note that we have adopted the convention that both utilities have the same vector of coefficients $\beta^T = (\beta_1, \ldots, \beta_K)$. It is important to stress that this is only a notational convention because, by appropriately defining the various elements in x, we can effectively give each systematic utility function different coefficients. A coefficient appearing in all utility functions is *generic*, and a coefficient appearing in only one utility function is *alternative specific*. Consider a binary mode choice example, where one alternative is auto (A) and the other is transit (T), and where the utility functions are defined as

$$\begin{array}{rcl} V_{An} = 0.37 - 2.13 t_{An} \\ V_{Tn} = \quad\;\; - 2.13 t_{Tn}. \end{array} \tag{4.7}$$

In this case it appears as though the auto utility has an additional term equal to 0.37. We can "convert" this model into the form of equation (4.6) by defining our x's as follows:

$$\begin{aligned} x_{An1} &= 1, \\ x_{Tn1} &= 0, \\ x_{An2} &= t_{An}, \\ x_{Tn2} &= t_{Tn}, \end{aligned} \tag{4.8}$$

with $K = 2$, $\beta_1 = 0.37$ is alternative specific, and $\beta_2 = -2.13$ is generic. Thus,

$$\begin{aligned} V_{An} &= \beta^T x_{An} = \beta_1 x_{An1} + \beta_2 x_{An2} = 0.37 - 2.13 t_{An}, \\ V_{Tn} &= \beta^T x_{Tn} = \beta_1 x_{Tn1} + \beta_2 x_{Tn2} = - 2.13 t_{Tn}. \end{aligned} \tag{4.9}$$

In this example, the variable x_{An1} is an alternative specific (i.e., auto) dummy variable and β_1 is called an *alternative specific constant*.

A model with a linear-in-parameter formulation can be described in a *specification table*. A specification table has as many columns as alternatives in the model (two in the specific context of binary choice), plus one that contains the coefficients, and as many rows as coefficients (K). Entry (k, i) of the table contains x_{ik}, the variable k for alternative i. The specification table of the above example is reported in Table 4.1.

		Auto	Train
β_1	0.37	1	0
β_2	-2.13	t_{An}	t_{Tn}

Table 4.1: Simple binary example

Linearity in the parameters is not as restrictive an assumption as one might first think. *Linearity in the parameters is not equivalent to linearity in the variables* z *and* S. We allow for any function h of the variables so that polynomial, piecewise linear, logarithmic, exponential, and other transformations of the attributes are valid for inclusion as elements of x. Such functions are explored as part of examples described later in this chapter and in subsequent chapters.

Although linear-in-parameters function are widely used, nonlinear utility functions can be considered as well. Examples of nonlinear specifications are discussed later in this chapter.

We note that we have implicitly assumed that the parameters β_1, β_2,..., β_K are the same for all members of the population. Again this is not as

restrictive as it may seem at first glance. If different socioeconomic groups are believed to have entirely different parameters β, then it is possible to develop a distinct model for each subgroup. This is termed market segmentation. In the extreme case a market segment corresponds to a single individual, and a vector of parameters is specific to an individual. Clearly, the estimation of the parameters of such a model would require sufficient data for each individual. If the preferences or tastes of different members of the population vary systematically with some known socioeconomic characteristics, we can define some of the elements in x to reflect this. For example, it is not unusual to define as a variable cost divided by income, reflecting the a priori belief that the importance of cost declines as the inverse of income. As an advanced topic in chapter 13, we also consider the case where $\beta_1, \beta_2, \ldots, \beta_K$ are treated as random variables distributed across the population. We refer the reader to Section 5.4 for additional discussions about the specification of the systematic component.

4.1.3 Specification of the Disturbances

Our last remaining component of an operational binary choice model is the disturbance terms. As with the systematic components V_{in} and V_{jn}, we can discuss the specification of binary choice models by considering only the difference $\varepsilon_{jn} - \varepsilon_{in}$ rather than each element ε_{in} and ε_{jn} separately. Whereas in section 4.1 we chose to keep the V's separate, here we shall be somewhat more eclectic. We choose the interpretation that is most convenient and insightful for the purposes of exposition.

Remember from Section 3.7 that the disturbances represent various variables (attributes and characteristics) of the choice situation that are unknown to the analyst, as well as specification and measurement errors. Assumptions have to be made about how all these elements are combined together. We discuss the *mean* of the disturbances, that is their average value across the population of individuals, their *variance*, that is the spread of the values around the mean, and their *distribution*, characterizing the probability that different values may occur.

We begin first by considering the problem of the mean of the disturbances. As discussed earlier, only differences between utilities matter. Therefore, the choice probabilities (4.4) are unaffected by the addition of a constant to both utilities; we can add or subtract any number from all utilities, and their relative value (as measured by their difference) is unaffected. Another aspect of the mean of the disturbances is that there is no distinction between shifting the mean of the disturbance of one alternative's utility and shifting the systematic component by the same amount. Thus, if the mean of alternative i's

disturbance is some amount greater than that of alternative j's disturbance, we can fully represent the difference by adding that amount to V_{in}. *This implies that as long as one can add a constant to the systematic component, the means of disturbances can be defined as equal to any constant without loss of generality.* Usually the most convenient assumption is that all the disturbances have zero means. We maintain the convention of assuming that any nonzero means of the disturbances are "absorbed" into the systematic component of the utility function, unless noted otherwise. To do so, we introduce parameters a_{in} and a_{jn} such that (4.3) becomes

$$\begin{array}{rcl} U_{in} & = & V_{in} + a_{in} + \varepsilon_{in}, \\ U_{jn} & = & V_{jn} + a_{jn} + \varepsilon_{jn}. \end{array} \tag{4.10}$$

Therefore, assuming that the disturbances have zero means, we have

$$E[U_{in}] = V_{in} + a_{in}. \tag{4.11}$$

In general, it is assumed that the error components ε_{in} are identically distributed across n, so that $a_{in} = a_i$ and $a_{jn} = a_j$, for all decision makers n, and a_i and a_j are unknown parameters to be estimated. They are called *alternative specific constants*, and play the same role as intercepts in linear regression. As only the difference $\varepsilon_{jn} - \varepsilon_{in}$ matters in this context, only the difference between the two constants can be estimated. In practice, one of the two constants is normalized to 0 and the other one is estimated:

$$\begin{array}{rcll} U_{in} & = & V_{in} + a_i & +\varepsilon_{in}, \\ U_{jn} & = & V_{jn} & +\varepsilon_{jn}, \end{array} \tag{4.12}$$

or, equivalently

$$\begin{array}{rcll} U_{in} & = & V_{in} & +\varepsilon_{in}, \\ U_{jn} & = & V_{jn} + a_j & +\varepsilon_{jn}, \end{array} \tag{4.13}$$

where $a_i = -a_j$. This normalization sets the origin of the utilities.

In addition to the mean of the disturbances we must ensure that their variance is consistent with that of the V's. Note that any positive scaling of the utilities U_{in} and U_{jn} does not affect the choice probabilities. To see this, suppose the total utility of alternatives i and j are as given in equation (4.3). Note that

$$\begin{array}{rcl} P_n(i) & = & \Pr(U_{in} \geq U_{jn}) \\ & = & \Pr(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}) \\ & = & \Pr(\alpha V_{in} + \alpha\varepsilon_{in} \geq \alpha V_{jn} + \alpha\varepsilon_{jn}), \text{for any } \alpha > 0. \end{array} \tag{4.14}$$

Thus, just as adding a constant to each utility did not affect the choice probabilities, multiplying both by a positive constant does not. Hence, to

make our V's unique, we must "fix their scale." This can also be seen as fixing the *units* of the utility. This is typically done by placing some restrictions on the disturbances. For example, if we assume that $\alpha(\varepsilon_{jn} - \varepsilon_{in})$ is distributed with variance equal to 1, we in effect define the scale as follows:

$$\mathrm{Var}[\alpha(\varepsilon_{jn} - \varepsilon_{in})] = \alpha^2 \,\mathrm{Var}[\varepsilon_{jn} - \varepsilon_{in}] = 1,$$

which implies that

$$\alpha = \frac{1}{\sqrt{\mathrm{Var}[\varepsilon_{jn} - \varepsilon_{in}]}}. \tag{4.15}$$

Obviously any other assumed value for the variance would work as well. The choice is entirely arbitrary, and we usually use a scale that is analytically or computationally convenient. *It is important to emphasize that normalizing the scale does not remove it.* It is a parameter of the model, but is not identified together with the V's. Namely, it plays an important role when we compare models with different normalization conditions. It also comes into play for more advanced models, discussed later in this book. Also, we analyze limit values of the scale (0 and $+\infty$) for specific models later in this chapter.

Given an appropriate scaling of the disturbances, we can now ask, What is an appropriate functional form for the distribution of ε_{in} and ε_{jn}, or for $\varepsilon_{jn} - \varepsilon_{in}$? Varying the assumptions about the distributions of ε_{in} and ε_{jn} (or equivalently, assumptions about their difference) leads to different choice models. However, it makes little sense to think of the specification of the distribution of the ε's independently from the specification of the V's. In particular, since the ε's reflect the various sources of errors in the specification and measurements of the V's discussed in chapter 3, different specifications of V leads to different appropriate distributions for ε.

In actual applications the disturbances are a composite of a great number of unobserved effects, each of which contributes in some way to the disturbances' distribution. Although it is often difficult to make strong statements about the overall distribution of the disturbances, we are occasionally able to obtain insights into how to improve models by thinking more carefully about what "goes into" the ε's. This becomes increasingly relevant when we discuss choice models for more than two alternatives in chapter 5.

With the foregoing as background we now explore specific binary choice models. Our presentation is divided into two parts. First, in section 4.2 we derive the most common binary choice models and give some examples in section 4.3. Then in sections 4.4 and 4.5 (see also Appendix 4.B), we discuss the methods that can be used to estimate the models' parameters, giving further examples of actual applications of the procedures we derive.

4.2 Common Binary Choice Models

In this section, we finalize our derivation of operational models by introducing the most common binary choice models: the *binary probit* and the *binary logit* models. In each subsection we begin by making some assumptions about the distribution of the two disturbances, ε_{in} and ε_{jn}, or about the difference between them. Given one of these assumptions, we then solve for the probability that alternative i is chosen. (The probability that j is chosen is trivially equal to $1 - P_n(i)$.) As a final step we explore some of the properties of each model.

Recall the expression of the random utility model in (4.4), as follows:

$$\begin{aligned} P_n(i) &= \Pr(\varepsilon_{jn} - \varepsilon_{in} \leq V_{in} - V_{jn}) \\ &= \Pr(\varepsilon_n \leq V_{in} - V_{jn}), \end{aligned} \tag{4.16}$$

where $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$. It means that the probability for individual n to choose alternative i is equal to the probability that the difference $V_{in} - V_{jn}$ exceeds the value of ε_n. To calculate that quantity, we need to know how ε_n is distributed.

The distribution of a random variable is characterized by a function f_{ε_n}, called the *probability density function* (pdf), such as the bell-shaped function depicted in Figure 4.1, where the x-axis corresponds to the values ε that ε_n can take and the y-axis to $f_{\varepsilon_n}(\varepsilon)$. The value of this function is related to how values of ε_n are likely to occur. In this example, the higher values of the function close to zero mean that values of ε_n close to zero are more likely to be seen than those away from zero.

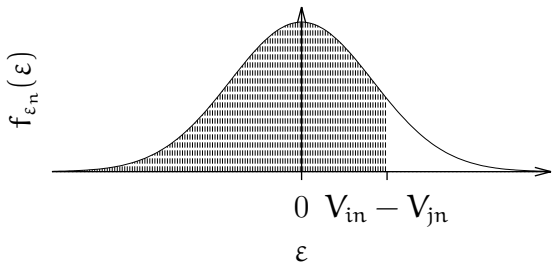


Figure 4.1: Probability density function of $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$

To calculate the choice probability (4.16), we must consider all values of ε_n that are inferior or equal to $V_{in} - V_{jn}$, and cumulate the values of the density function for such values of ε_n only. This quantity corresponds to the shaded area in Figure 4.1.

The function providing the probability that the value of a random variable ε_n is below a given threshold is called a *Cumulative Distribution Function* (CDF), and is denoted by F_{ε_n}:

$$\Pr(\varepsilon_n \leq c) = F_{\varepsilon_n}(c) = \int_{\varepsilon=-\infty}^{c} f_{\varepsilon_n}(\varepsilon)d\varepsilon. \tag{4.17}$$

The CDF is an increasing function, with values between 0 and 1. The CDF of the pdf from Figure 4.1 is represented in Figure 4.2.

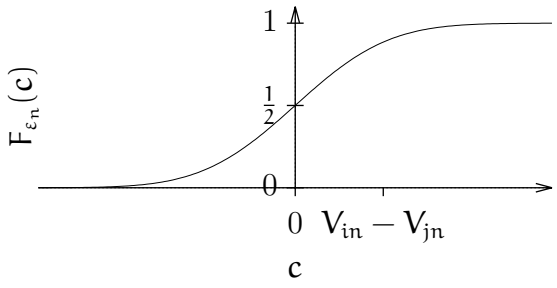


Figure 4.2: Cumulative distribution function

Therefore, the probability expression on the right hand side of (4.16) is equal to the cumulative distribution function (CDF) of ε_n evaluated at $V_{in} - V_{jn}$ as follows:

$$P_n(i) = F_{\varepsilon_n}(V_{in} - V_{jn}). \tag{4.18}$$

Then, the choice model is obtained by deriving the CDF of ε_n. In the following sections, this derivation is illustrated for commons binary choice models. For the definition of CDF, see Definition B.2 in Appendix B.

4.2.1 Binary Probit

The disturbances represent various variables (attributes and characteristics) of the choice situation that are unknown to the analyst, as well as specification and measurement errors. One possible assumption is that all these elements are independent from each other and add up to form the disturbance. In other words, we can view the disturbances as the sum of a large number of unobserved components. In this case, their distribution would follow the bell shaped function depicted in Figure 4.1. From a theoretical point of view, we can invoke the central limit theorem (theorem C.2) that states that it is the distribution of a *normal* random variable.

To be more specific, suppose that ε_{in} and ε_{jn} are both normal with zero means (according to the above discussion about the mean) and variances σ_i^2 and σ_j^2, respectively. Suppose further that they have covariance σ_{ij}. Note the absence of index n on the σ parameters, emphasizing that these quantities are assumed to be the same across all individuals. Under these assumptions the term $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$ is also normally distributed with mean zero and variance $\sigma_i^2 + \sigma_j^2 - 2\sigma_{ij} = \sigma^2$. Note that we implicitly assume here that the random variables $\varepsilon_{jn} - \varepsilon_{in}$ are independent and identically distributed (i.i.d.) across individuals, and independent of the attributes x_n. We can use this result to solve for the choice probabilities as follows:

$$\begin{aligned} P_n(i) &= \Pr(\varepsilon_{jn} - \varepsilon_{in} \le V_{in} - V_{jn}) \\ &= \int_{\varepsilon=-\infty}^{V_{in}-V_{jn}} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\varepsilon}{\sigma}\right)^2\right] d\varepsilon, \ \sigma > 0, \\ &= \frac{1}{\sqrt{2\pi}} \int_{u=-\infty}^{(V_{in}-V_{jn})/\sigma} \exp\left[-\frac{1}{2}u^2\right] du \\ &= \Phi\left(\tfrac{V_{in}-V_{jn}}{\sigma}\right), \end{aligned} \tag{4.19}$$

where the second equation is the CDF of a normal random variable with mean zero and variance σ^2, the third equation is obtained from the change of variables $u = \varepsilon/\sigma$ to convert it to standard normal, and $\Phi(\cdot)$ denotes the standardized cumulative normal distribution. This model is called *binary probability unit* or, more commonly, *binary probit*. In the case where $V_{in} = \beta^T x_{in}$ and $V_{jn} = \beta^T x_{jn}$,

$$P_n(i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta^T(x_{in}-x_{jn})/\sigma} \exp\left[-\frac{1}{2}u^2\right] du = \Phi\left(\frac{\beta^T(x_{in} - x_{jn})}{\sigma}\right). \tag{4.20}$$

The binary probit function is sketched in Figure 4.3. Note that the choice function has a characteristic sigmoidal shape and that the choice probabilities are never zero or one. They approach zero and one as the systematic components of the utilities become more and more different.

The binary probit choice probabilities depend only on σ, not on σ_i, σ_j, and σ_{ij}. Thus the variances and covariance of the individual disturbances are irrelevant to the choice probabilities. Moreover even the choice of σ is arbitrary since, by rescaling σ and β by any positive constant α, we do not affect the choice probabilities at all:

$$P_n(i) = \Phi\left(\frac{\beta^T(x_{in} - x_{jn})}{\sigma}\right) = \Phi\left(\frac{\alpha\beta^T(x_{in} - x_{jn})}{\alpha\sigma}\right), \forall \alpha > 0. \tag{4.21}$$

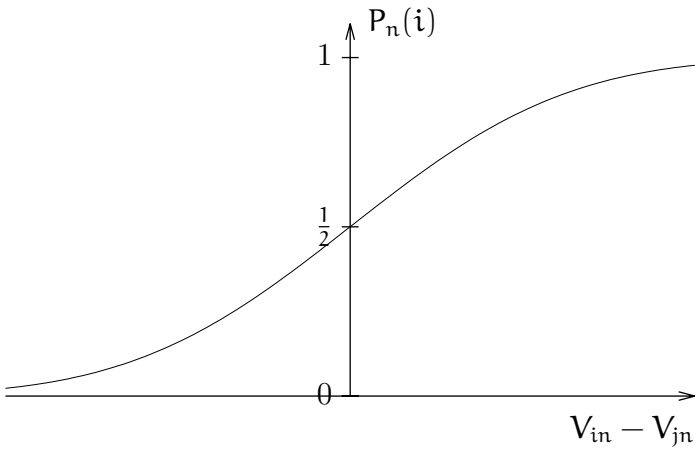


Figure 4.3: The binary probit model

Actually, $1/\sigma$ in (4.20) is the scale of the utility function that can be set to an arbitrary positive value, usually $\sigma = 1$. From (4.15), this is equivalent to normalize the variance of ε_n to 1. Any other value would serve as well. Another value for σ would simply affect the estimated values of β, to obtain the same ratio β/σ.

Remember that normalizing the scale does not remove it. If the β's are set, varying the parameter σ modifies the choice probability. There are actually two limiting cases of a probit model of special interest, both involving extreme values of the scale parameter. The first case is for $\sigma \to 0$:

$$\lim_{\sigma \to 0} P_n(i) = \begin{cases} 1 & \text{if } V_{in} - V_{jn} > 0, \\ 0 & \text{if } V_{in} - V_{jn} < 0; \end{cases}$$

that is, as $\sigma \to 0$, the variance vanishes and the choice model is deterministic. On the other hand, when $\sigma \to \infty$, the choice probability of i becomes $1/2$. Intuitively the model predicts equal probability of choice for each alternative, irrespectively of V_{in} and V_{jn} (see Figure 4.4).

Binary probit has been widely used in diverse fields. It has its origin in psychophysics, the branch of psychology that deals with the relationships between physical stimuli and mental phenomena. Thurstone (1927), introducing his law of comparative judgment, proposed a model of imperfect discrimination in which an alternative with true "stimulus level" V_{in} is perceived with a normal error as $V_{in} + \varepsilon_{in}$. He showed that the choice probability

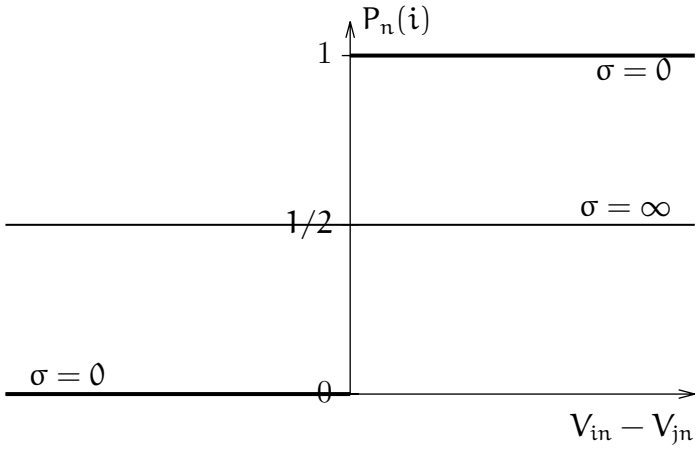


Figure 4.4: The limiting cases of the binary probit model

has a form that we now call binomial probit. Early applications are found in toxicology, where $V_{in} - V_{jn}$ represents the dosage of some toxic substance to which a population is exposed and $P_n(i)$ is interpreted as the probability that a member of the population falls ill or dies (see Finney, 1971). One of the early application of this model to travel mode choice is due to Lisco (1967) who analyzed the behavior of commuters in Chicago.

4.2.2 Binary Logit

Although binary probit is intuitively reasonable and there are at least some theoretical grounds for its assumptions about the distribution of ε_{in} and ε_{jn}, it has the unfortunate property of not having a closed form. Instead, we must express the choice probability as an integral that cannot be solved to an explicit analytical equation. Although it is not really an issue in the binary case where the integral can easily be evaluated with numerical methods, it becomes problematic when we consider more alternatives. This aspect of binary probit provides the motivation for searching for a choice model that is more convenient analytically. One such model is *binary logit*.

Its derivation from the random utility model follows from viewing the disturbances as the *maximum* (and not a sum, like for probit) of a large number of unobserved independent components. Similarly to the Central Limit Theorem (theorem C.1) which justifies the normal distribution as the

limit distribution of the sum of many random variables, a theorem due to Gumbel (1958) justifies the use of the *extreme value*, or Gumbel, distribution (see Appendix B). The extreme value distribution[1], while not as known as the normal distribution, is a proper distribution and, as such, has a pdf and a CDF, with location parameter η and scale parameter $\mu > 0$. We denote an extreme value distributed random variable ξ by

$$\xi \sim \mathrm{EV}(\eta, \mu). \tag{4.22}$$

Its probability density function is given by

$$f_\xi(\varepsilon) = \mu e^{-\mu(\varepsilon-\eta)} e^{\left(-e^{-\mu(\varepsilon-\eta)}\right)}, \tag{4.23}$$

and its cumulative distribution function (CDF) is given by

$$\begin{aligned} F_\xi(c) &= \int_{\varepsilon=-\infty}^{c} f_\xi(\varepsilon)\,d\varepsilon \\ &= e^{-e^{-\mu(c-\eta)}}. \end{aligned} \tag{4.24}$$

Note that, in this case, the CDF is defined by a closed form equation.

Therefore, we assume now that ε_{in} and ε_{jn} are independent and identically extreme value distributed (i.i.d.) across alternatives i and individuals n, with location parameter $\eta = 0$ and scale parameter $\mu > 0$ Note that μ has no index i or n as all error terms ε_{in} have the same distribution. As with the binary probit, we also assume that the random variables $\varepsilon_{jn} - \varepsilon_{in}$ are independent of the attributes x_n.

The distribution of $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$ must be derived from the assumptions about ε_{in} and ε_{jn}. From the property 5 of the extreme value distribution, presented in Appendix 4.A, ε_n follows a *logistic* distribution with location parameter 0 and scale parameter μ. This distribution has pdf

$$f_{\varepsilon_n}(\varepsilon) = \frac{\mu e^{-\mu\varepsilon}}{(1 + e^{-\mu\varepsilon})^2} \tag{4.25}$$

and CDF

$$F_{\varepsilon_n}(c) = \frac{1}{1 + e^{-\mu c}}. \tag{4.26}$$

Similarly to the derivation (4.19) of the probit model, but now under the assumption that ε_n is logistically distributed, the choice probability for

[1]Note that there are actually several types of extreme value distributions. The distribution described here is called "Extreme Value Type I" in the statistics literature.

E. J. Gumbel was born in Munich in 1891, in a Jewish family of bankers. He is considered to be the father of extreme value theory (see Gumbel, 1958). As a politically involved left-wing pacifist, he strongly condemned the right wing's campaign of organized assassination that started in 1919 in Germany. He was the first German professor to be expelled from university under the pressure of the Nazis. He left Heidelberg to Paris, where he met Borel and Fréchet. In 1940, he had to escape to New-York, where he continued his fight against Nazism by helping the US secret service. He died in 1966.

Figure 4.5: Emil Julius Gumbel

alternative i is given by

$$\begin{aligned} P_n(i) &= \Pr(\varepsilon_n \leq V_{in} - V_{jn}) \\ &= F_{\varepsilon_n}(V_{in} - V_{jn}) \\ &= \frac{1}{1 + e^{-\mu(V_{in} - V_{jn})}} \\ &= \frac{e^{\mu V_{in}}}{e^{\mu V_{in}} + e^{\mu V_{jn}}}. \end{aligned} \tag{4.27}$$

This is the binary logit model, which is illustrated in Figure 4.6. Here the choice probability of alternative i is depicted for two different values of μ. Note that if V_{in} and V_{jn} are linear in their parameters,

$$\begin{aligned} P_n(i) &= \frac{e^{\mu \beta^T x_{in}}}{e^{\mu \beta^T x_{in}} + e^{\mu \beta^T x_{jn}}} \\ &= \frac{1}{1 + e^{-\mu \beta^T (x_{in} - x_{jn})}}. \end{aligned} \tag{4.28}$$

The parameter μ cannot be distinguished from the overall scale of the β's. For convenience we generally make an arbitrary assumption that $\mu = 1$.

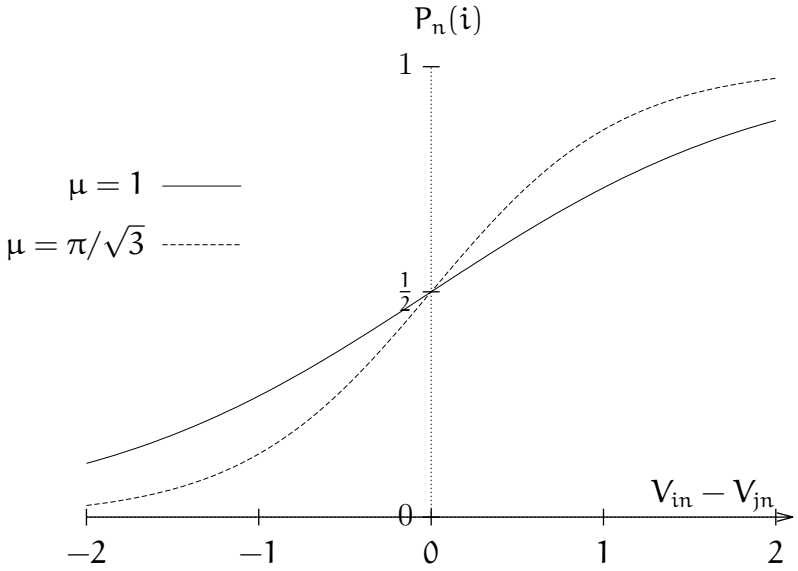


Figure 4.6: The binary logit model

This corresponds to assuming the variances of ε_{in} and ε_{jn} are both $\pi^2/6$, implying that $\mathrm{Var}(\varepsilon_{jn} - \varepsilon_{in}) = \pi^2/3$ (see property 3 in Appendix 4.A). Note that this differs from the standard scaling of binary probit models, where we set $\mathrm{Var}(\varepsilon_{jn} - \varepsilon_{in}) = 1$, and it implies that the scaled logit coefficients are $\pi/\sqrt{3}$ times larger than the scaled probit coefficients. A rescaling of either the logit or probit utilities is therefore required when comparing estimated coefficients from the two models.

As with probit, if the V's are set, there are two limiting cases of a binary logit model that are of special interest. The first case is for $\mu \to \infty$:

$$\lim_{\mu \to \infty} P_n(i) = \begin{cases} 1 & \text{if } V_{in} - V_{jn} > 0, \\ 0 & \text{if } V_{in} - V_{jn} < 0; \end{cases}$$

that is, as $\mu \to \infty$, the variance vanishes and the choice model is deterministic. On the other hand, when $\mu \to 0$, the choice probability of i becomes $1/2$ (see Figure 4.7).

The above derivation of the logit model from random utility theory is due to McFadden (1974). As discussed in Section 3.7.2, the logit model has been first derived by Luce (1959) using a constant utility approach and the choice axiom, which is now known as the independence from irrelevant alternatives (IIA) property.

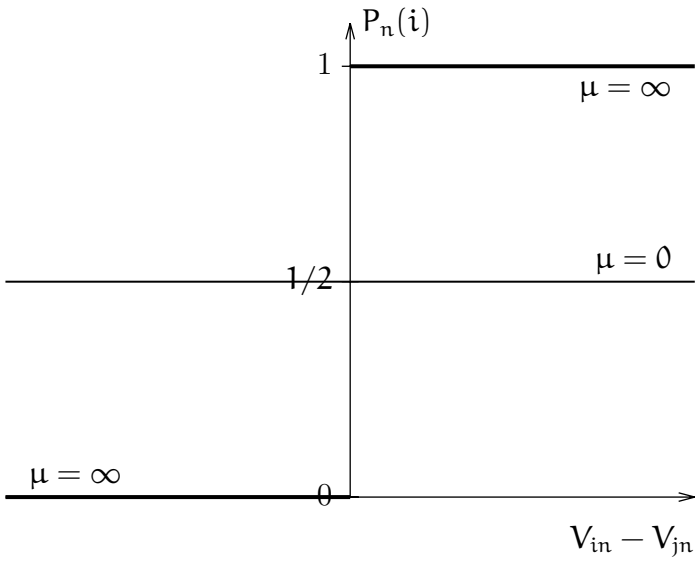


Figure 4.7: The limiting cases of the binary logit model

4.3 Example of Binary Choice Models

To clarify some of the aspects of specifying binary choice models, we consider two examples. The first concerns the choice of transportation modes in the Netherlands (Section 4.3.1), using revealed preference data. The second deals with the choice of airline itinerary (Section 4.3.2), using stated preference data. At this stage we consider only the structure of the particular models and not how their coefficients are estimated. In section 4.5 we return to these examples for a further discussion of statistical issues.

4.3.1 Mode choice in the Netherlands

The example deals with mode choice behavior for intercity travelers in the city of Nijmegen (the Netherlands) using revealed preference data. The survey was conducted during 1987 for the Netherlands Railways to assess factors that influence the choice between rail and car for intercity travel. The city has typical rail connections with the major cities in the western metropolitan area called the Randstad (that contains Amsterdam, Rotterdam and The Hague). Trips from Nijmegen to the Randstad take approximately two hours by both rail and car. The sample consisted of residents of Nijmegen who made a trip in the previous three months to Amsterdam, Rotterdam or The Hague; did not use a yearly rail pass, or other types of pass, which would eliminate the marginal cost of the trip; had the possibility of using a car, namely, possessed a driver's license and had a car available in the household; and had the possibility of using rail, namely, did not have heavy baggage, were not handicapped, and did not need to visit multiple destinations.

Qualifying residents of Nijmegen were identified in a random telephone survey and requested to participate in a home interview. 235 interviews were conducted out of the 365 people who were reached by telephone and satisfied the above criteria. The respondents were requested to report the characteristics of the above-mentioned trip, and those of a trip to the same destination but with the unchosen mode. So the attribute values of both modes were provided by the respondents rather than calculated from network data. After discarding improper data, a total of 228 observations were used to develop the model.

We consider a model where the systematic part of the utility functions is described by the specification table 4.2. The model contains $\mathsf{K} = 9$ coefficients. Some of them are *generic*, meaning that they are included in both alternatives with equal coefficients, some are *alternative specific*. In the specification table, these latter coefficients are simply multiplied by 0 for the irrelevant alternative. We discuss now the role and the nature of the

coefficients.

- Coefficient β_1 is the alternative specific constant. We have arbitrarily decided to define a constant for the car alternative, and constrained the other constant to zero. As described above, it captures the mean of $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$, that is, the difference in the utility of alternative i from that of j when all else is equal. In the specification table, the associated attribute is 1 for the alternative with the constant, and 0 for the other one.

- β_2 is the coefficient of travel cost, measured in Dutch guilders[2]. This coefficient is *generic*, that is identical for both alternatives. It represents an assumption that a guilder has the same marginal (dis)utility whether it is incurred on the car or train mode. We expect coefficient β_2 to be negative since, all else being equal, travelers prefer lower cost alternatives.

- β_3 and β_4 are coefficients of car travel time. These coefficients are *alternative specific*, and capture the marginal (dis)utility of travel time for the car alternative only. Moreover, we explicitly assume that this marginal disutility is different for travelers who are going to work and others, justifying the use of two coefficients. For a given individual, if her trip purpose is "work", then coefficient β_4 is multiplied by 0 in the utility function, and does not play any role. The same is true for coefficient β_3 for non-work related trips. We expect these coefficients to be negative.

- β_5 is the coefficient of train travel time. This coefficient is alternative specific, and is assumed to be the same across the population. Again, we expect this coefficient to be negative.

- Coefficient β_6 measures the impact on the utility of the train if the class preference for rail travel is first class. This coefficient is by nature alternative specific. The associated variable is typically categorical, with two levels (first class / second class). The reference level has been arbitrarily chosen as the second class. Unlike travel time, there is no clear-cut expectation about the sign of this coefficient. A positive value would indicate that a traveler with a preference for first class would, all else being equal, prefer train to car. We refer the reader to Section 5.4 for a detailed description of such specifications.

[2] 1 Dutch guilder =0.45378€

- β_7, β_8 and β_9 are coefficients of *alternative specific socioeconomic* variables. Basically they reflect the differences in preference for car and train as functions of characteristics of the decision-maker. One can view such variables much in the same way as the alternative specific constants, except that instead of using a 0 or 1, each individual has a *different* value of the variables. For any given individual the variable is in effect a constant and corrects the alternative specific constants for that individual. For instance, the constant for a female who is not the main earner in the family and has a fixed arrival time is captured by $1\beta_1 + 0\beta_7 + 0\beta_8 + 1\beta_9 = \beta_1 + \beta_9$, and the constant for a male who is the main earner in the family and has a fixed arrival time is $\beta_1 + \beta_7 + \beta_8 + \beta_9$. As with the alternative specific constants, it would make no sense to define these variables in both alternatives because only difference matters. We have decided to associate them with the car alternative, to be consistent with the alternative specific constant. This is good practice, as it helps in interpreting the value of these coefficients. Assigning them to the other alternative would simply change their sign.

We used the dataset to estimate a binary probit model with this specification. The estimation procedure is described later on in this chapter. In order to illustrate the use of the model, we have considered 3 hypothetical individuals with the attributes and characteristics described in Table 4.3. The estimated values for the coefficients and the corresponding variables for the three individuals are reported in Table 4.4.

Note that the signs of β_2, β_3, β_4 and β_5 are consistent with our expectations. The positive sign of β_1 indicates that, all else being equal, the car benefits from a higher utility than train. The positive sign of β_6 indicates a positive effect of first class preference on train choice. The negative value of β_7 indicates that (all else being equal) male commuters are less likely to use car (compared to train) than female commuters. The positive value of β_8 indicates that the main earner in the family is more likely to use car than other members. Finally, the negative value of β_9 indicates that commuters with a fixed arrival time associate less utility to car.

The probability that individual $n = 1$ chooses to commute by car is

$$\begin{aligned} P_1(\text{car}) &= \Pr(-0.3120 + \varepsilon_{\text{car1}} \geq -1.9551 + \varepsilon_{\text{train1}}) \\ &= \Pr(1.6431 \geq \varepsilon_1), \end{aligned}$$

where $\varepsilon_{\text{train1}} - \varepsilon_{\text{car1}} = \varepsilon_1 \sim N(0, 1)$. Therefore, $P_1(\text{car}) = \Phi(1.6431) = 0.950$. We compute similarly that $P_2(\text{car}) = 0.0792$ and $P_3(\text{car}) = 0.756$ (see Table 4.4).

We have also estimated the parameters of a binary logit model with the exact same specification. The estimated values are reported in the second column of Table 4.5. The interpretation of the signs is consistent with the binary probit model. However, the parameters are different from the parameters of the probit model, so that the models *appear* to be very different. But it is not the case. It is an issue due to the different normalization conditions between the binary probit, where $\sigma = 1$ means that the variance of the difference of the error terms is normalized to 1, and the binary logit, where $\mu = 1$ means that the variance of the difference of the error terms is normalized to $\pi^2/3$. As discussed at the end of Section 4.2, it means that the coefficients of the binary logit must be divided by $\pi/\sqrt{3}$ in order to be compared to the coefficients of the binary probit model. These scaled coefficients are reported in Table 4.6, where it is seen that they are similar. When we apply the binary logit model to individual 1 from Table 4.3, we obtain

$$P_1(\text{car}) = \frac{e^{-0.6642}}{e^{-0.6642} + e^{-3.5504}} = 0.947,$$

and $P_1(\text{train}) = 1 - P_1(\text{car}) = 0.0528$. The probabilities for the other individuals are computed in a similar way, and reported in the last row of Table 4.5. These probabilities are similar to the probit probabilities, although not exactly the same, due to the different distributional assumptions.

4.3.2 Airline itinerary choice

This example deals with the choice of airline itinerary. We use the data collected from the survey conducted by the Boeing Company described in Section 2.2. The respondent was offered three hypothetical choices based on the origin-destination market request that the respondent entered into the itinerary search engine. The first alternative is always a non stop flight (1), the second is always a flight with 1 stop on the same airline (2), and the third is always a flight with 1 stop and a change of airline (3). The respondent was asked to rank the available choices. Demographic data collected included age, gender, income, occupation, and education. Situational variables that were collected included: a) the desired departure time; b) the trip purpose; c) who is paying for the trip; and d) the number of passengers in the travel party. All trips were for origin-destination city pairs in the United States. There are 3609 respondents, each providing one stated preference (SP) response, so the data contains 3609 observations. Note that descriptive statistics are shown in Appendix E.3. In this example, we consider only data corresponding to leisure trips.

In order to illustrate the binary logit model, we focus on the choice between the options "non stop flight" and "one stop–same airline". Therefore, observations in which the respondent selected option 3 are discarded for this analysis, resulting in a data set of 2143 observations.

We consider a model where the systematic part of the utility functions is described in the specification table 4.7. The model contains $K = 9$ coefficients. Some of them are *generic*, meaning that they are included in both alternatives with equal coefficients, some are *alternative specific*. In the specification table, these latter coefficients are simply multiplied by 0 for the irrelevant alternative. We discuss now the role and the nature of the coefficients.

- β_1 is the alternative specific constant. We have arbitrarily decided to define a constant for alternative "one stop–same airline", and to constrain the other constant to zero. As described above, it captures the mean of $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$, that is, the difference in the utility of alternative i from that of j when all else is equal. In the specification table, the associated attribute is 1 for the alternative with the constant, and 0 for the other one.

- β_2 is the coefficient of the round trip fare, measured in hundreds of dollars. This coefficient is *generic*, that is identical for both alternatives. It represents an assumption that \$100 has the same marginal (dis)utility whether it is incurred on alternative "non stop" or "one stop–same airline". We expect coefficient β_2 to be negative since, all else being equal, as travelers prefer lower cost alternatives. Note that the choice of the units (here, hundreds of dollars) is arbitrary. It is good practice to select units in such a way that all pieces data have a similar level of magnitude.

- β_3 is the coefficient of elapsed time, that is the total time of the journey, including transfers. The elapsed time is expressed in hours. This coefficient is *generic* and is assumed to be the same across the population. Again, we expect this coefficient to be negative.

- β_4 and β_5 are coefficients of the leg room. The leg room is expressed in inches. These coefficients are *generic*. We explicitly assume that the marginal utility is different for men and women, justifying the use of two coefficients. β_4 captures the perception of men, whereas β_5 captures the perception of women. For a given individual, if he is a man, β_5 is multiplied by 0 in the utility function, and does not play any role. The same applies for coefficient β_4 for a woman. We expect

these coefficients to be positive, as individuals are supposed to prefer alternatives with a larger room for the legs.

- β_6 is the coefficient of the "being early" variable, expressed in hours. The respondents where asked about the most important between: *the ability to depart your home at a particular time*, and *the need to arrive at your destination at a particular time*. For respondents giving the importance to the departure time, the "being early" variable is defined as the difference between the desired and the scheduled departure times, if the former precedes the latter, and 0 otherwise. For respondents giving the importance to the arrival time, the "being early" variable is defined in the same way, but with the arrival time instead of the departure time. The coefficient is *generic* and expected to be negative.

- β_7 is the coefficient of the "being late" variable, expressed in hours. It is defined as the difference between the desired and scheduled departure or arrival time, if the desired time is after the scheduled time, 0 otherwise, using the same conventions as for the "being early" variable. Its coefficient is *generic* and expected to be negative.

- β_8, β_9 are coefficients of *alternative specific socioeconomic* variables. They reflect the differences in preference for alternatives (1) and (2) as functions of characteristics of the decision-maker. One can view such variables much in the same way as the alternative specific constants, except that instead of using a 0 or 1, each individual has a *different* value of the variables. For any given individual the variable is in effect a constant and corrects the alternative specific constants for that individual. β_8 is associated with a variable which is equal to 1 if the respondent makes more than 2 leisure trips per year, 0 otherwise. β_9 is related to a variable which is equal to 1 if the respondent is a male, 0 otherwise. For instance, the constant for a female who makes more than 2 leisure trips per year is captured by $\beta_1 + 1\beta_8 + 0\beta_9 = \beta_1 + \beta_8$, and the constant for a male who makes less than 2 leisure trips per year is $\beta_1 + 0\beta_8 + 1\beta_9 = \beta_1 + \beta_9$. As with the alternative specific constants, it would make no sense to define these variables in both alternatives because only difference matters. We have decided to associate them with the "one stop–same airline" alternative, to be consistent with the alternative specific constant. This is good practice, as it helps in interpreting the value of these coefficients. Assigning them to the other alternative would simply change their sign.

We used the dataset to estimate a binary probit model with this specification. In order to illustrate the use of the model, we have considered 3

hypothetical individuals with the attributes and characteristics described in Table 4.8. The estimated values for the coefficients and the corresponding variables for the three individuals are reported in Table 4.9.

Note that the signs of β_2, β_3, β_4, β_5, β_6 and β_7 are consistent with our expectations. The negative sign of β_1 indicates that, all else being equal, the alternative "one stop–same airline" benefits from a lower utility than the alternative "non stop". The negative sign of β_8 indicates that the frequent travelers are less likely to choose "one stop–same airline", compared to the other travelers. Finally, the positive sign of β_9 indicates that the men have more propensity to choose alternative "one stop–same airline", compared to women.

The probability that individual $n = 1$ chooses alternative (1) is

$$\begin{aligned} P_1(\text{non stop}) &= \Pr(-2.4299 + \varepsilon_{\text{non stop1}} \geq -3.4029 + \varepsilon_{\text{one stop1}}) \\ &= \Pr(0.937 \geq \varepsilon_1), \end{aligned}$$

where $\varepsilon_{\text{non stop1}} - \varepsilon_{\text{one stop1}} = \varepsilon_1 \sim N(0, 1)$. Therefore, $P_1(\text{non stop}) = \Phi(0.937) = 0.8347$. We compute similarly that $P_2(\text{non stop}) = 0.1155$ and $P_3(\text{non stop}) = 0.2889$ (see Table 4.9).

We have also estimated the parameters of a binary logit model with the exact same specification. The estimated values are reported in the second column of Table 4.10. The interpretation of the signs is consistent with the binary probit model. However, the parameters are significantly different from the parameters of the probit model, so that the models *appear* to be very different. As discussed at the end of Section 4.2, the coefficients of the binary logit must be divided by $\pi/\sqrt{3}$ in order to be compared to the coefficients of the binary probit model. These scaled coefficients are reported in Table 4.11, where it is seen that they are similar. When we apply the binary logit model to individual 1 from Table 4.8, we obtain

$$P_1((1)) = \frac{e^{-4.445}}{e^{-4.445} + e^{-6.0874}} = 0.8379,$$

and $P_1(\text{train}) = 1 - P_1(\text{car}) = 0.0528$. The probabilities for the other individuals are computed in a similar way, and reported in the last row of Table 4.10. These probabilities are similar to the probit probabilities, although not exactly the same, due to the different distributional assumptions.

	Car	Train
β_1	1	0
β_2	cost of trip by car (in Guilders)	cost of trip by train (in Guilders)
β_3	travel time by car (hours) if trip purpose is work, 0 otherwise	0
β_4	travel time by car (hours) if trip purpose is not work, 0 otherwise	0
β_5	0	travel time by train (hours)
β_6	0	1 if first class is preferred, 0 otherwise
β_7	1 if commuter is male, 0 otherwise	0
β_8	1 if commuter is the main earner in the family, 0 otherwise	0
β_9	1 if commuter had a fixed arrival time, 0 otherwise	0

Table 4.2: Specification table of the binary mode choice model in the Netherlands

	Individual 1	Individual 2	Individual 3
Train cost	40.00	7.80	40.00
Car cost	5.00	8.33	3.20
Train travel time	2.50	1.75	2.67
Car travel time	1.17	2.00	2.55
Gender	M	F	F
Trip purpose	Not work	Work	Not work
Class	Second	First	Second
Main earner	No	Yes	Yes
Arrival time	Variable	Fixed	Variable

Table 4.3: Three hypothetical individuals for the choice of transportation mode

			Individual 1		Individual 2		Individual 3	
Variables	Coef.	Value	Car	Train	Car	Train	Car	Train
Car dummy	β_1	1.77	1	0	1	0	1	0
Cost	β_2	-0.0296	5.00	40.00	8.33	7.80	3.20	40.00
Travel time by car (work)	β_3	-1.51	0	0	2.00	0	0	0
Travel time by car (not work)	β_4	-1.26	1.17	0	0	0	2.55	0
Travel time by train	β_5	-0.308	0	2.50	0	1.75	0	2.67
First class dummy	β_6	0.545	0	0	0	1	0	0
Male dummy	β_7	-0.471	1	0	0	0	0	0
Main earner dummy	β_8	0.213	0	0	1	0	1	0
Fixed arrival time dummy	β_9	-0.355	0	0	1	0	0	0
V_{in}			-0.3120	-1.9551	-1.6354	-0.2252	-1.3126	-2.0065
$P_n(i)$			0.950	0.0502	0.0792	0.921	0.756	0.244

Table 4.4: Coefficients of the binary probit model of mode choice in the Netherlands

			Individual 1		Individual 2		Individual 3	
Variables	Coef.	Value	Car	Train	Car	Train	Car	Train
Car dummy	β_1	3.04	1	0	1	0	1	0
Cost	β_2	-0.0527	5.00	40.00	8.33	7.80	3.20	40.00
Travel time by car (work)	β_3	-2.66	0	0	2	0	0	0
Travel time by car (not work)	β_4	-2.22	1.17	0	0	0	2.55	0
Travel time by train	β_5	-0.576	0	2.50	0	1.75	0	2.67
First class dummy	β_6	0.961	0	0	0	1	0	0
Male dummy	β_7	-0.850	1	0	0	0	0	0
Main earner dummy	β_8	0.383	0	0	1	0	1	0
Fixed arrival time dummy	β_9	-0.624	0	0	1	0	0	0
V_{in}			-0.6642	-3.5504	-2.9596	-0.4589	-2.4072	-3.6464
$P_n(i)$			0.947	0.0528	0.0758	0.924	0.775	0.225

Table 4.5: Coefficients of the binary logit model of mode choice in the Netherlands

	Logit	Scaled logit	Probit
β_1	3.04	1.68	1.77
β_2	-0.0527	-0.0291	-0.0296
β_3	-2.66	-1.47	-1.51
β_4	-2.22	-1.22	-1.26
β_5	-0.576	-0.318	-0.308
β_6	0.961	0.530	0.545
β_7	-0.85	-0.469	-0.471
β_8	0.383	0.211	0.213
β_9	-0.624	-0.344	-0.355

Table 4.6: Comparison of logit and probit estimated coefficients, for the choice of transportation mode

	Non stop flight	One stop flight with the same airline
β_1	0	1
β_2	Round trip fare ($100) for non stop flight	Round trip fare ($100) for one stop flight
β_3	Elapsed time (hours) for non stop flight	Elapsed time (hours) for one stop flight
β_4	Leg room (inches) for non stop flight, if male	Leg room (inches) for one stop flight, if male
β_5	Leg room (inches) for non stop flight, if female	Leg room (inches) for one stop flight, if female
β_6	Being early (hours) for non stop flight, at departure or arrival depending on the preference of the respondent	Being early (hours) for one stop flight, at departure or arrival depending on the preference of the respondent
β_7	Being late (hours) for non stop flight, at departure or arrival depending on the preference of the respondent	Being late (hours) for one stop flight, at departure or arrival depending on the preference of the respondent
β_8	0	1 if the respondent makes more than 2 air trips per year
β_9	0	1 if male, 0 otherwise

Table 4.7: Specification table of the binary choice model for the choice of airline itinerary

	Individual 1	Individual 2	Individual 3
Round trip fare for non stop ($100)	2.00	4.00	9.00
Round trip fare for one stop ($100)	2.00	2.00	7.00
Elapsed time for non stop(hours)	1.50	6.00	3.00
Elapsed time for one stop(hours)	2.00	8.00	4.00
Room for the legs for non stop (inches)	1	2	1
Room for the legs for one stop (inches)	1	1	4
Being early for non stop (hours)	1	0	0
Being late for non stop (hours)	0	0	2
Being early for one stop (hours)	0	0	1
Being late for one stop (hours)	2	0	
0 Number of air trips per year	4	3	1
Gender	F	M	F

Table 4.8: Three hypothetical individuals for the choice of airline itinerary

			Individual 1		Individual 2		Individual 3	
Variables	Coef.	Value	Non stop	One stop	Non stop	One stop	Non stop	One stop
One stop–same airline dummy	β_1	-0.685	0	1	0	1	0	1
Round trip fare ($100)	β_2	-1.13	2.00	2.00	4.00	2.00	9.00	8.00
Elapsed time (hours)	β_3	-0.109	1.50	2.00	6.00	8.00	3.00	4.00
Leg room (inches), if male	β_4	0.0752	0	0	2	1	0	0
Leg room (inches), if female	β_5	0.0607	1	1	0	0	1	4
Being early (hours)	β_6	-0.0671	1.00	0	0	0	0	1.00
Being late (hours)	β_7	-0.0528	0	2	0	0	2	0
More than 2 air trips per year, one stop–same airline	β_8	-0.195	0	1	0	1	0	0
Male	β_9	0.111	0	0	0	1	0	0
V_{in}			-2.4299	-3.4029	-5.0236	-3.8258	-10.5419	-9.9853
$P_n(i)$			0.8347	0.1653	0.1155	0.8845	0.2889	0.7111

Table 4.9: Coefficients of the binary probit model of airline itinerary choice

			Individual 1		Individual 2		Individual 3	
Variables	Coef.	Value	Non stop	One stop	Non stop	One stop	Non stop	One stop
One stop–same airline dummy	β_1	-1.10	0	1	0	1	0	1
Round trip fare ($100)	β_2	-2.06	2.00	2.00	4.00	2.00	9.00	8.00
Elapsed time (hours)	β_3	-0.214	1.50	2.00	6.00	8.00	3.00	4.00
Leg room (inches), if male	β_4	0.133	0	0	2	1	0	0
Leg room (inches), if female	β_5	0.122	1	1	0	0	1	4
Being early (hours)	β_6	-0.126	1.00	0	0	0	0	1.00
Being late (hours)	β_7	-0.0922	0	2	0	0	2	0
More than 2 air trips per year, one stop–same airline	β_8	-0.377	0	1	0	1	0	0
Male	β_9	0.182	0	0	0	1	0	0
V_{in}			-4.445	-6.0874	-9.258	-6.994	-19.2444	-18.074
$P_n(i)$			0.8379	0.1621	0.0941	0.9059	0.2368	0.7632

Table 4.10: Coefficients of the binary logit model of airline itinerary choice

	Logit	Scaled logit	Probit
β_1	-1.10	-0.606	-0.685
β_2	-2.06	-1.14	-1.13
β_3	-0.214	-0.118	-0.109
β_4	0.133	0.0733	0.0752
β_5	0.122	0.0673	0.0607
β_6	-0.126	-0.0695	-0.0671
β_7	-0.0922	-0.0508	-0.0528
β_8	-0.377	-0.208	-0.195
β_9	0.182	0.100	0.111

Table 4.11: Comparison of logit and probit estimated coefficients, for the models dealing with the choice of airline itinerary

4.4 Maximum Likelihood Estimation of Binary Choice Models

The model coefficients reflect the sensitivity of the behavior to the variables. To identify them, we use data on behavioral choices describing individuals, what they faced, and what they chose, as described in Chapter 2. Therefore, we turn now to the problem of estimating the values of the unknown parameters $\beta_1,\ldots,\beta_K$ from a sample of observations. We restrict our discussion in this section to the most typical case where our data consists of individuals drawn at random from the population[3].

Each observation consists of the following:

1. An indicator variable defined as
$$y_{in} = \begin{cases} 1 & \text{if person } n \text{ chose alternative } i, \\ 0 & \text{if person } n \text{ chose alternative } j. \end{cases}$$
(Note that y_{jn} is defined trivially by the identity $y_{in} + y_{jn} = 1$.)
2. Two vectors of attributes $x_{in} = h(z_{in}, S_n)$ and $x_{jn} = h(z_{jn}, S_n)$, each containing K values of the relevant variables.

Consider again the three hypothetical individuals introduced in Section 4.3.1, and the choice situation described in Table 4.4. In Section 4.3.1, the value of the coefficients were given, and the model was applied to predict the choice. Here, we do the opposite: we observe the choice made by each individual, and infer the value of the parameters of the model from these observations. For the sake of the example, we assume that individual 1 has chosen the car (alternative i), and individuals 2 and 3 have both chosen the train (alternative j). Using the above notations, we have

$$y_{i1} = 1, y_{j1} = 0, y_{i2} = 0, y_{j2} = 1, y_{i3} = 0, y_{j3} = 1.$$

The values of the variables x are provided in Table 4.10, with $K = 9$:

$$\begin{array}{lccccccccc}
x_{i1} = (1 & 5 & 0 & 1.17 & 0 & 0 & 1 & 0 & 0), \\
x_{j1} = (0 & 40 & 0 & 0 & 2.5 & 0 & 0 & 0 & 0), \\
x_{i2} = (1 & 8.33 & 2 & 0 & 0 & 0 & 0 & 1 & 1), \\
x_{j2} = (0 & 7.8 & 0 & 0 & 1.75 & 1 & 0 & 0 & 0), \\
x_{i3} = (1 & 3.2 & 0 & 2.55 & 0 & 0 & 0 & 1 & 0), \\
x_{j3} = (0 & 40 & 0 & 0 & 2.67 & 0 & 0 & 0 & 0).
\end{array}$$

[3]Extensions to other types of sampling procedures are developed in chapter 11. It is shown there that the results of this chapter are also valid to a wider range of typical stratified sampling strategies.

Given a sample of N observations, our problem then becomes one of finding estimates $\hat{\beta}_1, \ldots, \hat{\beta}_K$ that have some or all of the desirable properties of statistical estimators. We consider in detail the most widely used estimation procedure — maximum likelihood. The maximum likelihood estimators have the following desired properties:

1. They are consistent in the sense of convergence to true values as sample size gets larger.

2. They are asymptotically normally distributed in the sense of the Central Limit Theorem.

3. They are asymptotically efficient, and hence their variance attains the Cramer-Rao lower bound.

These concepts, introduced in Section 1.2, are discussed also in Appendix B.

4.4.1 General Formulation for Maximum Likelihood Estimation of Binary Choice Models

The maximum likelihood estimation (MLE) procedure is conceptually quite straightforward. It consists in identifying the value of the unknown parameters such that the joint probability of the observed choices as predicted by the model is the highest possible. This joint probability is called the *likelihood* of the sample. And it is a function of the parameters of the model.

In the above example, the likelihood of the sample of 3 hypothetical individuals is calculated as follows:

- individual 1 has chosen the car, and this choice is predicted by the model with probability $P_1(i)$,
- individual 2 has chosen the train, and this choice is predicted by the model with probability $P_2(j)$,
- individual 3 has chosen the train, and this choice is predicted by the model with probability $P_3(j)$.

Consequently, the probability that the model predicts all three observations is

$$\mathcal{L}^*(\beta_1, \ldots, \beta_9) = P_1(i)P_2(j)P_3(j). \tag{4.29}$$

If this value is calculated for $\beta_k = 0$, $k = 1, \ldots, K$, we obtain

$$\mathcal{L}^* = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 0.125. \tag{4.30}$$

If this value is calculated for the values of β reported in Table 4.5, we have

$$\mathcal{L}^* = 0.947 \cdot 0.924 \cdot 0.225 = 0.197. \tag{4.31}$$

This can be generalized to a sample of N observations assumed to be independently drawn from the population. As discussed above, the likelihood of the sample is the product of the likelihoods (or probabilities) of the individual observations. It is defined as follows:

$$\mathcal{L}^*(\beta_1, \beta_2, \ldots, \beta_K) = \prod_{n=1}^{N} P_n(i)^{y_{in}} P_n(j)^{y_{jn}}, \tag{4.32}$$

where $P_n(i)$ and $P_n(j)$ are functions of $\beta_1,\ldots,\beta_K$. Note that each factor represents the choice probability of the chosen alternative. Indeed,

$$P_n(i)^{y_{in}} P_n(j)^{y_{jn}} = \begin{cases} P_n(i) & \text{if } y_{in} = 1, y_{jn} = 0 \\ P_n(j) & \text{if } y_{in} = 0, y_{jn} = 1. \end{cases}$$

It is more convenient to analyze the logarithm of $\mathcal{L}^*$, denoted as $\mathcal{L}$ and called the *log likelihood*, because the logarithm of a product of elements is easier to manipulate, being equal to the sum of the logarithms of the elements. Morever, the value of the likelihood is always between 0 and 1, and usually very small, especially when N is large. The range of values of the log likelihood is much larger, as it can take any negative value (from $-\infty$ to 0) and can be represented better in computers. The log likelihood is written as follows:

$$\mathcal{L}(\beta_1, \ldots, \beta_K) = \sum_{n=1}^{N} (y_{in} \ln P_n(i) + y_{jn} \ln P_n(j)), \tag{4.33}$$

or, noting that $y_{jn} = 1 - y_{in}$ and $P_n(j) = 1 - P_n(i)$,

$$\mathcal{L}(\beta) = \mathcal{L}(\beta_1, \ldots, \beta_K) = \sum_{n=1}^{N} (y_{in} \ln P_n(i) + (1 - y_{in}) \ln(1 - P_n(i))), \tag{4.34}$$

where β is the vector with entries $\beta_1, \ldots, \beta_K$. We seek estimates $\hat{\beta}_1$, $\hat{\beta}_2,\ldots,\hat{\beta}_K$ that solve

$$\max \mathcal{L}(\hat{\beta}) = \mathcal{L}(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K), \tag{4.35}$$

where $\hat{\beta}$ is the vector with entries $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K$. The optimization problem is solved using dedicated algorithms, as explained with an example in a subsequent section of this chapter and in Appendix B.7.2.

If a solution exists, it must satisfy the necessary first order conditions (B.95), that is

$$\frac{\partial \mathcal{L}}{\partial \beta_k}(\widehat{\beta}) = \sum_{n=1}^{N} \left(y_{in} \frac{\partial P_n(i)/\partial \beta_k}{P_n(i)} + y_{jn} \frac{\partial P_n(j)/\partial \beta_k}{P_n(j)} \right) = 0, \; k = 1, \ldots, K, \tag{4.36}$$

or in vector form

$$\frac{\partial \mathcal{L}}{\partial \beta}(\widehat{\beta}) = 0. \tag{4.37}$$

The term $\partial \mathcal{L}(\widehat{\beta})/\partial \beta$ is the vector of first derivatives of the log likelihood function with respect to the unknown parameters, evaluated at the estimated value of the parameters. Each entry k of the vector $\partial \mathcal{L}(\widehat{\beta})/\partial \beta$ represents the slope of the multi-dimensional log likelihood function along the corresponding k^{th}axis. If $\widehat{\beta}$ corresponds to a maximum of the function, all these slopes must be zero, justifying (4.36).

Solving the optimization problem requires an iterative procedure. It starts with arbitrary values for the parameters (provided by the analyst, or all set to zero if no value can be guessed). If the first derivatives of the log likelihood function are zero, a solution has been found. If not, they provide information about the slope of the function, and a direction of "hill-climbing" can be identified. This direction is followed for a while, until a new set of values is found, corresponding to a higher log likelihood. The process is restarted from this new set of values, until convergence to the maximum is reached.

A family of algorithms commonly used in practice is called *Newton*'s method. At each iteration ℓ, a quadratic model of the log likelihood function is built around the current iterate $\beta^{(\ell)}$. This quadratic model is such that the value of the model and of its first and second derivatives are the same at $\beta^{(\ell)}$ as the log likelihood function:

$$m(\beta; \beta^{(\ell)}) = \mathcal{L}(\beta^{(\ell)}) + (\beta - \beta^{(\ell)})^T \nabla \mathcal{L}(\beta^{(\ell)}) + \frac{1}{2} (\beta - \beta^{(\ell)})^T \nabla^2 \mathcal{L}(\beta^{(\ell)}) (\beta - \beta^{(\ell)}), \tag{4.38}$$

where $\nabla \mathcal{L}(\beta^{(\ell)})$ is the gradient, that is the vector of the first derivatives of the log likelihood function evaluated at $\beta^{(\ell)}$, and $\nabla^2 \mathcal{L}(\beta^{(\ell)})$ is the matrix of the second derivatives of the log likelihood function evaluated at $\beta^{(\ell)}$. The kth entry of $\mathcal{L}(\beta^{(\ell)})$ is $\partial \mathcal{L}(\beta^{(\ell)})/\partial \beta_k$, and the entry in the kth row and the mth column of $\nabla^2 \mathcal{L}(\beta^{(\ell)})$ is

$$\frac{\partial^2 \mathcal{L}(\beta^{(\ell)})}{\partial \beta_k \partial \beta_m}. \tag{4.39}$$

The approximation of the log likelihood function by the quadratic model is illustrated in Figure 4.8 for a log likelihood function with only one parameter, where both the log likelihood function and the quadratic model at $\beta^{(\ell)}$ are displayed. Note that both functions coincide at $\beta^{(\ell)}$, and have the same slope (first derivative) and curvature (second derivative) at that point. The next iterate is selected as the value of the parameters maximizing the quadratic model, that is

$$\beta^{(\ell+1)} = \beta^{(k)} - \nabla^2\mathcal{L}(\beta^{(\ell)})^{-1}\nabla(\beta^{(\ell)}), \tag{4.40}$$

as illustrated in Figures 4.8 and 4.9 for two successive iterations.

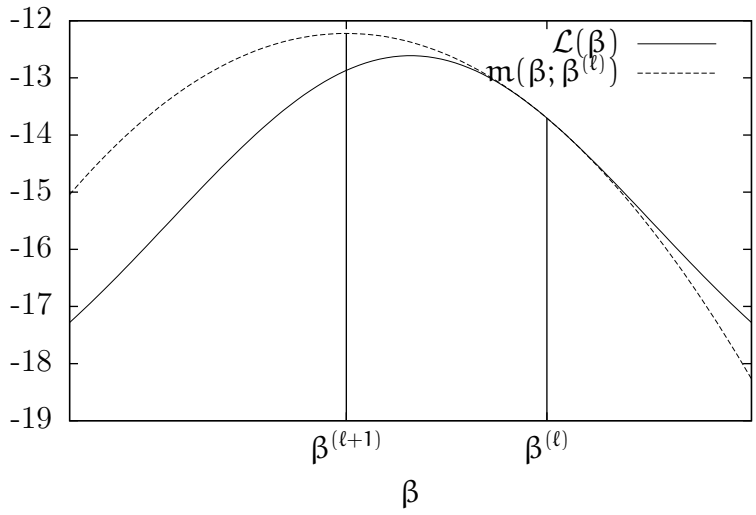


Figure 4.8: Illustration of Newton's method for optimization

It is numerically obtained by solving the system of linear equations

$$\nabla^2\mathcal{L}(\beta^{(\ell)})d = -\nabla(\beta^{(\ell)}), \tag{4.41}$$

to obtain the direction d, and then calculating

$$\beta^{(\ell+1)} = \beta^{(\ell)} + d. \tag{4.42}$$

The procedure continues until the gradient is sufficiently close to zero, depending on the level of precision that is required. In practice, it happens when the norm of the gradient is below a user-specified threshold Γ, that is

$$\left\| \frac{\partial\mathcal{L}(\beta)}{\partial\beta} \right\| = \sqrt{\sum_k \left(\frac{\partial\mathcal{L}(\beta)}{\partial\beta_k} \right)^2} \leq \Gamma.$$

A typical value for Γ is 10^{-6}.

Actually, the method described above is not guaranteed to converge, and variants involving a scaled version of d have to be used, that is

$$\beta^{(\ell+1)} = \beta^{(\ell)} + \alpha d, \ \alpha > 0. \tag{4.43}$$

We refer the reader to Appendix B.7.2 and Bierlaire (2015) for more details on optimization algorithms.

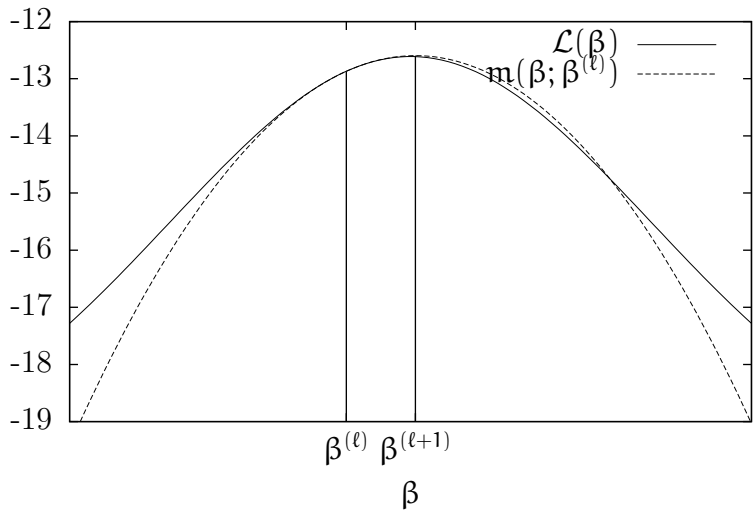


Figure 4.9: Illustration of Newton's method for optimization: second iteration

To illustrate the procedure, consider again the mode choice in the Netherlands described in Section 4.3.1, and consider a simple specification with only two parameters: the alternative specific constant of the car alternative (β_1), and the generic coefficient of travel time (β_2), as described in Table 4.12. The value of the log likelihood, as a function of the two parameters, is depicted at Figure 4.10, where the level curves of the function are also reported.

	Car	Train
β_1	1	0
β_2	travel time by car (hours)	travel time by train (hours)

Table 4.12: Mode choice in the Netherlands: specification table with two coefficients

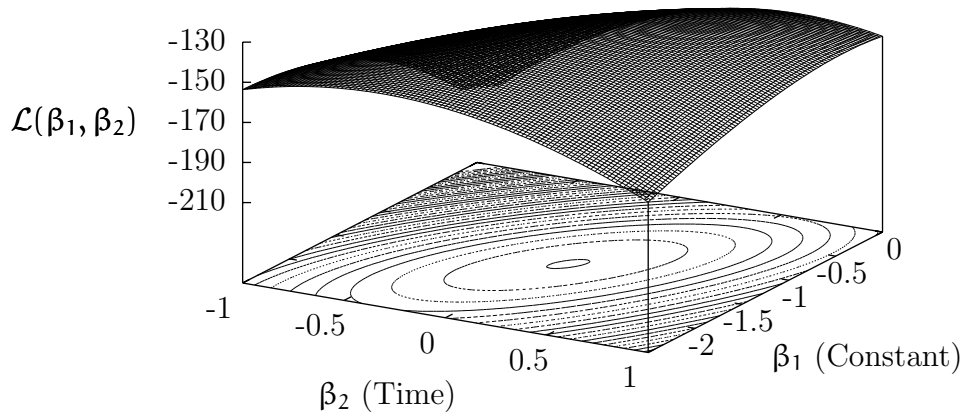


Figure 4.10: Log likelihood function for the binary logit model of mode choice in the Netherlands with two parameters (Table 4.12)

The iterates generated by Newton's method, started from $\beta^{(0)} = (0, 0)$, are reported in Table 4.13. The two first iterations are represented by two arrows superposed on the level curves of the function, in Figure 4.11.

Iteration	β_1	β_2
0	0.0	0.0
1	0.0855	-1.0
2	0.0707	-1.31
3	0.0676	-1.33

Table 4.13: Iterations of Newton's method to find the maximum likelihood estimates of the binary logit model of mode choice in the Netherlands with two parameters (Table 4.12)

In many cases of practical interest, including logit and probit with linear-in-parameters specification, we can show that the likelihood function is globally concave, so that if a solution to the first order conditions exists, it is unique. However, it is quite possible that there are multiple solutions to the first order conditions.

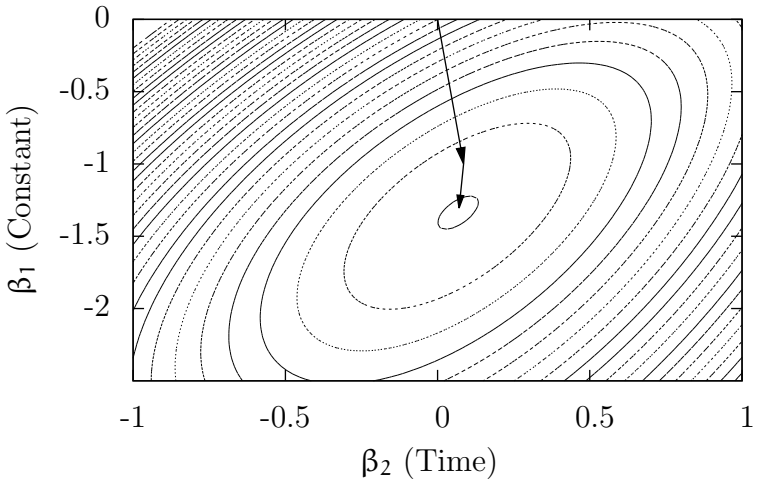


Figure 4.11: Level curves of the log likelihood function for the binary logit model of mode choice in the Netherlands with two parameters (Table 4.12), and two iterations

4.4.2 Variance-covariance of the estimates

In addition to the role that the second derivatives matrix of the log likelihood function $\nabla^2\mathcal{L}(\beta)$ has in some optimization algorithms, it is also used to compute an estimate of the variance-covariance matrix of the parameter estimates, from which standard errors and p values are generated.

Under relatively general conditions, the asymptotic variance-covariance matrix of the maximum likelihood estimates is given by the Cramer-Rao bound

$$-\mathrm{E}\left[\nabla^2\mathcal{L}(\beta)\right]^{-1} = \left\{-\mathrm{E}\left[\frac{\partial^2\mathcal{L}(\beta)}{\partial\beta\partial\beta^\intercal}\right]\right\}^{-1}. \tag{4.44}$$

From the second order optimality conditions (see Section B.7), this matrix is negative definite if the local maximum is unique, which is the algebraic equivalent of the local strict concavity of the log likelihood function.

Since we do not know the actual values of the parameters at which to evaluate the second derivatives, or the distribution of x_{in} and x_{jn} over which to take their expected value, we estimate the variance-covariance matrix by evaluating the second derivatives at the estimated parameters $\hat{\beta}$ and the sample distribution of x_{in} and x_{jn} instead of their true distribution. Thus we

use

$$\mathrm{E}\left[\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta_k \partial \beta_m}\right] \approx \sum_{n=1}^{N} \left[\frac{\partial^2 \left(y_{in} \ln P_n(i) + y_{jn} \ln P_n(j)\right)}{\partial \beta_k \partial \beta_m}\right]_{\beta=\hat{\beta}}, \tag{4.45}$$

as a consistent estimator of the matrix of second derivatives. Denote this matrix as $\hat{A}$. Therefore, an estimate of the Cramer-Rao bound (4.44) is given by

$$\widehat{\Sigma}_{\beta}^{\mathrm{CR}} = -\hat{A}^{-1}. \tag{4.46}$$

If the matrix $\hat{A}$ is negative definite then $-\hat{A}$ is invertible and the Cramer-Rao bound is positive definite. However, this is not guaranteed.

Another consistent estimator of the (negative of the) second derivatives matrix can be obtained by the matrix of the cross-products of first derivatives as follows:

$$-\mathrm{E}\left[\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta \partial \beta^{\mathsf{T}}}\right] \approx \sum_{n=1}^{n} \nabla L_n(\hat{\beta}) \nabla L_n(\hat{\beta})^{\mathsf{T}} = \hat{B}, \tag{4.47}$$

where

$$\nabla L_n(\hat{\beta}) = \nabla(y_{in} \ln P_n(i) + y_{jn} \ln P_n(j)) \tag{4.48}$$

is the gradient vector of the log likelihood of observation n. As the gradient $\nabla L_n(\hat{\beta})$ is a column vector of dimension $K \times 1$, and its transpose $\nabla L_n(\hat{\beta})^{\mathsf{T}}$ is a row vector of size $1 \times K$, the product $\nabla L_n(\hat{\beta}) \nabla L_n(\hat{\beta})^{\mathsf{T}}$ appearing for each observation n in (4.47) is a rank one matrix of size $K \times K$. The approximation $\hat{B}$ is employed by the BHHH algorithm and replaces the second derivative matrix $\nabla^2 \mathcal{L}(\beta^{(\ell)})$ in (4.40) and (4.41) (Berndt et al., 1974). It can also provide an estimate of the variance-covariance matrix:

$$\widehat{\Sigma}_{\beta}^{\mathrm{BHHH}} = \hat{B}^{-1}, \tag{4.49}$$

although this estimate is rarely used. Instead, $\hat{B}$ is used to derive a third consistent estimator of the variance-covariance matrix of the parameters, defined as

$$\widehat{\Sigma}_{\beta}^{\mathrm{R}} = (-\hat{A})^{-1}\, \widehat{B}\, (-\hat{A})^{-1} = \widehat{\Sigma}_{\beta}^{\mathrm{CR}}\, (\widehat{\Sigma}_{\beta}^{\mathrm{BHHH}})^{-1}\, \widehat{\Sigma}_{\beta}^{\mathrm{CR}}. \tag{4.50}$$

It is called the *robust* estimator, or sometimes the *sandwich* estimator, due to the form of equation (4.50). An example of these estimators is given in Section 4.5.

When the true likelihood function is maximized, these estimators are asymptotically equivalent, and the Cramer-Rao bound (4.44) should be preferred (Kauermann and Carroll, 2001). When other consistent estimators

are used, different from the maximum lilelihood, the robust estimator (4.50) must be used (White, 1982). Consistent non-maximum likelihood estimators, known as pseudo maximum likelihood estimators, are often used when the true likelihood function is unknown or difficult to compute. In such cases, it is often possible to obtain consistent estimators by maximizing an objective function based on a simplified probability distribution. We discuss the need for estimators different from the maximum likelihood when we investigate various sampling strategies in Chapter 11.

4.4.3 Binary logit

In the case of binary logit the log likelihood function is obtained by substituting (4.27) into (4.33):

$$\mathcal{L} = \sum_{n=1}^{N} \left(y_{in} \ln \left(\frac{e^{V_{in}}}{e^{V_{in}} + e^{V_{jn}}} \right) + y_{jn} \ln \left(\frac{e^{V_{jn}}}{e^{V_{in}} + e^{V_{jn}}} \right) \right). \tag{4.51}$$

Note that the scale parameter μ is normalized to 1, as discussed in Section 4.1. For the sake of notational simplicity we define

$$V_n = V_{in} - V_{jn}. \tag{4.52}$$

In this notation

$$P_n(i) = \frac{1}{1 + e^{-V_n}} \text{ and } P_n(j) = \frac{e^{-V_n}}{1 + e^{-V_n}}.$$

To find the maximum, we must set the derivatives to zero. First, note that

$$\frac{\partial \ln P_n(i)}{\partial V_n} = \frac{e^{-V_n}}{(1 + e^{-V_n})^2}(1 + e^{-V_n}) = P_n(j), \tag{4.53}$$

and

$$\frac{\partial \ln P_n(j)}{\partial V_n} = -P_n(i). \tag{4.54}$$

So, applying the chain rule, we can calculate the first derivatives as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \beta_k} &= \sum_{n=1}^{N} \left(y_{in} P_n(j) \frac{\partial V_n}{\partial \beta_k} - y_{jn} P_n(i) \frac{\partial V_n}{\partial \beta_k} \right) \\ &= \sum_{n=1}^{N} \left(y_{in}(1 - P_n(i)) - (1 - y_{in}) P_n(i) \right) \frac{\partial V_n}{\partial \beta_k} \qquad (4.55) \\ &= \sum_{n=1}^{N} \left(y_{in} - P_n(i) \right) \frac{\partial V_n}{\partial \beta_k}, \; k = 1, \ldots, K. \end{aligned}$$

If the model is linear in the parameters, we define $x_n = x_{in} - x_{jn}$ so that

$$V_n = \sum_{k=1}^{K} \beta_k (x_n)_k \text{ and } \frac{\partial V_n}{\partial \beta_k} = (x_n)_k, \tag{4.56}$$

where $(x_n)_k$ denotes the kth entry of vector x_n.

The first order optimality condition (4.36) of the optimization problem is a system of K nonlinear equations in K unknowns $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K$ of the form

$$\sum_{n=1}^{N} (y_{in} - P_n(i))(x_n)_k = 0, \; k = 1, \ldots, K. \tag{4.57}$$

The entries of the second derivatives matrix can be straightforwardly derived from (4.55) and (4.56):

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_k \partial \beta_m} = -\sum_{n=1}^{N} P_n(i)(1 - P_n(i))(x_n)_k (x_n)_m. \tag{4.58}$$

Under certain regularity conditions, if there is a solution to the first-order conditions of equation (4.57), it is the only solution. To show this, we need only to prove that the log likelihood is globally concave. A sufficient condition for this is that the matrix of second derivatives is negative semidefinite for all values of β. If we denote D as an $N \times K$ matrix of real values with entries

$$d_{nk} = (x_n)_k \sqrt{P_n(i)(1 - P_n(i))},$$

then the matrix

$$D^T D = -\nabla^2 \mathcal{L}(\beta) = -\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta \partial \beta^T}.$$

Since all the entries in D are real, $D^T D$ is positive semidefinite, and $\nabla^2 \mathcal{L}(\beta)$ negative semidefinite for any $\beta \in \mathbb{R}^K$, implying that $\mathcal{L}$ is concave. Thus, we have shown that any vector $\hat{\beta}$ such that $\partial \mathcal{L}/\partial \beta(\hat{\beta}) = 0$ is a maximum likelihood estimate (see Section B.7.1). Note that this property applies to logit models with more than two alternatives, when the utility function is linear in parameters.

A few aspects of the maximum likelihood estimates for logit are worth noting. First, suppose that $(x_n)_k$ always equals 1 and the associated coefficient is an alternative specific constant. In this instance the kth first-order condition simplifies to

$$\sum_{n=1}^{N} (y_{in} - P_n(i)) = 0, \tag{4.59}$$

implying that

$$\sum_{n=1}^{N} y_{in} = \sum_{n=1}^{N} P_n(i). \tag{4.60}$$

In words, this means that the total number of individuals in the sample who are observed choosing alternative i equals the sum (taken over the entire sample) of the choice probabilities when evaluated at the maximum likelihood estimates. (Since $y_{jn} = 1 - y_{in}$ and $P_n(j) = 1 - P_n(i)$, this holds for alternative j as well.) This property holds regardless of the specification of the other variables. Viewed another way, the maximum likelihood estimators have the desirable property that the predicted share choosing i (of the sample on which the model was estimated) equals the observed share in the sample. This property extends to the case of alternative specific dummy variables. Suppose a variable is defined as

$$(x_{in})_k = \begin{cases} 1 & \text{for some subset of the sample,} \\ 0 & \text{for the remainder of the sample,} \end{cases}$$

$$(x_{jn})_k = 0.$$

In this case $(x_n)_k = (x_{in})_k - (x_{jn})_k$ is 1 for the relevant subset of the sample, and 0 otherwise. Now let us order the observations so that the first N' falls into the group for which $(x_n)_k = 1$. The kth first-order condition then reduces to

$$\sum_{n=1}^{N'} (y_{in} - P_n(i)) = 0, \tag{4.61}$$

or

$$\sum_{n=1}^{N'} y_{in} = \sum_{n=1}^{N'} P_n(i). \tag{4.62}$$

In words, this implies that *within the subset* of the sample defined by the alternative specific dummy variable, the sum of the predicted choice probabilities for alternative i equals the number of people actually choosing it.

We can also use the first-order conditions to obtain a slightly different perspective on why it is meaningless to have two alternative specific constants, one in each alternative. Basically one can show that having two constants produces a situation where there are multiple solutions to the first-order conditions. If the constants are the first two variables in x_n, and if $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K)$ is a solution to equation (4.57), then it is easy to show that $(\hat{\beta}_1 + \alpha, \hat{\beta}_2 + \alpha, \hat{\beta}_3, \ldots, \hat{\beta}_K)$ is also a solution for any constant α. Indeed, the choice probabilities are unaffected by a shift of both constants, and so are the first

order conditions (4.57). In a geometric sense the log likelihood function has a "ridge" of points, all of which are maxima.

4.4.4 Binary probit

Solution for the maximum likelihood estimates for binary probit exactly parallels the solution for logit. Using the same notational convention,

$$\mathcal{L} = \sum_{n=1}^{N} \left(y_{in} \ln \Phi(\beta^T x_n) + (1 - y_{in}) \ln(1 - \Phi(\beta^T x_n)) \right). \tag{4.63}$$

Using the fact that $\Phi'(x) = \phi(x)$, where $\phi(\cdot)$ denotes the probability density function of the standardized normal distribution, the first derivatives are

$$\frac{\partial \mathcal{L}}{\partial \beta_k} = \sum_{n=1}^{N} \left(\frac{y_{in} \phi(\beta^T x_n)}{\Phi(\beta^T x_n)} - \frac{(1 - y_{in}) \phi(\beta^T x_n)}{1 - \Phi(\beta^T x_n)} \right) (x_n)_k, \; k = 1, \ldots, K, \tag{4.64}$$

where ϕ is the standardized normal density function, and Φ the corresponding CDF, so that

$$\frac{\partial \Phi(\beta^T x_n)}{\partial \beta_k} = \phi(\beta^T x_n)(x_n)_k, \tag{4.65}$$

and

$$\frac{\partial \phi(\beta^T x_n)}{\partial \beta_m} = -\beta^T x_n \phi(\beta^T x_n)(x_n)_m. \tag{4.66}$$

Using the fact that $\phi'(x) = -x\phi(x)$, the second derivatives are, for $k, m = 1, \ldots, K$,

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_k \partial \beta_m} = \sum_{n=1}^{N} \Bigg(-y_{in} \left(\frac{\beta^T x_n}{\Phi(\beta^T x_n)} + \frac{\phi(\beta^T x_n)}{\Phi(\beta^T x_n)^2} \right) \tag{4.67}$$

$$+ (1 - y_{in}) \left(\frac{\beta^T x_n}{1 - \Phi(\beta^T x_n)} - \frac{\phi(\beta^T x_n)}{(1 - \Phi(\beta^T x_n))^2} \right) \Bigg) \tag{4.68}$$

$$\phi(\beta^T x_n)(x_n)_k (x_n)_m. \tag{4.69}$$

We again use these derivatives in the iterative algorithm (4.40). As in the case of binary logit, under certain regularity conditions this solution can be shown to be unique (see Daganzo, 1980).

4.5 Examples of Maximum Likelihood Estimation

4.5.1 Simple Example Revisited

To illustrate the application of maximum likelihood to binary logit estimation, we return to the simple two-variable problem described in equation (4.7):

$$\begin{array}{rcccl} V_{An} & = & \beta_1 & + & \beta_2 t_{An} \\ V_{Tn} & = & & & \beta_2 t_{Tn}. \end{array}$$

In this model the two alternatives are labeled auto A and transit T. The two variables correspond to an alternative specific constant and a single generic travel time variable. To estimate the model, we need a sample of N observations containing each the value of t_{An}, t_{Bn} and the choice.

We have extracted 25 observations from the database used for the mode choice example in Section 4.3.2. This small sample is shown in Table 4.14. The log likelihood function is depicted at Figure 4.12. At the first iteration, the parameters are initialized to $\beta_1 = \beta_2 = 0$. The contribution of each observation to the log likelihood function and to its derivatives are reported in Table 4.15. We obtain the log likelihood function $\mathcal{L}(0,0) = -17.328680$, its first derivatives:

$$\nabla \mathcal{L}(0,0) = \begin{pmatrix} \frac{\partial \mathcal{L}}{\partial \beta_1}(0,0) \\ \frac{\partial \mathcal{L}}{\partial \beta_2}(0,0) \end{pmatrix} = \begin{pmatrix} 5.500000 \\ -4.381500 \end{pmatrix}, \tag{4.70}$$

and its second derivatives:

$$\nabla^2 \mathcal{L}(0,0) = \begin{pmatrix} \frac{\partial^2 \mathcal{L}}{\partial \beta_1^2}(0,0) & \frac{\partial^2 \mathcal{L}}{\partial \beta_1 \partial \beta_2}(0,0) \\ \frac{\partial^2 \mathcal{L}}{\partial \beta_1 \partial \beta_2}(0,0) & \frac{\partial^2 \mathcal{L}}{\partial \beta_2^2}(0,0) \end{pmatrix} = \begin{pmatrix} -6.250000 & 2.200250 \\ 2.200250 & -2.291771 \end{pmatrix}. \tag{4.71}$$

Therefore, 4.41 is written:

$$\begin{pmatrix} -6.250000 & 2.200250 \\ 2.200250 & -2.291771 \end{pmatrix} d = - \begin{pmatrix} 5.500000 \\ -4.381500 \end{pmatrix}, \tag{4.72}$$

so that

$$d = \begin{pmatrix} 0.31261 \\ -1.61171 \end{pmatrix} \text{ and } \beta^{(1)} = \beta^{(0)} + d = \begin{pmatrix} 0.31261 \\ -1.61171 \end{pmatrix}, \tag{4.73}$$

where $\beta^{(1)}$ is the new estimate of the parameter proposed by the algorithm.

Obs.	transit time ($\mathtt{t}_{\mathtt{Tn}}$)	car time ($\mathtt{t}_{\mathtt{An}}$)	Choice
1	1.916	1.283	Car
2	1.833	1.416	Car
3	2.084	1.417	Car
4	2.25	1.533	Car
5	2.083	1.517	Car
6	1.583	1.583	Rail
7	1.25	1.517	Car
8	1.917	2.084	Car
9	2.416	1.283	Rail
10	1.75	1.583	Car
11	1.883	1.667	Car
12	2.416	1.583	Car
13	1.717	1.583	Rail
14	2.75	2.083	Car
15	1.866	1.583	Car
16	1.834	2.583	Rail
17	2.5	1.517	Car
18	2.167	1.583	Car
19	1.5	2.333	Rail
20	2.25	1.434	Car
21	2.25	1.533	Car
22	2.333	1.55	Car
23	2.75	2.583	Car
24	2.117	2.25	Rail
25	2.5	2.033	Rail

Table 4.14: Data for simple binary example

For the second iteration, the contribution of each observation to the log likelihood function and to its derivatives are reported in Table 4.16. Now, 4.41 is written:

$$\begin{pmatrix} -4.674359 & 1.222222 \\ 1.222222 & -1.382656 \end{pmatrix} \mathrm{d} = - \begin{pmatrix} 0.800482 \\ -0.689803 \end{pmatrix}, \tag{4.74}$$

so that

$$\mathrm{d} = \begin{pmatrix} 0.053067 \\ -0.451988 \end{pmatrix} \text{ and } \beta^{(2)} = \beta^{(1)} + \mathrm{d} = \begin{pmatrix} 0.36568 \\ -2.06370 \end{pmatrix}. \tag{4.75}$$

Obs.	$\mathcal{L}$	$\frac{\partial \mathcal{L}}{\partial \beta_1}$	$\frac{\partial \mathcal{L}}{\partial \beta_2}$	$\frac{\partial^2 \mathcal{L}}{\partial \beta_1^2}$	$\frac{\partial^2 \mathcal{L}}{\partial \beta_1 \beta_2}$	$\frac{\partial^2 \mathcal{L}}{\partial \beta_2^2}$
1	-0.693147	0.500000	-0.316500	-0.250000	0.158250	-0.100172
2	-0.693147	0.500000	-0.208500	-0.250000	0.104250	-0.043472
3	-0.693147	0.500000	-0.333500	-0.250000	0.166750	-0.111222
4	-0.693147	0.500000	-0.358500	-0.250000	0.179250	-0.128522
5	-0.693147	0.500000	-0.283000	-0.250000	0.141500	-0.080089
6	-0.693147	-0.500000	-0.000000	-0.250000	-0.000000	-0.000000
7	-0.693147	0.500000	0.133500	-0.250000	-0.066750	-0.017822
8	-0.693147	0.500000	0.083500	-0.250000	-0.041750	-0.006972
9	-0.693147	-0.500000	0.566500	-0.250000	0.283250	-0.320922
10	-0.693147	0.500000	-0.083500	-0.250000	0.041750	-0.006972
11	-0.693147	0.500000	-0.108000	-0.250000	0.054000	-0.011664
12	-0.693147	0.500000	-0.416500	-0.250000	0.208250	-0.173472
13	-0.693147	-0.500000	0.067000	-0.250000	0.033500	-0.004489
14	-0.693147	0.500000	-0.333500	-0.250000	0.166750	-0.111222
15	-0.693147	0.500000	-0.141500	-0.250000	0.070750	-0.020022
16	-0.693147	-0.500000	-0.374500	-0.250000	-0.187250	-0.140250
17	-0.693147	0.500000	-0.491500	-0.250000	0.245750	-0.241572
18	-0.693147	0.500000	-0.292000	-0.250000	0.146000	-0.085264
19	-0.693147	-0.500000	-0.416500	-0.250000	-0.208250	-0.173472
20	-0.693147	0.500000	-0.408000	-0.250000	0.204000	-0.166464
21	-0.693147	0.500000	-0.358500	-0.250000	0.179250	-0.128522
22	-0.693147	0.500000	-0.391500	-0.250000	0.195750	-0.153272
23	-0.693147	0.500000	-0.083500	-0.250000	0.041750	-0.006972
24	-0.693147	-0.500000	-0.066500	-0.250000	-0.033250	-0.004422
25	-0.693147	-0.500000	0.233500	-0.250000	0.116750	-0.054522
Total	-17.328680	5.500000	-4.381500	-6.250000	2.200250	-2.291771

Table 4.15: Log likelihood and its derivatives, $\beta_1 = \beta_2 = 0$

Obs.	$\mathcal{L}$	$\frac{\partial \mathcal{L}}{\partial \beta_1}$	$\frac{\partial \mathcal{L}}{\partial \beta_2}$	$\frac{\partial^2 \mathcal{L}}{\partial \beta_1^2}$	$\frac{\partial^2 \mathcal{L}}{\partial \beta_1 \beta_2}$	$\frac{\partial^2 \mathcal{L}}{\partial \beta_2^2}$
1	-0.234069	0.208693	-0.132103	-0.165140	0.104534	-0.066170
2	-0.317401	0.271962	-0.113408	-0.197998	0.082565	-0.034430
3	-0.222878	0.199788	-0.133259	-0.159873	0.106635	-0.071126
4	-0.207289	0.187215	-0.134233	-0.152166	0.109103	-0.078227
5	-0.257588	0.227086	-0.128531	-0.175518	0.099343	-0.056228
6	-0.861618	-0.577522	-0.000000	-0.243990	-0.000000	-0.000000
7	-0.753737	0.529395	0.141349	-0.249136	-0.066519	-0.017761
8	-0.671656	0.489138	0.081686	-0.249882	-0.041730	-0.006969
9	-2.250049	-0.894606	1.013589	-0.094286	0.106826	-0.121034
10	-0.443987	0.358526	-0.059874	-0.229985	0.038408	-0.006414
11	-0.416385	0.340574	-0.073564	-0.224583	0.048510	-0.010478
12	-0.174844	0.160412	-0.133623	-0.134680	0.112189	-0.093453
13	-0.991962	-0.629152	0.084306	-0.233320	0.031265	-0.004189
14	-0.222878	0.199788	-0.133259	-0.159873	0.106635	-0.071126
15	-0.380902	0.316755	-0.089642	-0.216421	0.061247	-0.017333
16	-0.342729	-0.290170	-0.217337	-0.205971	-0.154272	-0.115550
17	-0.139788	0.130457	-0.128240	-0.113438	0.111510	-0.109614
18	-0.251073	0.222035	-0.129668	-0.172735	0.100877	-0.058912
19	-0.305296	-0.263095	-0.219158	-0.193876	-0.161499	-0.134528
20	-0.179290	0.164137	-0.133936	-0.137196	0.111952	-0.091353
21	-0.207289	0.187215	-0.134233	-0.152166	0.109103	-0.078227
22	-0.188216	0.171565	-0.134335	-0.142130	0.111288	-0.087138
23	-0.443987	0.358526	-0.059874	-0.229985	0.038408	-0.006414
24	-0.743480	-0.524543	-0.069764	-0.249398	-0.033170	-0.004412
25	-1.361399	-0.743698	0.347307	-0.190611	0.089015	-0.041570
Total	-12.569793	0.800482	-0.689803	-4.674359	1.222222	-1.382656

Table 4.16: Log likelihood and its derivatives, $\beta_1 = 0.31261$, $\beta_2 = -1.61171$

The iterations of the algorithm on the simple example with 25 observations are summarized in Table 4.17, reporting the log likelihood $\mathcal{L}(\beta)$, its first derivative

$$\nabla\mathcal{L}(\beta) = \begin{pmatrix} \partial\mathcal{L}/\partial\beta_1 \\ \partial\mathcal{L}/\partial\beta_2 \end{pmatrix},$$

and its second derivatives

$$\nabla^2\mathcal{L}(\beta) = \begin{pmatrix} \frac{\partial^2\mathcal{L}}{\partial\beta_1^2}(\beta_1,\beta_2) & \frac{\partial^2\mathcal{L}}{\partial\beta_1\partial\beta_2}(\beta_1,\beta_2) \\ \frac{\partial^2\mathcal{L}}{\partial\beta_1\partial\beta_2}(\beta_1,\beta_2) & \frac{\partial^2\mathcal{L}}{\partial\beta_2^2}(\beta_1,\beta_2) \end{pmatrix}.$$

The progress of the iterations on the level curves of the function is represented on Figure 4.13. Note that, in this example, the first two steps are large, after which the parameters adjust marginally.

The raw output of the estimation procedure is

- The optimal solution $\widehat{\beta} = (0.371513, -2.130979)$;
- The gradient $\partial\mathcal{L}(\widehat{\beta})/\partial\beta = (-3.33067e-16, -2.22045e-16)$;
- The second derivatives matrix

$$\frac{\partial^2\mathcal{L}(\widehat{\beta})}{\partial\beta\partial\beta^\intercal} = \begin{pmatrix} -4.02971 & 0.885865 \\ 0.885865 & -1.04576 \end{pmatrix};$$

- The BHHH matrix (4.47)

$$\widehat{B} = \begin{pmatrix} 3.84142 & -1.36155 \\ -1.36155 & 1.49193 \end{pmatrix}.$$

From these raw outputs, we can derive an estimate of the variance-covariance matrix, used to calculate t statistics and p values. We report here two estimates, one from the Cramer-Rao bound (4.44):

$$\widehat{\Sigma}_\beta^{CR} = \left(-\begin{pmatrix} -4.02971 & 0.885865 \\ 0.885865 & -1.04576 \end{pmatrix}\right)^{-1} = \begin{pmatrix} 0.304944 & 0.25832 \\ 0.25832 & 1.17507 \end{pmatrix},$$

and one from the robust estimator (4.50):

$$\widehat{\Sigma}_\beta^{R} = \widehat{\Sigma}_\beta^{CR}\begin{pmatrix} 3.84142 & -1.36155 \\ -1.36155 & 1.49193 \end{pmatrix}\widehat{\Sigma}_\beta^{CR} = \begin{pmatrix} 0.242265 & 0.176726 \\ 0.176726 & 1.4898 \end{pmatrix}.$$

Typically, the results of the estimation are presented in two sections: summary statistics about the estimation, and details on the parameters estimates. The exact list of summary statistics depend on the estimation software. The following are often reported (see Table 4.18). Their use is briefly introduced with an example later in this chapter, and detailed in Chapter 6.

Number of parameters The number K of estimated parameters.

Number of observations The number N of observations actually used for the estimation.

Null log likelihood the value $\mathcal{L}(0)$ of the log likelihood function when all the parameters are zero. In binary choice models it is the log likelihood of the most naive possible model, that is, one in which the choice probabilities are $1/2$ for each of the two alternatives. Consequently, $\mathcal{L}(0) = -N \ln 2$.

Constant log likelihood the value $\mathcal{L}(c)$ of the log likelihood function when only an alternative specific constant is included. This corresponds to the log likelihood for another naive model in which the choice probability for each alternative simply equals the fraction of the sample choosing the alternative. In this simple case 7 of the 25 observed chose rail, and 18 chose auto, so for the naive model $P_n(\text{auto}) = 18/25 = 0.72$ and $P_n(\text{rail}) = 7/25 = 0.28$. Consequently, $\mathcal{L}(c) = 18 \ln(0.72) + 7 \ln(0.28) =$ -14.824. $\mathcal{L}(c)$ is greater than or equal to $\mathcal{L}(0)$.

Final log likelihood the value of the log likelihood function at its maximum, $\mathcal{L}(\widehat{\beta})$.

Likelihood ratio test statistic used to test the null hypothesis that all the parameters are zero, and is defined as $-2(\mathcal{L}(0) - \mathcal{L}(\hat{\beta}))$. It is asymptotically distributed as χ^2 with K degrees of freedom (see section 6.5.5). Here $K = 2$, and the value of the statistic is 9.90, which indicates that we can reject the null hypothesis that all the parameters are zero at the 0.01 level of significance. Note that, in practice, this test is rarely useful, as the null hypothesis is almost always rejected. We discuss more useful applications of the likelihood ratio test in Chapter 6.

Rho-square Denoted by ρ^2, it is an informal goodness-of-fit index that measures the fraction of an initial log likelihood value explained by the model. It is defined as $1 - (\mathcal{L}(\hat{\beta})/\mathcal{L}(0))$. For a binary choice model with an alternative specific constant, ρ^2 must lie between 0 and 1. ρ^2 is analogous to R^2 used in regression, but it should be used with somewhat more caution. While R^2 has a concrete meaning (it is the percentage of the dependent variable variation that is explained by a linear model), ρ^2 does not have such a meaning. Values of ρ^2 depend on the type of model being estimated. The measure is only useful in comparing two specifications developed on the exact same data. Here $\rho^2 = 0.576$. Although ρ^2 is most often used informally, it can be shown

that $\rho^2/(1-\rho^2)$ is asymptotically distributed as F with (K, K) degrees of freedom. ρ^2 is occasionally defined as $1-(\mathcal{L}(\hat{\beta})/\mathcal{L}(c))$. In this case, $(K/(K-1))(\rho^2/(1-\rho^2))$ is approximately F distributed with $(K-1, K)$ degrees of freedom under the null hypothesis that $\beta = c$.

Adjusted rho-square The ρ^2 statistics is monotonic in the number of variables in the model, a limitation it shares with regression statistic R^2. Therefore, we also consider an adjusted version of the statistics, denoted by $\bar{\rho}^2$. It is another informal goodness-of-fit measure that is corrected for the number of parameters estimated. As shown in chapter 6, this measure is defined as $\bar{\rho}^2 = 1-(\mathcal{L}(\hat{\beta})-K)/\mathcal{L}(0)$.

The list of output related to parameters estimates depends also on the estimation software package. We report here typical outputs for coefficient k (see Table 4.19).

Value Estimated value $\widehat{\beta}_k$.

Std. Err. Estimated standard error. It is the square root of entry (k, k) of $\widehat{\Sigma}_\beta^{CR}$.

t-test Ratio between the estimated value of the parameter and the estimated standard error. It is used to test the null hypothesis that the true value of the parameter is 0. At the α significance level, the null hypothesis is rejected if the t-test is larger than $\phi^{-1}((1-\alpha)/2)$ or lower than $\phi^{-1}(\alpha/2)$, where $\phi(\cdot)$ denotes the probability density function of the standardized normal distribution. For the traditionnal value of the significance level $\alpha = 5\%$, these threshold values are 1.96 and -1.96 respectively (see Section 6.5.2 for more details).

p-value It is used as well to test the null hypothesis that the true value of the parameter is 0. If the p-value is smaller than the significance level of the test (tradionnally 5%, say), the null hypothesis is rejected. If t represents the t-test, the p-value is equal to $2(1-\Phi(t))$, where $\Phi(\cdot)$ is the cumulative density function of the univariate normal distribution.

Rob. Std. Err. Estimated robust standard error. It is the square root of entry (k, k) of $\widehat{\Sigma}_\beta^{R}$.

Rob. t-test same as t-test, using the robust estimator for the estimated standard error.

Rob. p-value same as p-value, using the robust estimator for the estimated standard error.

The standard presentation that we adopt throughout the book is presented at Table 4.19, where only the "robust" statistics are reported.

A third measure of goodness of fit known as "% right" is sometimes considered. This statistic is defined as $(100/N)\sum_n \hat{y}_n$, where $\hat{y}_n$ is 1 if the highest predicted probability corresponds to the chosen alternative, and 0 otherwise.

We do not recommend the use of this statistic for evaluating model performance. It can mask poor goodness of fit. For example, suppose that we have a sample of 100 observations, 90 of which chose alternative 1 and 10 of which chose alternative 2. Now suppose we adopt the model $P_n(1) = 0.9$. The "% right" statistic is 90% for this model, despite the fact that it completely misclassifies every observation choosing alternative 2!

In general, it is not good practice to use an indicator like $\hat{y}_n$ in any circumstance. Indeed, considering the alternative with the largest probability, that is with the largest utility, amounts to ignore the error term in the utility function, and may lead to significant errors. A typical example is the case where one alternative has a predicted probability slightly over 50%. The indicator always favors that alternative, while the other one is almost as likely to be preferred.

A different approach to calculate the "% right" is to use the calculated choice probabilities as follows:

$$\frac{100}{N}\sum_n \sum_i P_n(i)^{y_{in}}.$$

For the simple model assumed before this measure is equal to

$$90 \times 0.9 + 10 \times 0.1 = 82\%,$$

which is smaller than the 90% right obtained with the first method. However, with this method of prediction we maintain the desirable property of replicating the shares of the alternatives as follows:

$$\frac{1}{N}\sum_n P_n(i) = \frac{1}{N}\sum_n y_{in}.$$

The value of the log likelihood function should be preferred to this last measure of "% right" for evaluating a model. Indeed, it is the objective function being maximized and is more sensitive to low values of predicted probabilities for the chosen alternative. Transforms of it, such as ρ^2 and $\bar{\rho}^2$, can be used as well. The insensitivity of the "% right" statistic, in addition to its potential for completely misleading indications, argue against its use.

Still, the "% right" statistics is used in fields such as machine learning and classification, where it may have desired mathematical properties. We refer the interested reader to the literature on proper scoring rules (Savage, 1971, Schervish, 1989, Gneiting and Raftery, 2007) for more information about these properties.

4.5.2 Mode choice in the Netherlands, Revisited

To show how actual empirical results from maximum likelihood estimation may be presented, we return to the binary logit example given in section 4.3.1 (see table 4.2 for a summary of the specification).

Table 4.20 reports the estimation results using the same format as table 4.19. Note that for β_1, β_2, β_3, β_4 and β_7, we can reject the null hypothesis that the true value is zero at the 0.05 significance level, as the $\mathfrak{t}$-tests are, in absolute value, larger than 1.96 or, equivalently, because the $\mathfrak{p}$-values are lower than 0.05. For β_9, we can reject the null hypothesis at the 0.10 significance level (For a two-tailed test the critical values of the quasi-$\mathfrak{t}$ statistic are $\pm$ 1.65 and $\pm$ 1.96 for the 0.10 and 0.05 significance levels, respectively.) Note that the fact that a parameter is not significant does not automatically mean that it should be removed from the model. For instance, the null hypothesis that the true value of β_5 cannot be rejected at the 0.10 significance level. This parameter corresponds to the travel time by train. It is an important policy variable, and we have a strong belief that it is an important explanatory variable. Removing it from the model is likely to lead to a specification error, which may be worse than keeping it with a low $\mathfrak{t}$-test. This is discussed in more details in Chapter 6.

Table 4.21 reports the detailed estimation results of the binary probit model, for which similar interpretations can be made.

4.5.3 Airline itinerary choice, Revisited

We continue with the second example and return to the binary logit model given in section 4.3.2 (see table 4.7 for a summary of the specification).

Table 4.22 reports the estimation results using the same format as table 4.19. Note that for β_1, β_3, β_4, β_5, β_6, β_7 and β_8, we can reject the null hypothesis that the true value is zero at the 0.05 significance level. In addition we can reject the null hypothesis that all the parameters are jointly zero at the 0.01 level. The null hypothesis that the true value of β_9 is 0, cannot be rejected at the 0.10 significance level. This parameter corresponds to the difference of a priori preference of men compared to women, for the "one stop–same airline" alternative.

Table 4.23 reports the detailed estimation results of the binary probit model, for which similar interpretation can be made.

We return to the problem of interpreting estimation results in subsequent chapters. At this point the reader should gain familiarity with reading tables 4.18 to 4.23 and understand the mechanics of the various statistical tests we have discussed. If necessary, review the appropriate sections of chapter B before proceeding. Further formal and informal means of evaluating model specifications is developed in chapter 6.

4.6 Summary

This chapter has taken the abstract theory developed in chapter 3 and shown how it can be translated into an operational method for analyzing binary choice situations. We have shown that it is useful to partition the utility of an alternative into two parts: a systematic component (V) and a random disturbance (ε). Different assumptions about the distributions of the disturbances lead to different binary choice models.

The most common situations are where the systematic component of the utility is a linear function of the parameters. We expressed this as

$$V_{in} = \beta^T x_{in} \text{ and } V_{jn} = \beta^T x_{jn},$$

where x_{in} and x_{jn} are variables that are functions of the attributes of i and j, respectively, as well as those of decision maker n.

Two practical models were derived using different assumptions about the disturbance terms. These were as follows:

- $\varepsilon_{jn} - \varepsilon_{in}$ normal leads to the binary probit model:

$$P_n(i) = \Phi\left(\frac{V_{in} - V_{jn}}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{u=-\infty}^{(V_{in}-V_{jn})/\sigma} \exp\left[-\frac{1}{2}u^2\right] du; \quad (4.76)$$

- $\varepsilon_{jn} - \varepsilon_{in}$ logistic leads to the binary logit model:

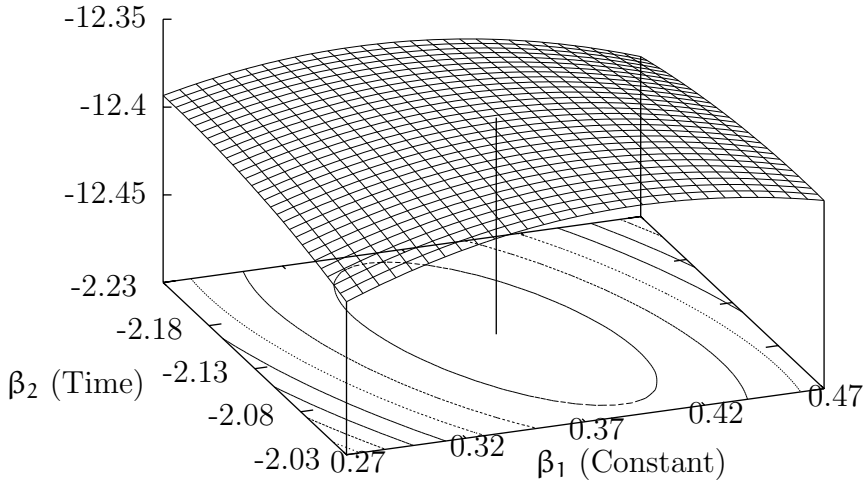
$$P_n(i) = \frac{e^{\mu V_{in}}}{e^{\mu V_{in}} + e^{\mu V_{jn}}}. \quad (4.77)$$

The maximum likelihood estimator for the unknown parameters was derived. Some of the computational aspects were presented along with some examples of hypothetical and actual binary choice models.

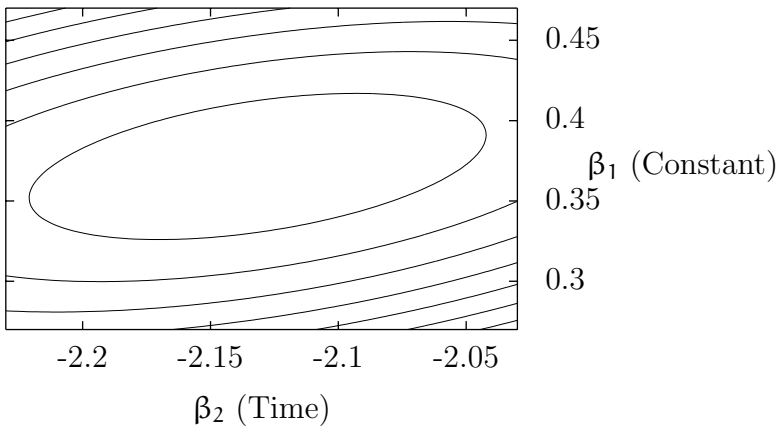
In the next chapter we extend much of our analysis to the more general case of interest with more than two alternatives. We focus primarily on

extensions of binary logit, exploiting the ease with which it can be generalized beyond binary situations.

Before that, we describe other estimation methods and other binary choice models in the chapter appendix.



(a) Log likelihood

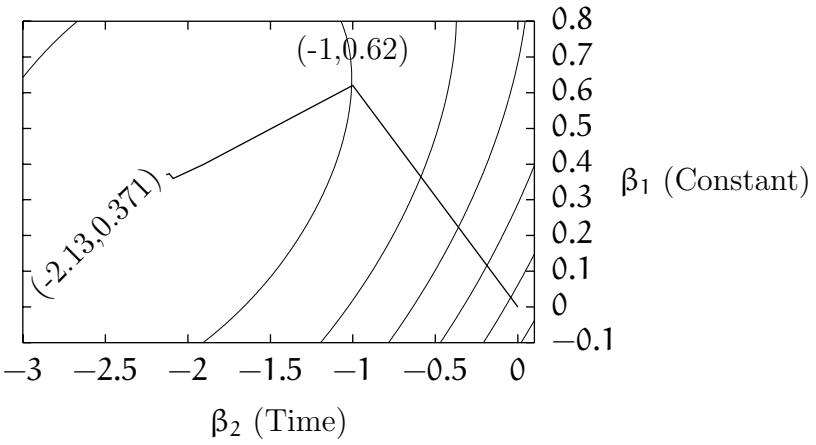


(b) Level curves

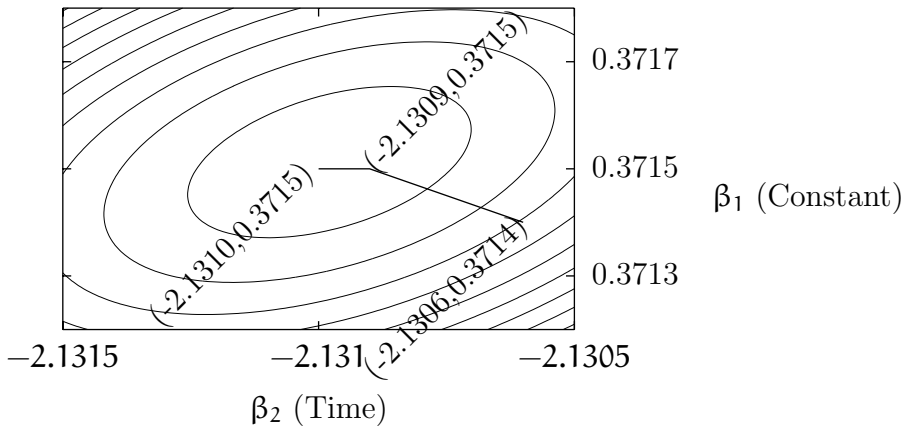
Figure 4.12: Log likelihood of small sample with binary logit model

Iter.	$\mathcal{L}(\beta)$	β_1	β_2	$\partial\mathcal{L}/\partial\beta_1$	$\partial\mathcal{L}/\partial\beta_2$	$\partial^2\mathcal{L}/\partial\beta_1^2$	$\partial^2\mathcal{L}/\partial\beta_2^2$	$\partial^2\mathcal{L}/\partial\beta_1\partial\beta_2$
1	-17.328680	0.000000	0.000000	5.5	-4.3815	-6.25	-2.29177	2.20025
2	-12.569793	0.312610	-1.611710	0.800482	-0.689803	-4.67436	-1.38266	1.22222
3	-12.379423	0.365677	-2.063698	0.0846084	-0.076959	-4.10703	-1.08518	0.923655
4	-12.376605	0.371429	-2.129720	0.00145062	-0.00139071	-4.03111	-1.04647	0.886535
5	-12.376605	0.371512	-2.130978	4.79376e-07	-4.78298e-07	-4.02971	-1.04576	0.885866
6	-12.376605	0.371513	-2.130979	5.59552e-14	-5.701e-14	-4.02971	-1.04576	0.885865
7	-12.376605	0.371513	-2.130979	-3.33067e-16	-2.22045e-16	-4.02971	-1.04576	0.885865

Table 4.17: Example of the maximum likelihood estimation iterations



(a) Zoomed in



(b) Zoomed out

Figure 4.13: Maximum likelihood iterations

Number of estimated parameters	:	2
Number of observations	:	25
$\mathcal{L}(0)$	:	-17.329
$\mathcal{L}(c)$	:	-14.824
$\mathcal{L}(\widehat{\beta})$	:	-12.377
$-2(\mathcal{L}(0) - \mathcal{L}(\widehat{\beta}))$	:	9.904
ρ^2	:	0.286
$\bar{\rho}^2$	:	0.170

Table 4.18: General indicators for the simple example

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	Robust t-stat	Robust p-value
1	Auto constant	0.372	0.492	0.75	0.45
2	Travel time	-2.13	1.22	-1.75	0.08

Summary statistics

Number of observations = 25

$\mathcal{L}(0)$	=	-17.329
$\mathcal{L}(c)$	=	-14.824
$\mathcal{L}(\widehat{\beta})$	=	-12.377
$-2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})]$	=	9.904
ρ^2	=	0.286
$\bar{\rho}^2$	=	0.170

Table 4.19: Standard presentation of the results for the simple example

Param. number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	Car dummy	3.04	1.09	2.78	0.01
2	Cost	-0.0527	0.0127	-4.17	0.00
3	Travel time by car (work)	-2.66	0.578	-4.60	0.00
4	Travel time by car (not work)	-2.22	0.499	-4.46	0.00
5	Travel time by train	-0.576	0.460	-1.25	0.21
6	First class dummy	0.961	0.768	1.25	0.21
7	Male dummy	-0.850	0.358	-2.37	0.02
8	Main earner dummy	0.383	0.353	1.09	0.28
9	Fixed arrival time dummy	-0.624	0.370	-1.69	0.09

Summary statistics

Number of observations = 228

$$\begin{aligned} \mathcal{L}(0) &= -158.038 \\ \mathcal{L}(c) &= -148.347 \\ \mathcal{L}(\hat{\beta}) &= -108.836 \\ -2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 98.404 \\ \rho^2 &= 0.311 \\ \bar{\rho}^2 &= 0.254 \end{aligned}$$

Table 4.20: Binary logit model of mode choice in the Netherlands: detailed estimation results in standard form

Param. number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	Car dummy	1.77	0.632	2.81	0.00
2	Cost	-0.0296	0.00706	-4.20	0.00
3	Travel time by car (work)	-1.51	0.347	-4.35	0.00
4	Travel time by car (not work)	-1.26	0.312	-4.03	0.00
5	Travel time by train	-0.308	0.258	-1.20	0.23
6	First class dummy	0.545	0.414	1.32	0.19
7	Male dummy	-0.471	0.206	-2.29	0.02
8	Main earner dummy	0.213	0.208	1.02	0.31
9	Fixed arrival time dummy	-0.355	0.211	-1.68	0.09

Summary statistics

Number of observations = 228

$$\mathcal{L}(0) = -158.038$$
$$\mathcal{L}(c) = -148.347$$
$$\mathcal{L}(\hat{\beta}) = -109.544$$
$$-2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] = 96.987$$
$$\rho^2 = 0.307$$
$$\bar{\rho}^2 = 0.250$$

Table 4.21: Binary probit model of mode choice in the Netherlands: detailed estimation results

Param. number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-1.10	0.294	-3.75	0.00
2	Round trip fare ($100)	-2.06	0.138	-14.91	0.00
3	Elapsed time (hours)	-0.214	0.115	-1.86	0.06
4	Leg room (inches), if male	0.133	0.0319	4.17	0.00
5	Leg room (inches), if female	0.122	0.0304	4.02	0.00
6	Being early (hours)	-0.126	0.0245	-5.13	0.00
7	Being late (hours)	-0.0922	0.0221	-4.17	0.00
8	More than 2 air trips per year	-0.377	0.141	-2.67	0.01
9	Male dummy	0.182	0.129	1.40	0.16

Summary statistics

Number of observations = 2143

$$\mathcal{L}(0) = -1485.414$$
$$\mathcal{L}(c) = -1094.709$$
$$\mathcal{L}(\hat{\beta}) = -818.421$$
$$-2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] = 1333.987$$
$$\rho^2 = 0.449$$
$$\bar{\rho}^2 = 0.443$$

Table 4.22: Binary logit model for the choice of airline itinerary: detailed estimation results

Param. number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.685	0.165	-4.16	0.00
2	Round trip fare ($100)	-1.13	0.0728	-15.55	0.00
3	Elapsed time (hours)	-0.109	0.0645	-1.69	0.09
4	Leg room (inches), if male	0.0752	0.0178	4.22	0.00
5	Leg room (inches), if female	0.0607	0.0166	3.65	0.00
6	Being early (hours)	-0.0671	0.0141	-4.76	0.00
7	Being late (hours)	-0.0528	0.0120	-4.39	0.00
8	More than 2 air trips per year	-0.195	0.0798	-2.45	0.01
9	Male dummy	0.111	0.0716	1.55	0.12

Summary statistics

Number of observations = 2143

$$\begin{aligned} \mathcal{L}(0) &= -1485.414 \\ \mathcal{L}(c) &= -1094.709 \\ \mathcal{L}(\hat{\beta}) &= -822.073 \\ -2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 1326.682 \\ \rho^2 &= 0.447 \\ \bar{\rho}^2 &= 0.441 \end{aligned}$$

Table 4.23: Binary probit model for the choice of airline itinerary: detailed estimation results

Chapter Appendix

4.A Properties of the extreme value distribution

The extreme value distribution introduced in Section 4.2.2 has the following properties:

1. The mode is η.
2. The mean is $\eta + \frac{\gamma}{\mu}$, where

$$\gamma = -\int_0^{+\infty} e^{-x} \ln x dx = \lim_{n\to\infty} \left(\sum_{k=1}^{n} \frac{1}{k} - \ln n \right) \approx 0.5772 \tag{4.78}$$

   is Euler's constant.
3. The variance is $\frac{\pi^2}{6\mu^2}$.
4. If $\varepsilon \sim \text{EV}(\eta,\mu)$, then

$$a\varepsilon + b \sim \text{EV}(a\eta + b, \frac{\mu}{a}),$$

   where $a, b \in \mathbb{R}$, $a > 0$.
5. If $\varepsilon_a \sim \text{EV}(\eta_a, \mu)$ and $\varepsilon_b \sim \text{EV}(\eta_b, \mu)$ are independent with the same scale parameter μ, then

$$\varepsilon = \varepsilon_a - \varepsilon_b \sim \text{Logistic}(\eta_a - \eta_b, \mu),$$

   namely

$$f_\varepsilon(\xi) = \frac{\mu e^{-\mu(\xi-\eta_a+\eta_b)}}{(1+e^{-\mu(\xi-\eta_a+\eta_b)})^2}, \tag{4.79}$$

$$F_\varepsilon(\xi) = \frac{1}{1+e^{-\mu(\xi-\eta_a+\eta_b)}}, \ \mu > 0, -\infty < \xi < \infty. \tag{4.80}$$

$$\tag{4.81}$$

6. If $\varepsilon_i \sim \text{EV}(\eta_i, \mu)$, for $i = 1, \ldots, J$, and ε_i are independent with the same scale parameter μ, then

$$\varepsilon = \max_{i=1,\ldots,J} \varepsilon_i \sim \text{EV}(\eta, \mu) \tag{4.82}$$

where

$$\eta = \frac{1}{\mu} \ln \sum_{i=1}^{J} e^{\mu \eta_i}. \tag{4.83}$$

It is important to note that this property holds only if all ε_i have the same scale parameter μ. As ε follows an extreme value distribution, its expected value is

$$E[\varepsilon] = \eta + \frac{\gamma}{\mu}.$$

Equivalently,

$$\eta = E[\varepsilon] - \frac{\gamma}{\mu}.$$

Therefore, (4.83) provides the expected value of the maximum, up to a constant. This interpretation is useful in the derivation of random utility models such as the nested logit model, in Chapter 7.

4.B Least Squares and Berkson's Method

Least squares procedures can also be used to estimate the unknown parameters of probit, logit, or other choice models. In general, the estimates are the parameters $\hat{\beta}$ that solve

$$\min_{\hat{\beta}} Q = \sum_{n=1}^{N} (y_{in} - P_n(i))^2. \tag{4.84}$$

The necessary conditions for a solution to equation (4.84) are

$$\frac{\partial Q}{\partial \hat{\beta}} = -2 \sum_{n=1}^{N} (y_{in} - P_n(i)) \frac{\partial P_n(i)}{\partial \hat{\beta}} = 0, \ \forall k, \tag{4.85}$$

or more simply

$$\sum_{n=1}^{N} (y_{in} - P_n(i)) \frac{\partial P_n(i)}{\partial \hat{\beta}} = 0, \ k = 1, \ldots, K. \tag{4.86}$$

This approach yields consistent estimates (see White, 1980). However, it can be computationally difficult and has no theoretical advantage over maximum likelihood. It has therefore not been used in actual practice. An alternative least squares procedure developed by Berkson (1953) does offer some significant advantages over maximum likelihood in certain instances. Berkson's

procedure is based on the observation that linear-in-parameters binary choice models can easily be transformed to put them in a form amenable to standard regression analysis. Any such model can be written as

$$P_n(i) = F(\beta^T x_n), \tag{4.87}$$

where x_n is defined again as $x_{in} - x_{jn}$. Thus as long as F, a proper cumulative distribution function, is strictly monotonically increasing, we can write

$$F^{-1}(P_n(i)) = \beta^T x_n, \tag{4.88}$$

in which the right-hand side is a simple linear function. To be more specific, consider binary logit and probit. Here we have the following:

$$\ln\left(\frac{P_n(i)}{P_n(j)}\right) = \beta^T x_n \text{ for the binary logit model} \tag{4.89}$$

and

$$\Phi^{-1}(P_n(i)) = \beta^T x_n \text{ for the binary probit.} \tag{4.90}$$

The natural problem with applying this result is that $P_n(i)$ is of course not observed; all we observe is y_{in}, a (0,1) indicator of whether the person chose alternative i or j. Berkson's procedure divides the sample into homogeneous subgroups and uses the share of each group choosing each alternative as estimates of the choice probabilities. These shares, appropriately transformed, become the left-hand side variables in the regression.

More formally, Berkson's procedure assumes that the sample of N decision makers can be divided into G ($< N$) subgroups of size $N_1, N_2, N_3, \ldots, N_G$, where

$$\sum_{g=1}^{G} N_g = N.$$

Each subgroup is assumed to be homogeneous in terms of $x_n = x_{in} - x_{jn}$. We denote x_g, $g = 1, \ldots, G$ as the value of x_n for the gth group. We also define R_{ig} and R_{jg} as the share of the gth group choosing alternatives i and j, respectively. We use R_{ig} and R_{jg} as estimates of $P_n(i)$ and $P_n(j)$, respectively, for group g. For logit this implies that we estimate the parameters via the regression

$$\ln\left(\frac{R_{ig}}{R_{jg}}\right) = \beta^T x_n + \xi_g, \tag{4.91}$$

where ξ_g is a disturbance term in the regression attributable to the fact that R_{ig} is only an estimate of $P_n(i)$.

We illustrate the procedure using the example introduced in Section 1.2, analyzing the market penetration of smartphones. We consider the data in Table 4.14 (same data as in Table 1.1), and estimate a model with the following specification:

$$\begin{array}{rcl} V_{in} & = & \beta_\ell x_{low} + \beta_m x_{medium} + \beta_0, \\ V_{jn} & = & 0, \end{array} \tag{4.92}$$

where i is the alternative corresponding to the choice of using the smartphone. We define three subgroups, corresponding to each level of education, so that we can compute the quantities in (4.91):

Group g	R_{ig}	$R_{jg} = 1 - R_{ig}$	$\ln(R_{ig}/R_{jg})$
Low	0.25	0.75	-1.1
Medium	0.5	0.5	0
High	0.25	0.75	-1.1

We obtain the following set of equations:

$$\begin{array}{rcll} -1.1 & = & \beta_\ell + \beta_0 & +\xi_\ell, \\ 0 & = & \beta_m + \beta_0 & +\xi_m, \\ -1.1 & = & \beta_0 & +\xi_h. \end{array} \tag{4.93}$$

With this simple example, the estimates of the parameters are trivially derived from these equations: $\beta_0 = -1.1$, $\beta_\ell = 0$, $\beta_m = 1.1$.

	Education			
VoIP	Low ($k = 1$)	Medium ($k = 2$)	High ($k = 3$)	
Yes ($i = 1$)	10	100	90	200
No ($i = 2$)	140	200	60	400
	150	300	150	600

Figure 4.14: Survey responses from the simple example in Section 1.2

Cox (1970) notes that the disturbance term ξ_g is heteroscedastic and suggests a two-stage estimator to gain efficiency. The heteroscedasticity arises because $R_{ig}N_g$ is binomially distributed (Domencich and McFadden, 1975, p. 109). For probit we use

$$\Phi^{-1}(R_{ig}) = \beta^T x_g + \xi_g,$$

where ξ_g is again a disturbance due to the approximations of $P_n(i)/P_n(j)$ by R_{ig}/R_{jg}. In both cases we have G grouped observations.

The main advantage of Berkson's method is that it allows use of standard regression packages and reduces the number of data points from the original N to a smaller number G. It can be shown that this procedure yields consistent estimates of β, where consistency is defined here as $N_g \to \infty$ for all g. Despite its obvious appeal Berkson's method is not widely used in practice for a number of significant reasons. First, it requires division of the sample into homogeneous groups. As Domencich and McFadden (1975) note, even if each of the K variables in the model can take only 2 values, this still implies 2^K homogeneous groups. For typical values of K this implies a very large value of G. (For $K = 10$ there would be 1,024 different groups.) Second, data are usually quite expensive to gather, so the total sample may be small, and the number of observations per group is small if G is large. As a consequence the variance of R_{ig} and R_{jg} is large.

A third, related problem is that when N_g is small for many groups, it is not at all unusual in some groups for one of the observed shares choosing i or j to be zero. This makes $F^{-1}(R_{ig})$ undefined for logit or probit models, requiring that these cells not be used or that more aggregated groups be formed.

Fourth, when many of the attributes are continuous, forming groups requires an arbitrary categorization along some attributes and the use of the within-group mean as "representative" attributes for the group. This tends to introduce an error-in-variables problem into the model. Some simple simulation results reported by Domencich and McFadden (1975) indicate that this may not be of great practical importance.

Despite these difficulties Berkson's method should be considered when appropriate and is generally most useful under the following conditions:

1. The available sample is extremely large, which is more and more the case with data collected on the internet, or using technologies such as smartphones or GPS.

2. The data are only available in aggregated form, as is often the case with the U.S. Census data (aggregation in this case is to preserve the anonymity of respondents).

3. The model structure uses only a small number of categorized variables so that the number of cells is reasonably small.

4. Each respondent to a survey is observed making a large number of repeated decisions so that each individual's choices form a natural basis for grouping.

4.C Other Estimation Methods

Although both maximum likelihood and least squares estimations have been used in almost all actual applications of binary choice models, they by no means exhaust the full spectrum of possible estimation methods. There are other approaches for estimating choice models that differ fundamentally from these two procedures.

One interesting class of approaches includes methods that are similar to maximum likelihood but optimize a somewhat different function. An example of such a procedure is given in chapter 11, where we show that even though the likelihood function is extremely complicated, there exists another simpler function that yields consistent (though not necessarily as efficient) parameter estimates.

A second class of procedures are what might be termed *nonparametric* estimation methods (Tsybakov, 2008). By this we mean that they do not make a specific distributional assumption on the ε's but rather hypothesize that the distribution of the disturbances belongs to a class that has some very general properties.

One relatively intuitive procedure was proposed by Manski (1975). This approach, termed *maximum score estimation*, selects parameters so that the fraction of the sample who chose the alternative with greatest systematic component of the utility is maximized. (This is equivalent to maximizing the % right statistic discussed in section 4.5.) Manski shows that the maximum score estimates are not in general unique because more than one value of the parameter estimate $\hat{\beta}$ may yield the same % right value. As the sample gets larger, the problem of nonuniqueness typically becomes less significant.

Maximum score estimates are consistent under some very general conditions. The most relevant of these conditions is that for each individual in the population the choice probability must be monotonically increasing with its systematic component of utility — that is, the greater V_{in} is, the greater $P_n(i)$. We do not even require that each individual's disturbances come from the same distribution as long as all the disturbances have this property.

Another procedure proposed by Cosslett (1983) is slightly more restrictive than Manski's approach in that it requires all individual's disturbances to come from the same unknown distribution. Cosslett then derives estimates of the choice probabilities as a function of $V_{in} - V_{jn}$ along with $\hat{\beta}$. Though quite general, the procedure is somewhat difficult computationally. Actually Cosslett's method is a maximum likelihood estimator where both the choice model and the coefficients β are unknown. We include it in this section because it is so different from the usual maximum likelihood approaches.

4.D Ordinal binary choice model

An ordinal binary choice model is derived when ordinal responses are available, where the respondent not only reports the preference, but also the strength of the preference. For instance, if alternatives i and j are available, the respondent can report one of the following.

- definitely choose j;
- probably choose j;
- indifferent;
- probably choose i;
- definitely choose i.

As for the binary choice model, the selected category is explained by the difference $U_{in} - U_{jn}$ between the utilities of the two alternatives, as depicted in Figure 4.15.

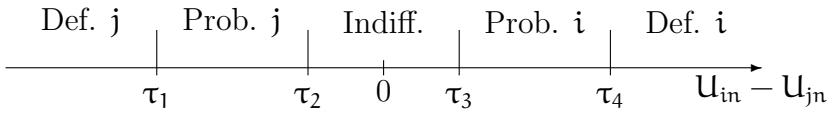


Figure 4.15: Categories for the ordinal binary choice model

Formally, we consider $Q \geq 2$ categories, ordered such that category q corresponds to a stronger preference towards alternative i compared to category $q-1$, for $q = 1, \ldots, Q$. We define $Q+1$ parameters τ_q, $q = 0, \ldots, Q$, such that $\tau_0 = -\infty$, $\tau_Q = +\infty$, and $\tau_{q-1} \leq \tau_q$, $q = 1, \ldots, Q$. A category q is associated with the interval $[\tau_{q-1}, \tau_q]$. Using (4.3), the probability for category q to be selected by the respondent is

$$\begin{aligned} P_n(q) &= \Pr(\tau_{q-1} \leq U_{in} - U_{jn} \leq \tau_q) \\ &= \Pr(\tau_{q-1} \leq (V_{in} - V_{jn}) - (\varepsilon_{jn} - \varepsilon_{in}) \leq \tau_q) \\ &= \Pr(V_{in} - V_{jn} - \tau_q \leq \varepsilon_n \leq V_{in} - V_{jn} - \tau_{q-1}) \\ &= F_{\varepsilon_n}(V_{in} - V_{jn} - \tau_{q-1}) - F_{\varepsilon_n}(V_{in} - V_{jn} - \tau_q) \end{aligned} \tag{4.94}$$

where $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$, and F_{ε_n} is the CDF of ε_n. This is illustrated by Figure 4.16, where the shaded area under the density function represents

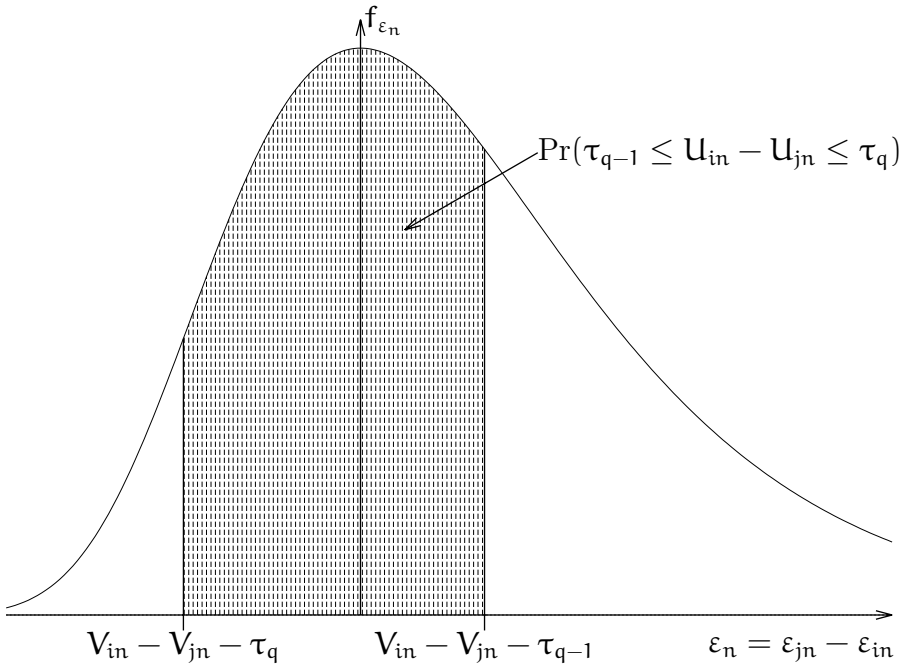


Figure 4.16: Ordinal choice: probability for category q

the probability for category q to be selected. By definition of the CDF, the probability for the extreme categories simplify to

$$\begin{array}{rcrcl} P_n(1) & = & F_{\varepsilon_n}(V_{in} - V_{jn} + \infty) & - & F_{\varepsilon_n}(V_{in} - V_{jn} - \tau_1) \\ & = & 1 & - & F_{\varepsilon_n}(V_{in} - V_{jn} - \tau_1), \\ P_n(Q) & = & F_{\varepsilon_n}(V_{in} - V_{jn} - \tau_{Q-1}) & - & F_{\varepsilon_n}(V_{in} - V_{jn} - \infty) \\ & = & F_{\varepsilon_n}(V_{in} - V_{jn} - \tau_{Q-1}). & & \end{array} \tag{4.95}$$

In particular, if ε_n is logistically distributed, we obtain the *ordinal logit* model. Similary, if ε_n is normally distributed, we obtain the *ordinal probit* model. We immediately note that binary choice models are specific instances of ordinal binary choice models, with two categories ($Q = 2$), and $\tau_1 = 0$.

For concrete applications, we refer the reader to McKelvey and Zavoina (1975), who analyze the congressional voting on the 1965 medicare bill in the USA, Chu and Anderson (1992), who estimate an ordinal logit model in a financial context, and de Palma and Picard (2005) who estimate an ordinal probit model in a transportation analysis context.

Chapter 5

Choice with multiple alternatives

Contents

We now turn to the development of models for the more general case where the choice set, $\mathcal{C}_n$, can consist of more than two alternatives. (Recall that the choice set is subscripted by the decision-maker's index, to indicate that choice sets may vary across individuals.) In such instances the derivation of useful choice models and appropriate estimation methods becomes considerably more complex than for binary choice analysis. In particular, it is not sufficient simply to specify the univariate distribution of the differences in disturbances, $\varepsilon_{jn} - \varepsilon_{in}$. Instead, we have to characterize the joint distribution of all the disturbances.

The choice among multiple alternatives is called a *polychotomous choice* (e.g. Gurland et al., 1960.) In the choice modeling literature, it is known as *multinomial choice* (e.g. Bock, 1968 and Theil, 1969.)

Our presentation proceeds as follows. In section 5.1 we analyze the general problem of choice with multiple alternatives for random utility models. Then, due to the ease with which it extends to multiple alternatives choice analysis, we focus most of our attention on the logit model. The derivation, properties, estimation, and an example of this model is covered in the next sections. Section 5.9 then briefly describes some models that differ from logit by allowing for less restrictive assumptions about the distribution of disturbance terms. A summary of the chapter is given in section 5.10.

5.1 Derivation from the Random Utility Model

We begin by assuming that for the problem being studied, the analyst can define some set $\mathcal{C}$ that includes all potential choices for some population. We call $\mathcal{C}$ the universal choice set, and define J to be the number of elements in it. Each member of the population has some subset of $\mathcal{C}$ as his or her choice set. For example, in a mode choice model, $\mathcal{C}$ may consist of eight elements:

1. driving alone,

2. sharing a ride,
3. taxi,
4. motorcycle,
5. bicycle,
6. walking,
7. bus,
8. rail.

However, for any particular traveler the actual choice set, $\mathcal{C}_n$, may be considerably smaller. A worker may live or work beyond the reasonable limits of transit service, thus eliminating options 7 and 8. Some workers may not own an automobile, making driving alone infeasible. For others the work trip may be too long for walking to be viewed as a viable option.

Obviously what constitutes a feasible alternative for any particular individual may be difficult for the analyst to determine. How far is "too far to walk"? Is bicycling really feasible in certain climates? Do some people know about the existence of transit services that they might use? These types of questions require the analyst to make informed judgments about what Manski (1977) terms the choice set generation process. At this stage we assume that each individual's choice set can be specified by the analyst using some reasonable, deterministic rules.

We must realize, however, that this imputation of the choice set by the analyst is in effect a potentially crude model of a complex interaction between an individual decision maker and his or her environment. It is possible to formulate choice models that explicitly account for choice set generation, albeit at a great cost in terms of complexity. This is discussed in Section 5.3.2.

Given that each individual has a feasible choice set denoted by $\mathcal{C}_n$, we define $J_n \leq J$ to be the number of feasible choices. Following the development of random utility theory of chapter 3, each individual considers a vector of J_n utilities,

$$U_n = \begin{pmatrix} U_{1n} \\ \vdots \\ U_{J_n n} \end{pmatrix} = V_n + \varepsilon_n \tag{5.1}$$

and the probability that any element i in $\mathcal{C}_n$ is chosen by decision maker n is given by

$$P_n(i) = \Pr(U_{in} \geq U_{jn}, \forall j \in \mathcal{C}_n), \tag{5.2}$$

or, equivalently,

$$P_n(i) = \Pr(U_{jn} - U_{in} \leq 0, \forall j \in \mathcal{C}_n). \tag{5.3}$$

In order to write (5.3) in vector notations, we use the $(J_n - 1) \times J_n$ matrix Δ_i that transforms the utilities to differences with respect to alternative i, defined in Section 3.B. Equation (5.3) can now be written as

$$P_n(i) = \Pr(\Delta_i U_n \leq 0) = \Pr(\Delta_i V_n + \Delta_i \varepsilon_n \leq 0) = \Pr(\Delta_i \varepsilon_n \leq -\Delta_i V_n), \tag{5.4}$$

where the inequality between two n-dimensional vectors applies to all pairs of elements with the same index. In other words, the choice probability of alternative i in (5.4) is equal to the joint probability of all utility differences, $U_{jn} - U_{in}$, for all $j \in \mathcal{C}_n$ being less or equal to zero, as expressed in (5.3).

Any particular choice model can be derived using equation (5.4) given specific assumptions on the joint distribution of the disturbances. Let $f(\varepsilon_n) = f(\varepsilon_{1n}, \varepsilon_{2n}, \ldots, \varepsilon_{J_n n})$ denote the joint probability density function (pdf) of the disturbance terms. To evaluate the probability in Equation (5.4), we need to derive the joint cumulative distribution function (CDF) of $\Delta_i \varepsilon_n$ from the joint pdf of ε_n.

We illustrate this derivation for the probit case, where a normal distribution is assumed. We consider a model with three alternatives, so that

$$U = \begin{pmatrix} U_1 \\ U_2 \\ U_3 \end{pmatrix} = \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} = V + \varepsilon,$$

where $U \sim N(V, \Sigma)$ or, equivalently, $\varepsilon \sim N(0, \Sigma)$, and $\Sigma \in \mathbb{R}^{3\times 3}$ is the variance-covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22}^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33}^2 \end{pmatrix},$$

and

$$f(\varepsilon) = (2\pi)^{-3/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} \varepsilon^T \Sigma^{-1} \varepsilon\right), \tag{5.5}$$

where $|\Sigma|$ denotes the determinant of Σ. In order to compute the probability of choosing alternative 2, say, we need to consider (see Section 3.B)

$$\Delta_2 U = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} U = \begin{pmatrix} U_1 - U_2 \\ U_3 - U_2 \end{pmatrix} = \begin{pmatrix} V_1 - V_2 \\ V_3 - V_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 - \varepsilon_2 \\ \varepsilon_3 - \varepsilon_2 \end{pmatrix}.$$

Pre-multiplication of the vector U by the matrix Δ_2 yields a linear transformation of U which by assumption is normally distributed. A linear transformation of a normally distributed random vector is also normally distributed, as follows:

$$\Delta_2 U \sim N(\Delta_2 V, \Delta_2 \Sigma \Delta_2^T),$$

and

$$\Delta_2 \Sigma \Delta_2^T = \begin{pmatrix} \sigma_{11}^2 + \sigma_{22}^2 - 2\sigma_{12} & \sigma_{22}^2 + \sigma_{13} - \sigma_{12} - \sigma_{23} \\ \sigma_{22}^2 + \sigma_{13} - \sigma_{12} - \sigma_{23} & \sigma_{33}^2 + \sigma_{22}^2 - 2\sigma_{23} \end{pmatrix}.$$

The density function of the random vector $\Delta_i U$ is given by

$$f(\Delta_i \varepsilon) = \frac{1}{2\pi} \left|\Delta_i \Sigma \Delta_i^T\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\Delta_i \varepsilon - \Delta_i V)^T (\Delta_i \Sigma \Delta_i^T)^{-1} (\Delta_i \varepsilon - \Delta_i V)}, \tag{5.6}$$

where $|\Delta_i \Sigma \Delta_i^T|$ denotes the determinant of the matrix $\Delta_i \Sigma \Delta_i^T$. Therefore, from (5.4), we have

$$\begin{aligned} P_n(2) &= \Pr(\Delta_2 \varepsilon \leq -\Delta_2 V) \\ &= \int_{-\infty}^{V_2 - V_3} \int_{-\infty}^{V_2 - V_1} \frac{1}{2\pi} \left|\Delta_2 \Sigma \Delta_2^T\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\nu - \Delta_2 V)^T (\Delta_2 \Sigma \Delta_2^T)^{-1} (\nu - \Delta_2 V)} d\nu_1 d\nu_2, \end{aligned} \tag{5.7}$$

where

$$\nu = \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} = \begin{pmatrix} \varepsilon_1 - \varepsilon_2 \\ \varepsilon_3 - \varepsilon_2 \end{pmatrix}.$$

Unfortunately, the double integral in (5.7) does not have a closed form. In practice, the double integral can be reduced to a single integral, and computed using Gauss-Legendre numerical integration (Donnelly, 1963, Drezner and Wesolowsky, 1989, Genz, 2004.) For larger choice sets, numerical integration imposes prohibitive computational burden. Monte Carlo integration is then the only viable option.

Equivalently, (5.2) can be written as

$$\begin{aligned} P_n(i) &= \Pr(U_{in} \geq U_{jn}, \forall j \in \mathcal{C}_n, j \neq i) \\ &= \Pr(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}, \forall j \in \mathcal{C}_n, j \neq i) \\ &= \Pr(\varepsilon_{jn} \leq V_{in} - V_{jn} + \varepsilon_{in}, \forall j \in \mathcal{C}_n, j \neq i). \end{aligned} \tag{5.8}$$

Without loss of generality consider alternative i to be the first alternative in $\mathcal{C}_n$. Then

$$\begin{aligned} P_n(1) = & \int_{\varepsilon_{1n}=-\infty}^{+\infty} \int_{\varepsilon_{2n}=-\infty}^{V_{1n} - V_{2n} + \varepsilon_{1n}} \cdots \\ & \int_{\varepsilon_{J_n n}=-\infty}^{V_{1n} - V_{J_n n} + \varepsilon_{1n}} f(\varepsilon_{1n}, \varepsilon_{2n}, \ldots, \varepsilon_{J_n n}) d\varepsilon_{J_n n} \cdots d\varepsilon_{1n}. \end{aligned} \tag{5.9}$$

Obviously, if we are interested in cases other than $i = 1$, we can simply reorder the choices in $\mathcal{C}_n$ appropriately to use equation (5.9). Note that the integration is carried over a subspace of the disturbances where

$$U_{in} = \max(U_{1n}, U_{2n}, \ldots, U_{J_n n}).$$

Although equation (5.9) is the most direct way of expressing the choice probability in the abstract, it is often not the most convenient way to derive $P_n(i)$ for a particular situation. Two other forms for the choice probability can also be used. In the first we denote $F_i(\varepsilon_{1n}, \varepsilon_{2n}, \ldots, \varepsilon_{J_n n})$ as the partial derivative of F (the cumulative distribution function of the disturbances) with respect to ε_{in}. Using this notation (and again assuming for convenience that the alternatives are ordered so that $i = 1$), we can express $P_n(1)$ (see the derivation in Section 3.A):

$$\begin{aligned} P_n(1) = \int_{\varepsilon_{1n}=-\infty}^{+\infty} & F_1(\varepsilon_{1n}, V_{1n} - V_{2n} + \varepsilon_{1n}, V_{1n} - V_{3n} + \varepsilon_{1n}, \\ & \ldots, V_{1n} - V_{J_n n} + \varepsilon_{1n}) d\varepsilon_{1n}. \end{aligned} \tag{5.10}$$

In words, equation (5.10) can be interpreted as follows. Set the disturbance ε_{1n} at some given value. The integrand is then the probability that ε_{1n} equals that value and all the other disturbances satisfy the condition $V_{1n} + \varepsilon_{1n} \geq V_{jn} + \varepsilon_{jn}, \forall j \in \mathcal{C}_n$. By integrating over all possible values of ε_{1n}, we obtain the total probability that alternative 1 is chosen.

The third, and perhaps the most insightful way to express $P_n(i)$, is to reduce the choice problem with multiple alternatives to a problem with only two. To do this, we note that the condition

$$U_{in} \geq U_{jn}, \ \forall j \in \mathcal{C}_n, j \neq i, \tag{5.11}$$

is in fact equivalent to

$$U_{in} \geq \max_{j \in \mathcal{C}_n, j \neq i} U_{jn}. \tag{5.12}$$

Thus we can create what is in effect a "composite" alternative out of all the elements in $\mathcal{C}_n$ other than i, and we use the utility of the best alternative in the composite to represent the entire composite. If U_{in} exceeds the utility of the composite alternative, then i is chosen; otherwise, it is not. Thus

$$P_n(i) = \Pr\left(V_{in} + \varepsilon_{in} \geq \max_{j \in \mathcal{C}_n, j \neq i} (V_{jn} + \varepsilon_{jn})\right). \tag{5.13}$$

Of course, since U_{jn} is a random variable, $\max_{j \in \mathcal{C}_n, j \neq i} U_{jn}$ is also random. Thus, to utilize equation (5.13), we have to derive the distribution of the

utility of the composite alternative from the underlying distribution of the disturbances, F. In most instances this is a formidable task. However, in the logit case it is feasible. This facet of the logit model leads to some of its most valuable properties and has made it the most widely used method for discrete choice analysis.

5.2 The Logit Model

The logit model is also known as the *multinomial logit model* or the *conditional logit model*. It is expressed as

$$P_n(i) = \frac{e^{\mu V_{in}}}{\sum_{j \in C_n} e^{\mu V_{jn}}}, \tag{5.14}$$

where μ is the scale parameter. This model reduces to binary logit when $J_n = 2$. It defines a proper probability mass function since

$$0 \leq P_n(i) \leq 1, \text{ for all } i \in C_n, \tag{5.15}$$

and

$$\sum_{i \in C_n} P_n(i) = 1. \tag{5.16}$$

If we assume that $U_{in} = V_{in} + \varepsilon_{in}$, for all $i \in C_n$, and that all the disturbances ε_{in} are (1) independently , (2) identically , and (3) extreme value (EV) distributed with a location parameter η, and a scale parameter $\mu > 0$, then the random utility model yields the logit model in (5.14)

As in the case of binary logit the assumption of a constant η for all alternatives, or $\eta = 0$, is not in any sense restrictive as long as the systematic utilities include alternative specific constants as appropriate. Note that the choice of the EV distribution is motivated by two aspects. First, as discussed for the binary logit model, the distribution captures the maximum of many i.i.d. random variables. We therefore assume that the disturbance term of the utility function represents the largest among many unknown factors. This assumption is consistent with the concept of utility maximization. Second, it is analytically convenient, as a simple closed form probability is obtained.

However, the assumption that the disturbances are independent and identically distributed (i.i.d.) represents an important restriction. First, it constrains all the disturbances to have the same scale parameter μ. Although the choice of μ is arbitrary since it simply sets the scale of the utilities (see section 4.2), the fact that each disturbance has the *same* value of μ across

Robert Duncan Luce is Distinguished Research Professor of Cognitive Sciences and Economics at the University of California, Irvine. As a child, Duncan Luce liked painting and was interested in airplanes. His parents discouraged him against an artistic career and an astigmatism kept him from becoming a pilot, but he obtained a B.S. in Aeronautical Engineering from the Massachusetts Institute of Technology in 1945. He stayed at MIT, where he obtained a PhD in mathematics in 1950. One of his most famous contributions is the *choice axiom*, which is equivalent to the "independence from irrelevant alternatives" (IIA), discussed in Section 5.3.1. Daniel McFadden cited Luce's work upon receiving his 2000 Nobel Prize: "In a fully just world there would be a Nobel Prize for psychology. And Duncan Luce would have long since received it."

Figure 5.1: R. Duncan Luce

individuals and alternatives implies that the variances of the random components of the utilities are equal across alternatives $\mathtt{i}$ and across individuals $\mathtt{n}$. As such, the model is said to have equal variances across alternatives and to be *homoscedastic* across individuals. Furthermore, as we discuss extensively in section 5.3, the assumption of independence of the disturbances may, in some situations, be difficult to defend.

Derivation of Logit

The logit model can be derived in a great number of ways. Its original formulation is due to Luce (1959), a mathematical psychologist. He derived a form of equation (5.14) by making assumptions about the choice probabilities rather than about random utilities. To be consistent with the rest of this book, we derive equation (5.14) by restating a result of McFadden (1974) whose origin is attributed to Marschak (1960) and Marley (as reported by Luce and Suppes, 1965.) The particular line of proof uses equation (5.13), though other forms, such as equations (5.9) or (5.10), could serve equally well. Actually, (5.10) is used to derive the logit model in Section 3.A.

The logit model can be derived using the properties of the extreme value distribution, described in Section 4.2.2. Our proof is a variant of one by Domencich and McFadden (1975); for convenience we assume a zero location parameter, $\eta = 0$ for all the disturbances. If we, as before, order the alternatives so that $i = 1$, then

$$P_n(1) = \Pr\left(V_{1n} + \varepsilon_{1n} \geq \max_{j=2,\ldots,J_n} (V_{jn} + \varepsilon_{jn})\right). \tag{5.17}$$

Define

$$U_n^* = \max_{j=2,\ldots,J_n} (V_{jn} + \varepsilon_{jn}). \tag{5.18}$$

From property 6 of the extreme value distribution, U_n^* is EV distributed with parameters

$$\left(\frac{1}{\mu} \ln \sum_{j=2}^{J_n} e^{\mu V_{jn}}, \mu\right).$$

Using property 4, we can write $U_n^* = V_n^* + \varepsilon_n^*$, where

$$V_n^* = \frac{1}{\mu} \ln \sum_{j=2}^{J_n} e^{\mu V_{jn}}$$

and ε_n^* is EV distributed with parameters $(0, \mu)$.

Since

$$\begin{aligned} P_n(1) &= \Pr(V_{1n} + \varepsilon_{1n} \geq V_n^* + \varepsilon_n^*) \\ &= \Pr((V_n^* + \varepsilon_n^*) - (V_{1n} + \varepsilon_{1n}) \leq 0), \end{aligned} \tag{5.19}$$

by property 5 we have that

$$\begin{aligned} P_n(1) &= \frac{1}{1 + e^{\mu(V_n^* - V_{1n})}} \\ &= \frac{e^{\mu V_{1n}}}{e^{\mu V_{1n}} + e^{\mu V_n^*}} \\ &= \frac{e^{\mu V_{1n}}}{e^{\mu V_{1n}} + e^{\ln \sum_{j=2}^{J_n} e^{\mu V_{jn}}}} \\ &= \frac{e^{\mu V_{1n}}}{\sum_{j=1}^{J_n} e^{\mu V_{jn}}}. \end{aligned} \tag{5.20}$$

Note the presence of the scale parameter μ in each of the terms of (5.20). This parameter is usually not identifiable, just as it was not identifiable in

the binary case (see section 4.2), so the usual procedure is to set it arbitrarily to a convenient value, such as 1. Although this is operationally necessary, we must not let it obscure our understanding of the role of μ in the model. Through it we reflect the assumption of disturbances with equal variance; if this assumption is inappropriate for the population in question, it is necessary to take suitable measures (see section 6.6) to model correctly the choice utilities.

Limiting Cases of the Logit Model

As in the binary case (see section 4.2), there are two limiting cases of the logit model that result from extreme values of μ:

$\mu \to 0$

$$\lim_{\mu \to 0} P_n(i) = \frac{1}{J_n}, \ \forall i \in C_n.$$

As $\mu \to 0$, the variance of the disturbances approaches infinity. The choice model then provides no information, so the alternatives are equally likely.

$\mu \to \infty$

$$\begin{aligned} \lim_{\mu \to +\infty} P_n(i) &= \lim_{\mu \to +\infty} \frac{1}{1 + \sum_{j \in C_n, j \neq i} e^{\mu(V_{jn} - V_{in})}} \\ &= \begin{cases} 1 & \text{if } V_{in} > \max_{j \in C_n, j \neq i} V_{jn}, \\ 0 & \text{if } V_{in} < \max_{j \in C_n, j \neq i} V_{jn}. \end{cases} \end{aligned}$$

In the event of a tie among the utilities for some of the alternatives,

$$V_{in} = \max_{j \in C_n, j \neq i} V_{jn},$$

the limit is $1/J_n^*$ for the J_n^* alternatives for which

$$V_{in} = \max_{j \in C_n} V_{jn}, \ i = 1, \ldots, J_n^*,$$

and is zero for the remaining $J_n - J_n^*$ alternatives. Note that as $\mu \to \infty$, the variance of the utility disturbances approaches zero and a deterministic choice model is obtained because all the information about individual preferences is included in the systematic utilities.

Linear-in-Parameters Logit Model

Up to this point in our discussion we have not imposed any functional form on V_{in}, the systematic component of the utility function. As in the case of

binary choice it is generally computationally convenient to restrict V_{in} to the class of linear-in-parameters functions. Following our convention of defining a single vector of coefficients β that applies to all the utility functions, we can write this restricted version of logit as

$$P_n(i) = \frac{e^{\mu\tilde{\beta}^T x_{in}}}{\sum_{j=1}^{J_n} e^{\mu\tilde{\beta}^T x_{jn}}} = \frac{e^{\beta^T x_{in}}}{\sum_{j=1}^{J_n} e^{\beta^T x_{jn}}}, \tag{5.21}$$

where x_{in} and x_{jn} are vectors describing the attributes of alternatives i and j, and $\beta = \mu\tilde{\beta}$ are unknown parameters to be estimated from data.

5.3 Properties of Logit

5.3.1 Independence from Irrelevant Alternatives Property (IIA)

One of the most widely discussed aspects of the logit model is the Independence from Irrelevant Alternatives property, or IIA.

Stated succinctly, the IIA property holds that for a specific individual the ratio of the choice probabilities of any two alternatives is entirely unaffected by the presence (or absence) of any other alternatives in the choice set and by the systematic utilities of any other alternatives. In other words, the odds ratio of i and j is unaffected by the choice set composition, provided, of course, that the choice set must include i and j. This can be easily shown to hold in the case of logit as follows:

$$\begin{aligned} \frac{P_n(i)}{P_n(\ell)} &= \frac{e^{\mu V_{in}} / \sum_{j\in C_n} e^{\mu V_{jn}}}{e^{\mu V_{\ell n}} / \sum_{j\in C_n} e^{\mu V_{jn}}} \\ &= \frac{e^{\mu V_{in}}}{e^{\mu V_{\ell n}}} \\ &= e^{\mu(V_{in} - V_{\ell n})}. \end{aligned} \tag{5.22}$$

This seemingly simple property has some important ramifications. In some instances it can give rise to somewhat odd and erroneous predictions. One of the most widely cited anomalies is the *red bus/blue bus paradox* described in Chapter 3.

To understand fully the IIA property and its implications, it is useful to go back to the assumptions from which the logit model is derived from the random utility model. Although all of the assumptions listed in section

5.2 are needed to produce the specific form of the IIA property, the core of the problem is the assumption that the disturbances are mutually independent. This assumption requires that the sources of errors contributing to the disturbances must do so in a way such that the total disturbances are independent. In the case of red buses and blue buses this is wholly implausible since both these alternatives share all the unobserved characteristics of buses. In fact, rather than being independent, the disturbances of the red and blue bus modes are more reasonably assumed to be perfectly correlated. *Although models other than logit might produce different numerical results, any model based on the assumption that all the disturbances are independent would necessarily yield counterintuitive forecasts for the red bus/blue bus problem.*

Another way to see why logit yields incorrect forecasts for this situation is to trace through the steps in its derivation. Let us define alternatives 1,2, and 3 as auto, red bus, and blue bus, respectively. By the construction of the original paradox, $V_{1n} = V_{2n} = V_{3n}$ so we let V_n be the systematic component of the utility for any of the three alternatives. If we follow the line of proof in section 5.2, we note that we relied on the fact that

$$U_n^* = \max(U_{2n}, U_{3n}) = V_n^* + \varepsilon_n^*, \tag{5.23}$$

where $V_n^* = \ln(e^{V_n} + e^{V_n}) = \ln 2 + V_n$ and ε_n^* is EV distributed with parameters $(0, 1)$. This step assumed that the disturbances for alternatives 2 and 3 were independent. However, if they are perfectly correlated, then the maximum of U_{2n} and U_{3n} is, for the case of the two bus modes with equal systematic utility, simply $V_n + \varepsilon_n$. In a sense the assumption of independence makes V_n^* too high by an amount equal to $\ln 2$. Moreover, if we correct this overestimate, then

$$P_n(\text{auto}) = \Pr(V_n + \varepsilon_{1n} \geq V_n^* + \varepsilon_n^*) = \frac{1}{1 + e^{V_n - V_n}} = \frac{1}{2}, \tag{5.24}$$

which is the intuitively correct answer. Thus it would be possible to use a logit model for the choice among the three alternatives in this example only if we define the systematic utilities for the two bus alternatives as $V_n - \ln 2$.

A common misinterpretation of the IIA property is that it applies to the population as a whole. Thus it is often interpreted as implying that the ratio of the *shares of the population* choosing any two alternatives is unaffected by the utilities of other alternatives. Except in one extreme (and wholly unrealistic) case, this is simply not true. We demonstrate this by the following simple counterexample.

Suppose that the destinations of shopping trips in a small city are equally divided between a suburban shopping center and the downtown area. If a

new suburban center with observed attributes that are identical to the first one is built, would the logit model predict that it would draw one-third of the population, taken equally from the trips to the downtown area and the original suburban center? The answer is, in general, no.

To see this, suppose that the entire population of trips is composed of only two, equal-sized groups that are internally homogeneous in their observed attributes. The first group has predicted choice probabilities of 0.95 and 0.05 for the downtown and the original suburban center, respectively; for the second group, the probabilities are reversed. Table 5.1 summarizes the predicted probabilities before and after the second suburban center is introduced. Note that though the IIA property does apply to each homogeneous group, *it does not apply to the population as a whole.* In this example, rather than forecasting a shift of the downtown share of shopping destinations from 1/2 to 1/3 (as in the red bus/blue bus case), the share decreases only to 0.4652.

	P_n(downtown)	P_n(suburb 1)	P_n(suburb 2)
Before new center			
Group 1	0.95	0.05	
Group 2	0.05	0.95	
Population share	0.50	0.50	
After new center			
Group 1	0.9048	0.0476	0.0476
Group 2	0.0256	0.4872	0.4872
Population share	0.4652	0.2674	0.2674

Table 5.1: Data for a heterogeneous population without the IIA property

The key here is that the choice probabilities before the new shopping center is added are not the same across the population. Rather, there are two distinct market segments in the population with very different systematic preferences.

The idea that IIA may be more or less believable depending on whether heterogeneities in the population are accounted for in the model is a significant one. It implies that whether or not logit is an appropriate model for a particular choice situation is not something that can be judged in the abstract. Rather, one must examine the particular specification of the systematic component of the utility function and ask if it reasonably accounts for population heterogeneities. As a practical consequence logit models that include in the systematic utility specification decision-makers's characteristics in an appropriate way stand a far better chance of yielding reasonable forecasts than those that omit such variables.

We return to the problem of identifying when the IIA property is violated in chapter 6 where we discuss statistical tests for specification errors.

5.3.2 Choice set generation

As described in Section 5.1, the choice set $\mathcal{C}_n$ may vary across individuals, justifying the index n. However, the process to identify which alternatives are actually considered by individual n is usually not straightforward.

In general, it is convenient to use deterministic rules. For example, in a transportation mode choice context, the alternative "driving a car" is declared not available to individuals not in possession of a driving licence, or not in possession of a car; the alternative "walking" is declared not available if the distance is larger than 3 kilometers, say; etc.

With known availability of the alternatives, it is useful to rewrite the logit model, introducing the availability variables into the utility functions as follows:

Let A_{in} be 1 if individual n considers or has available alternative i, and 0 otherwise. For instance, in the above example where "walking" is considered not available when the distance is larger than 3 kilometers, we have

$$A_{\text{Walk},n} = \begin{cases} 1 & \text{if } d_n < 3, \\ 0 & \text{if } d_n \geq 3, \end{cases} \tag{5.25}$$

where d_n is the distance in kilometers to be traveled by individual n. Using the A_{in}'s we rewrite the choice model based on the universal choice set:

$$\begin{aligned} P_n(i|C_n) &= \Pr(U_{in} \geq U_{jn}, \forall j \in \mathcal{C}_n) \\ &= \Pr(U_{in} + \ln A_{in} \geq U_{jn} + \ln A_{jn}, \forall j \in \mathcal{C}), \end{aligned} \tag{5.26}$$

which for logit means

$$P_n(i|C_n) = \frac{e^{\mu V_{in}}}{\sum_{j\in\mathcal{C}_n} e^{\mu V_{jn}}} = \frac{e^{\mu V_{in}+\mu \ln A_{in}}}{\sum_{j\in\mathcal{C}} e^{\mu V_{jn}+\mu \ln A_{jn}}} = \frac{A_{in}^{\mu} e^{\mu V_{in}}}{\sum_{j\in\mathcal{C}} A_{jn}^{\mu} e^{\mu V_{jn}}}. \tag{5.27}$$

It is not always possible to deterministically identify the choice set of each individual in the population. Probabilistic choice set generation procedures must then be used (see Manski, 1977, Swait and Ben-Akiva, 1987b, Swait and Ben-Akiva, 1987a, Siddarth et al., 1995.) Let $\mathcal{G}_n$ be the collection of all possible non-empty choice sets for individual n. Then, the choice model writes:

$$P_n(i) = \sum_{\mathcal{C}\in\mathcal{G}_n} P_n(i|\mathcal{C}) \Pr(\mathcal{C}|\mathcal{G}_n), \tag{5.28}$$

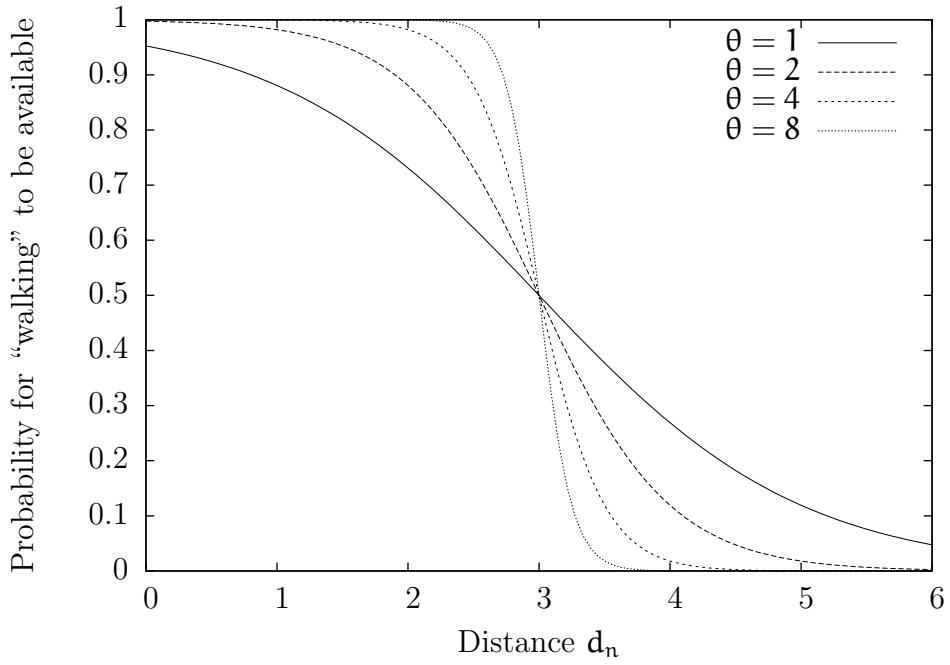


Figure 5.2: Probabilistic availability of an alternative

where $P_n(i|\mathcal{C})$ is the choice model conditional on choice set $\mathcal{C}$, and $\Pr(\mathcal{C}|\mathcal{G}_n)$ is the probability that individual n considers choice set $\mathcal{C}$ among the set of all feasible choice sets denoted by $\mathcal{G}_N$.

For the transportation mode choice example, we may include in the model two choice sets: $\mathcal{C}$, the universal choice set, and $\tilde{\mathcal{C}} = \mathcal{C} \setminus \{\text{walking}\}$, that is the universal choice set without the "walking" alternative. We must define the probability for each of these choice sets to be actually considered. For instance, we may define a binary logit model:

$$\Pr(\mathcal{C}|\{\mathcal{C}, \tilde{\mathcal{C}}\}) = \frac{1}{1 + e^{\theta(d_n - 3)}}, \tag{5.29}$$

where θ is a parameter that can be predefined, or estimated from data. This probability function is represented for various values of θ in Figure 5.2. As there are only two choice sets, $\Pr(\tilde{\mathcal{C}}) = 1 - \Pr(\mathcal{C})$, and (5.28) writes

$$P_n(i) = \frac{1}{1 + e^{\theta(d_n - 3)}} P_n(i|\mathcal{C}) + \frac{e^{\theta(d_n - 3)}}{1 + e^{\theta(d_n - 3)}} P_n(i|\tilde{\mathcal{C}}).$$

Model (5.28) becomes very complex in most practical cases, due to the combinatorial nature of $\mathcal{G}_n$. Ben-Akiva (1977) proposes an application of

the logit captivity model which is the same as the "dogit" model by Gaudry and Dagenais (1979). A parametrized logit captivity model was estimated by Swait and Ben-Akiva (1987a). Swait and Ben-Akiva (1987b) also derive (5.28) from random constraints. Such a model was specified and estimated by Ben-Akiva and Boccara (1995). Swait (2001) derives an operational model by showing that the choice set generation model (5.28) can be consistent with the family of Multivariate Extreme Value models (described in Chapter 8.)

Some authors (such as Ben-Akiva and Gershenfeld, 1998, Cascetta and Papola, 2001, Martinez et al., 2009) have exploited the formulation (5.27) to propose an approximation of a choice set generation model which is more tractable. The idea is to replace the 0/1 variable A_{in} by a probability distribution function. This can somehow be interpreted as the probability for alternative i to be considered by individual n. Using again the walking distance example, A_{in} can be defined for example in the same way as (5.29):

$$A_{in} = \frac{1}{1 + e^{\theta(d_n - 3)}},$$

where i corresponds to the "walking" alternative, and θ is a parameter that can be predefined, or estimated from data. As $\ln(A_{in}) = -\ln(1 + e^{\theta(d_n-3)})$, the utility of "walking" in the model is decreased by $\ln(1 + e^{\theta(d_n-3)})$. If θ is estimated, we obtain a tractable model with a nonlinear utility function. However, it is only a convenient approximation, which may happen to be poor, as illustrated by Bierlaire et al. (2010). A correctly specified model would be in the form of (5.28).

5.3.3 Expected maximum utility

As discussed in Section 3.7.3, the *expected maximum utility* (EMU) is defined by

$$\mathrm{E}[\max_{i \in \mathcal{C}_n} U_{in}]. \tag{5.30}$$

For the logit model, it has a relatively tractable form. In particular, if $\mathcal{C}_n$ is a choice set, for logit

$$\mathrm{E}[\max_{i \in \mathcal{C}_n} U_{in}] = \frac{1}{\mu}(\ln \sum_{i \in \mathcal{C}_n} e^{\mu V_{in}} + \gamma), \tag{5.31}$$

where γ is Euler's constant. In general, the expected maximum utility is used to compute differences, so that the γ cancels out and can be ignored.

5.4 Specification of the systematic component

The specification of the systematic utilities of a choice model with multiple alternatives is a direct extension of the specification of a binary choice model (see Section 4.1.) As in binary choice, adding the same constant to all J_n utilities would not change the probability. Similarly, only differences in utilities matter, as shown by (5.3) and (5.4). Consequently, alternative specific constants may be included in the utilities of $J-1$ alternatives, one alternative being used as a reference (see Bierlaire et al., 1997 for other possible normalizations.) The selection of the referent or base alternative has no effect on the model other than to shift the values of the alternative specific constants, preserving their differences. This property holds even when the choice set varies across observations. Indeed, as described in Section 5.3.2, the model can always be rewritten based on the universal choice set.

As the ASCs capture the mean of the error term ε_{in}, they may vary with n. In practice, the population is divided into homogenous segments, defined by decision-maker characteristics such as age, income or gender. Each segment may be associated with a set of alternative specific constants. The utility function is specified as described in Section 4.1, and is illustrated by the following example.

Consider four education levels, with categories characterized by the highest degree (elementary, high school, college, post-graduate), two genders and two levels of income (low, high.) A reference level is arbitrary selected for each variable: elementary, male, low income. For each characteristic, we introduce a dummy variable for each level but the reference level, and include the following function for the alternative specific part of a utility:

$$\begin{aligned}\beta_0 \;+\; & \beta_{\text{college}}\text{DummyCollege} \;+\; \beta_{\text{highSchool}}\text{DummyHighSchool} \\ & \qquad\qquad\qquad\quad +\; \beta_{\text{postgraduate}}\text{DummyPostgraduate} \\ & +\; \beta_{\text{female}}\text{DummyFemale} \\ & +\; \beta_{\text{high}}\text{DummyHighIncome}.\end{aligned} \tag{5.32}$$

With this formulation, the constant associated with the reference segment (young, male, low income) is β_0. The constant associated with other segments is the sum of β_0 and relevant coefficients. For instance, the constant for the segment (adults, male, high income) is $\beta_0 + \beta_{\text{adult}} + \beta_{\text{high}}$.

Clearly, the terms with the dummy variables are also defined in all alternatives but one. Although not necessary from a specification viewpoint, using the same alternative as a reference for all these terms is helpful for the interpretation of estimation results.

5.4.1 Capturing nonlinearities in the utility function

The linear-in-parameters specification (5.21) can capture nonlinear influences of the variables on the utility function. In fact, x_{in} may be defined as any known nonlinear function of the original variables. For instance, x_{in} can be the logarithm of the price of the alternative, and its coefficients captures the effect on the utility of a relative change in price whereas a coefficient of linear price captures the effect of an absolute change in price.

More complex specifications can also be considered.

Splines A *piecewise linear*, or *spline*, specification can also be considered to capture nonlinearity. Let x be the variable under interest, and βx the corresponding term in the utility function (we drop the indices i and n for notational simplification.) We divide the range of values for x into M intervals $[a_m, a_{m+1})$, $m = 1, \ldots, M$. For each interval, we define a new variable

$$x_m = \begin{cases} 0 & \text{if } x < a_m \\ x - a_m & \text{if } a_m \leq x < a_{m+1} \\ a_{m+1} - a_m & \text{otherwise} \end{cases} \tag{5.33}$$

or, equivalently

$$x_m = \max(0, \min(x - a_m, a_{m+1} - a_m)). \tag{5.34}$$

We define also the variables

$$x_0 = \min(x, a_1) = \begin{cases} x & \text{if } x < a_1 \\ a_1 & \text{otherwise} \end{cases} \tag{5.35}$$

and

$$x_{M+1} = \max(0, x - a_{M+1}) = \begin{cases} 0 & \text{if } x < a_{M+1} \\ x - a_{M+1} & \text{otherwise} \end{cases} \tag{5.36}$$

corresponding to the intervals $(-\infty, a_1)$ and $[a_{M+1}, +\infty)$, respectively. Another, more intuitive way of defining the variables x_m is the following. Let $[a_\ell, a_{\ell+1})$ be the interval containing x, that is $x \in [a_\ell, a_{\ell+1})$. Then the variable x_ℓ is defined as $x - a_\ell$. Basically, the value a_ℓ is associated with previous intervals, and what is left is associated with x_ℓ. All the values of x_m before ($m \leq \ell$) are equal to the size of the corresponding interval, that is $x_m = a_{m+1} - a_m$, and all the values after

($m > \ell$) are zero. As a example, consider $M = 2$, $a_1 = -1$, $a_2 = 5$ and $a_3 = 9$. We present below the coding of four different values of x.

x	x_0	x_1	x_2	x_3
-5	-5	0	0	0
2	-1	3	0	0
8	-1	6	3	0
23	-1	6	4	14

Note that we always have

$$\sum_{m=0}^{M+1} x_m = x. \tag{5.37}$$

The term βx is replaced in the utility function by

$$\sum_{m=0}^{M+1} \beta_m x_m. \tag{5.38}$$

This is still a linear-in-parameter formulation. If the value of β_m is the same for all m, that is $\beta_m = \beta$ for all m, then (5.38) reduces to βx, as a consequence of (5.37). An example of a model with a piecewise linear specification is discussed in Section 5.8. Another one is reported in Table 6.7 and illustrated in Figure 6.3.

Power series The *power series* specification allows also to capture the potential nonlinear effect of a variable, while keeping a linear-in-parameter specification. Assume that the analyst believes that the variable x influences the utility function in a nonlinear way, that is

$$V = \cdots + \beta f(x) + \cdots, \tag{5.39}$$

where f is an unknown nonlinear function of x. From Taylor's theorem, we know that any function f, if sufficiently differentiable, can be approximated by the power series

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + \cdots + \frac{f^{(p)}(0)}{p!}x^p + \cdots. \tag{5.40}$$

Ignoring the constant, truncating the series after the pth term, and considering the coefficients of the power series as unknown parameters, (5.39) can be written as

$$V = \cdots + \beta_1 x + \beta_2 x^2 + \cdots \beta_p x^p + \cdots. \tag{5.41}$$

In practice, only a small number of such terms is included in the utility function (typically 2 or 3.) An example with the linear and the quadratic term is reported in Table 6.9.

Box-Cox The *Box-Cox* transformation is a parametric family of transformations from x to $x^{(\lambda)}$ proposed by Box and Cox (1964), where the parameter λ is possibly a vector. Two important instances are defined as

$$x^{(\lambda)} = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln x & \text{if } \lambda = 0, \end{cases} \tag{5.42}$$

and

$$x^{(\lambda)} = \begin{cases} \dfrac{(x + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if } \lambda_1 \neq 0, \\ \ln(x + \lambda_2) & \text{if } \lambda_1 = 0. \end{cases} \tag{5.43}$$

The transformation (5.42) holds for $x > 0$. It is illustrated in Figure 5.3 for various values of λ. It is seen that the transform is convex in x if $\lambda \geq 1$, and concave if $\lambda \leq 1$. By construction, it is continuous at $\lambda = 0$, as

$$\lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} = \ln x. \tag{5.44}$$

The transformation (5.43) holds for $x > -\lambda_2$.

Assume now that x is a positive explanatory variable of a choice model. It can be included in the utility function using transformation (5.42) as follows:

$$V = \cdots + \beta \frac{x^\lambda - 1}{\lambda} + \cdots, \tag{5.45}$$

where β and λ are unknown parameters to be estimated. The utility function is **not** linear-in-parameter anymore. Note that, when $\lambda = 1$, Eq. (5.45) becomes

$$V = \cdots + \beta(x - 1) + \cdots = \cdots + \beta x - \beta + \cdots, \tag{5.46}$$

which is the standard linear specification, up to a constant.

5.4.2 Interactions

Tastes may vary across the population. This is captured in the specification of the utility function by *interactions*. Let x be an explanatory variable of a

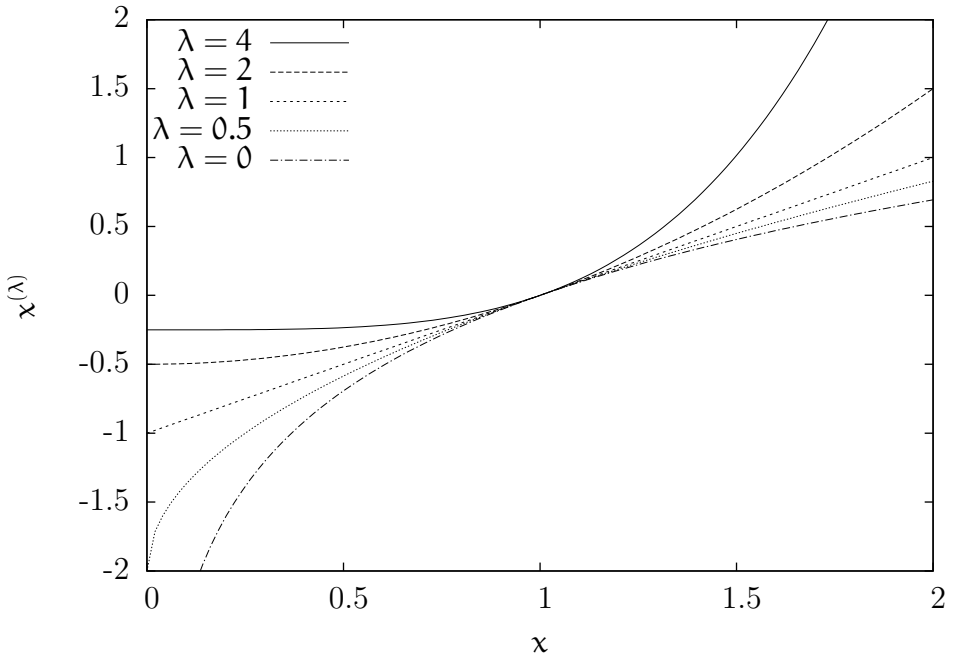


Figure 5.3: Box-Cox transformations

choice model, involved linearly in the utility function, that is

$$V = \cdots + \beta x + \cdots . \tag{5.47}$$

Let s_n be a socio-economic characteristic associated with individual n (e.g. age, income, education or profession.) If s_n is a categorical variable (like education or profession) with C categories, the interaction consists in defining an unknown parameter per category $\beta_1, \ldots, \beta_C$, and define the utility function as

$$V = \cdots + \sum_{c=1}^{C} \beta_c \delta_c(s_n) x + \cdots , \tag{5.48}$$

where $\delta_c(s_n)$ is 1 if the value of s_n belongs to category c, 0 otherwise.

If s_n is a continuous variable (like age or income), the analyst can assume how the unknown parameter β varies with s_n. The simplest case consists in a linear relationship, where

$$\beta = \widehat{\beta} \frac{s_n}{s_n^{\text{ref}}}, \tag{5.49}$$

where s_n^{ref} is an arbitrary reference value of the socio-economic characteristic, such as the population mean, mode, minimum or maximum. Therefore, the

specification of the utility function becomes

$$V = \cdots + \widehat{\beta}\, \frac{s_n}{s_n^{\text{ref}}}\, x + \cdots . \tag{5.50}$$

The unknown parameter $\widehat{\beta}$ represents the sensitivity of the utility function to the variable x when $s_n = s_n^{\text{ref}}$. Note that the specifications (5.48) and (5.50) are linear-in-parameter.

Nonlinear relations can be considered as well. A typical example consists in defining

$$\beta = \widehat{\beta} \left(\frac{s_n}{s_n^{\text{ref}}} \right)^{\lambda}, \tag{5.51}$$

where s_n^{ref} is defined as above, and $\widehat{\beta}$ and λ are unknown parameters to be estimated. Incorporating (5.51) into (5.47), we obtain

$$V = \cdots + \widehat{\beta} \left(\frac{s_n}{s_n^{\text{ref}}} \right)^{\lambda} x + \cdots . \tag{5.52}$$

The parameter λ represents the elasticity of β with respect to variations of s_n, that is

$$\lambda = \frac{\partial \beta}{\partial s_n} \frac{s_n}{\beta}. \tag{5.53}$$

Specification (5.52) is not linear-in-parameter.

5.4.3 Segments with different variances

One of the assumptions used to derive the logit model is that the error terms ε_{in} are identically distributed across i and n. In particular, it means that the variance of ε_{in} is the same for each individual n in the population. This assumption may be incorrect in some circumstances. For instance, when some individuals are experimented in the choice under interest, and some others not, it is fair to assume that the variance of the error term for the unexperienced individuals is higher. Also, it may happen that several sources of data are combined to estimate a choice model, such as revealed and stated preferences. Because the data has not been collected in the same context, it is fair to assume that the variance of the error term associated with the decision-makers in each data set is different.

We show that this situation can be modeled using a logit model and a nonlinear specification of the utility function. Consider that the population is partitioned into G groups, $g = 1, \ldots, G$, such that we can safely assume

that the variance of the error terms is constant within each group but may vary across groups. More precisely, assume that

$$\text{Var}(\varepsilon_{in}) = \alpha_g^2 \tag{5.54}$$

if individual n belongs to group g. Given the utility function

$$U_{in} = V_{in} + \varepsilon_{in}, \tag{5.55}$$

we know that multiplying the utility function of each alternative by a positive constant does not change the choice model. Indeed,

$$\Pr(U_{in} \geq U_{jn}, \forall j \in \mathcal{C}_n) = \Pr(\alpha U_{in} \geq \alpha U_{jn}, \forall j \in \mathcal{C}_n) \quad \forall \alpha > 0. \tag{5.56}$$

We arbitrarily select group 1 as a reference, and we multiply the utility functions of individual n by α_1/α_g, where g is the group containing n. We obtain

$$U'_{in} = \frac{\alpha_1}{\alpha_g} U_{in} = \frac{\alpha_1}{\alpha_g} V_{in} + \frac{\alpha_1}{\alpha_g} \varepsilon_{in}. \tag{5.57}$$

Now, we define the random variable

$$\varepsilon'_{in} = \frac{\alpha_1}{\alpha_g} \varepsilon_{in}, \tag{5.58}$$

which is such that

$$\text{Var}(\varepsilon'_{in}) = \text{Var}\left(\frac{\alpha_1}{\alpha_g} \varepsilon_{in}\right) = \frac{\alpha_1^2}{\alpha_g^2} \text{Var}\left(\varepsilon_{in}\right) = \alpha_1^2, \text{ for each } n.$$

Therefore, the variance of

$$U'_{in} = \frac{\alpha_1}{\alpha_g} V_{in} + \varepsilon'_{in} \tag{5.59}$$

is the same for every individual in the population, so that the i.i.d. assumption is now appropriate. Consequently, we define an unknown parameter λ_g for each group, except group 1, and use the following specification:

$$U'_{in} = \lambda_g V_{in} + \varepsilon'_{in}, \tag{5.60}$$

where λ_g is estimated from data for all $g > 1$. Clearly, even if V_{in} is linear-in-parameter, the new specification is not, as λ_g multiplies V_{in}. An estimated value of $\lambda_g > 1$ means that the variance of the error term for group g is smaller than the variance for group 1. Symetrically, $\lambda_g < 1$ means that the variance for group g is larger than for group 1.

5.4.4 Multiplicative error terms

The additive form of the random utility $U_{in} = V_{in} + \varepsilon_{in}$ is widely used, as it is the most natural and the most convenient. But alternative specifications can be considered. Fosgerau and Bierlaire (2009) propose a multiplicative specification, such that

$$U_{in} = V_{in}\varepsilon_{in}. \tag{5.61}$$

This specification is valid if the signs of V_{in} and ε_{in} are known. It leads to a model such that the choice probabilities are no longer invariant with respect to addition of a constant to all the V_j's, instead they are invariant with respect to multiplication of all V_j's by a positive constant.

To derive an operational model, assume that $V_{in} < 0$ is the systematic part of the utility function, and $\varepsilon_{in} > 0$ is a random variable, independent of V_{in}. We assume that the ε_{in} are i.i.d. across individuals. The sign restriction on V_{in} is a natural assumption in many applications, for example when it is defined as a generalized cost, that is, a linear combination of attributes with positive values such as travel time and cost and parameters that are a priori known to be negative.

The choice probabilities under this model are given by

$$P_n(i) = \Pr(V_{in}\varepsilon_{in} \geq V_{jn}\varepsilon_{jn}, \ \ \forall j \in \mathcal{C}_n). \tag{5.62}$$

The multiplicative specification is related to the classical specification with additive independent error terms, as can be seen from the following derivation. The logarithm is a strictly increasing function. Consequently,

$$\begin{aligned} P_n(i) &= \Pr(V_{in}\varepsilon_{in} \geq V_{jn}\varepsilon_{jn}, \ \ \forall j) \\ &= \Pr(-\ln(-V_{in}) - \ln(\varepsilon_{in}) \geq -\ln(-V_{jn}) - \ln(\varepsilon_{jn}), \ \ \forall j). \end{aligned}$$

We define

$$-\ln(\varepsilon_{jn}) = \xi_{jn}/\lambda, \tag{5.63}$$

where ξ_{jn} are random variables, and $\lambda > 0$ is a scale parameter associated with ξ_{jn}, which is constant across j and n from the i.i.d. assumption. We obtain

$$P_n(i) = \Pr(-\lambda \ln(-V_{in}) + \xi_{in} \geq -\lambda \ln(-V_{jn}) + \xi_{jn}, j \in \mathcal{C}). \tag{5.64}$$

Defining

$$\bar{V}_{in} = -\lambda \ln(-V_{in}), \tag{5.65}$$

we have

$$P_n(i) = \Pr(\bar{V}_{in} + \xi_{in} \geq \bar{V}_{jn} + \xi_{jn}, j \in \mathcal{C}), \tag{5.66}$$

which is a random utility model with an additive specification and a non linear utility function.

This specification is fairly general, as we are free to make assumptions regarding the error terms ξ_{in}. For instance, assuming that the ξ_{in} are i.i.d. extreme value, the choice model becomes

$$P_n(i) = \frac{e^{-\lambda \ln(-V_{in})}}{\sum_{j \in \mathcal{C}} e^{-\lambda \ln(-V_{jn})}} = \frac{-V_{in}^{-\lambda}}{\sum_{j \in \mathcal{C}} -V_{jn}^{-\lambda}}. \tag{5.67}$$

Fosgerau and Bierlaire (2009) report a case study such that the multiplicative formulation fits the data better in the majority of the cases they have looked at.

5.5 The example of a model for the airline itinerary choice

The specification of a logit model consists of a number of distinct steps. First, we must define the universal choice set $\mathcal{C}$ for the problem under study. This step may require some judgments about which alternatives can be ignored given the objectives of the analysis and the data availability.

The next step is to define the choice set for each individual. As discussed before, this is generally done by applying reasonable judgments about what constitutes the feasibility of an alternative in any particular situation.

Finally, the particular variables entering into the utility functions must be defined. Our goal in this section is to illustrate at least some of the issues involved in developing a choice model through an example based on a study about choice of airline itinerary. This is the example which has been already used in Section 4.3.2. A description of the data is also presented in the Appendix E.3.

We now consider the choice between the three alternatives: (1) a non stop flight, (2) one stop with the same airline, (3) one stop and a change of airline. These three alternatives constitutes the full choice set. In this stated preferences experiment, all alternatives have always been presented to the respondents. Therefore, the choise set $\mathcal{C}_n$ is the same for every observation n.

The specification of the model is described in Table 5.2. This table is identical in format to those used to describe binary choice models in chapter 4. The only difference is that here there are 3 rather than two columns, one for each alternative. As in the case of binary choice models, the variables include

alternative-specific constants, attributes of the alternatives, and alternative-specific socioeconomic characteristics.

Most of the issues raised in discussing the specification of binary models extend directly to choice with multiple alternatives. For example, in binary choice models it made sense to have only one constant. Its coefficient reflected the relative utility (all else equal) of the alternative in which the constant was included as compared to the one from which it was omitted. When there are more than two alternatives, we must again have an alternative that acts as a referent, so we can logically include two constants in the present example, where the referent alternative is (1), the non stop flight.

In the general case, we can include as many as $J_n - 1$ constants. As in binary choice, the selection of the referent alternative has no effect on the model other than to shift the values of the estimated constants, preserving their differences.

Similarly we can use any particular socioeconomic variable as alternative specific in as many as $J_n - 1$ alternatives. In our example the socioeconomic attributes "more than 2 leisure trips per year" and the gender appears in all alternatives but (1). The estimated model is discussed in Section 5.7. More specifications for this data set are discussed in Chapter 6.

5.6 Estimation of Logit

There is little additional material to be discussed regarding the estimation of choice models with multiple alternatives beyond what is said in chapter 4 for binary models. All of the general procedures extend straightforwardly, though their computational burden grows as the number of alternatives increases. Rather than simply repeat much of section 4.4 in more general form, we consider only the estimation of the logit model. Logit has some special properties that under certain circumstances greatly simplify estimation of its parameters. Most of this theory is attributable to McFadden (1974).

5.6.1 Maximum Likelihood

As in chapter 4, let N denote the sample size and define

$$y_{in} = \begin{cases} 1 & \text{if observation } n \text{ chose alternative } i, \\ 0 & \text{otherwise.} \end{cases}$$

The likelihood function for a general choice model is

$$\mathcal{L}^* = \prod_{n=1}^{N} \prod_{i \in \mathcal{C}_n} P_n(i)^{y_{in}}, \tag{5.68}$$

where for logit

$$P_n(i) = \frac{e^{V_{in}}}{\sum_{j \in \mathcal{C}_n} e^{V_{jn}}}. \tag{5.69}$$

Taking the logarithm of equation (5.69), we seek a maximum to

$$\begin{aligned} \mathcal{L} &= \sum_{n=1}^{N} \sum_{i \in \mathcal{C}_n} y_{in} \left(V_{in} - \ln \sum_{j \in \mathcal{C}_n} e^{V_{jn}} \right) \\ &= \sum_{n=1}^{N} \left(\sum_{i \in \mathcal{C}_n} y_{in} V_{in} - \ln \sum_{i \in \mathcal{C}_n} e^{V_{in}} \right), \end{aligned} \tag{5.70}$$

as $\sum_{i \in \mathcal{C}_n} y_{in} = 1$. Setting the first derivatives of $\mathcal{L}$ with respect to the coefficients equal to zero, we obtain, for $k = 1, \ldots, K$, the necessary first-order conditions:

$$\frac{\partial \mathcal{L}}{\partial \beta_k} = \sum_{n=1}^{N} \left(\sum_{i \in \mathcal{C}_n} y_{in} \frac{\partial V_{in}}{\partial \beta_k} - \sum_{i \in \mathcal{C}_n} P_n(i) \frac{\partial V_{in}}{\partial \beta_k} \right) = 0. \tag{5.71}$$

Rearranging the terms, we obtain, for $k = 1, \ldots, K$,

$$\frac{\partial \mathcal{L}}{\partial \beta_k} = \sum_{n=1}^{N} \sum_{i \in \mathcal{C}_n} (y_{in} - P_n(i)) \frac{\partial V_{in}}{\partial \beta_k} = 0. \tag{5.72}$$

For a linear-in-parameters logit, it is

$$\sum_{n=1}^{N} \sum_{i \in \mathcal{C}_n} (y_{in} - P_n(i)) x_{ink} = 0, \text{ for } k = 1, \ldots, K. \tag{5.73}$$

The reader can verify that, in this case, the second derivatives of $\mathcal{L}$ are given by

$$\frac{\partial^2 \mathcal{L}}{\beta_k \beta_\ell} = -\sum_{n=1}^{N} \sum_{i \in \mathcal{C}_n} P_n(i) \left(x_{ink} - \sum_{j \in \mathcal{C}_n} x_{jnk} P_n(j) \right) \left(x_{in\ell} - \sum_{j \in \mathcal{C}_n} x_{jn\ell} P_n(j) \right). \tag{5.74}$$

All of the properties of the maximum likelihood estimation of binary logit extend to the multiple alternatives case. Under relatively weak conditions McFadden (1974) shows that $\mathcal{L}$ in equation (5.70) is globally concave when the V's are linear-in-parameters. Consequently, if a solution to the equation (5.73) exists, it is unique. The maximum likelihood estimator of β is consistent, asymptotically normal, and asymptotically efficient.

The first-order conditions (5.73) can be rewritten as

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{i\in\mathcal{C}_n} y_{in}x_{ink} = \frac{1}{N}\sum_{n=1}^{N}\sum_{i\in\mathcal{C}_n} P_n(i)x_{ink},\ k = 1,\ldots,K. \tag{5.75}$$

This means that the average value of an attribute for the chosen alternatives is equal to the average value predicted by the estimated choice probabilities. In particular, if parameter ℓ is an alternative-specific constant defined for an alternative j, then for each n, $x_{ink} = 0$ for all $i \neq j$, $k \neq \ell$, and $x_{jn\ell} = 1$. Consequently, (5.75) simplifies to

$$\sum_{n=1}^{N} y_{in} = \sum_{n=1}^{N} P_n(i), \tag{5.76}$$

implying that, at the maximum likelihood estimates, the sum of the choice probabilities for alternative i (taken over the sample) equals the number in the sample that chose i. Similarly, when an alternative-specific dummy variable defines some segment of the population, then equation (5.76) applies, the sums being taken only over the relevant group.

The computational methods used for solving the maximization problem are identical to those used in the binary case.

5.6.2 Maximum Likelihood for Grouped Data

There are some instances where each individual is observed repeatedly while the attributes of the alternatives remain constant. For example, rather than observing travelers' mode choices for a single day, we might observe all their daily choices over an entire month. We could solve for the maximum likelihood estimates simply by treating each day for each individual as a separate observation[1]. However, this would be computationally inefficient because it would not exploit the fact that the attributes of the alternatives were the same for groups of observations defined by each individual's trips. A more efficient (in the computational sense) approach is to rewrite the likelihood function. To do this, we define D_{in} = number of times individual n chose alternative i, and $D_n = \sum_{i\in\mathcal{C}_n} D_{in}$.

The likelihood of the sample is then

$$\mathcal{L}^* = \prod_{n=1}^{N} \frac{D_n!}{\prod_{j\in\mathcal{C}_n} D_{jn}!} \prod_{i\in\mathcal{C}_n} P_n(i)^{D_{in}}. \tag{5.77}$$

[1]By doing so, we would assume that the error terms are independent. This assumption is usually not verified, as unobserved factors related to each individual persist other time, generating serial correlation.

Thus the log likelihood function is given by:

$$\mathcal{L} = \sum_{n=1}^{N} \left(\ln D_n! - \sum_{j \in \mathcal{C}_n} \ln D_{jn}! + \sum_{i \in \mathcal{C}_n} D_{in} \ln P_n(i) \right). \tag{5.78}$$

We can then proceed as before by taking first derivatives and solving for the estimate $\hat{\beta}$ for which they are equal to zero.

The procedures of treating each individual's trips as separate and independent observations and using equation (5.78) yield exactly the same results in all respects but one. This exception is that the values of the log likelihood functions differ by a constant that depends on the D_{in}'s and correspond to the first two terms in brackets in equation (5.78):

$$\sum_{n=1}^{N} \left(\ln D_n! - \sum_{j \in \mathcal{C}_n} \ln D_{jn}! \right). \tag{5.79}$$

The difference is usually of no real importance because most statistical tests that use the value of the log likelihood function involve *differences* in log likelihoods evaluated on the exact same data; hence the term in equation (5.79) would cancel. However, it affects values of ρ^2 and $\bar{\rho}^2$. In comparing results of different analyses of repeated choices, the reader should be careful to check whether the reported value of the log likelihood does or does not include the term in equation (5.79).

5.6.3 Weighting or not weighting

5.6.4 Least Squares

As discussed in chapter 4, least squares estimation is rarely used on disaggregate data. However, Berkson's method, extended by Theil (1969), can be applied to estimate the parameters of the logit model. To show this, we note that

$$\ln \left(\frac{P_n(i)}{P_n(J_n)} \right) = V_{in} - V_{J_n n} = \beta^T (x_{in} - x_{J_n n}). \tag{5.80}$$

(See equation (5.22) to prove this.) Thus we can group observations that share common values of $x_{in} - x_{J_n n}$ and use the share of the group choosing alternatives i and J_n as estimates of $P_n(i)$ and $P_n(J_n)$, respectively. If N_{ig} denotes the number of members of group g choosing i, and J_g is the number of alternatives available to members of group g, then we would use the regression

$$\ln \left(\frac{N_{ig}}{N_{J_g g}} \right) = \beta^T (x_{ig} - x_{J_g g}) + \xi_{ig} \tag{5.81}$$

to obtain an estimate of β. Each group would correspond to $(J_g - 1)$ observations in the regression. We have implicitly assumed here that each group consists of individuals with the same choice set. This is not essential, but it corresponds to the case where grouping is most likely to be feasible.

All of our comments on Berkson's method stated in chapter 4 apply to the multiple alternatives case. However, the problems of finding homogeneous groups are compounded by the fact that choice sets can vary across the population (making grouping more difficult), and the number of alternatives is greater. These factors tend to make the number of disaggregate observations needed to form groups considerably larger. Application of the Berkson-Theil method for logit is therefore extremely difficult except when the data base is large or when repeated observations on individuals' choices are available.

5.6.5 Other Estimators

Most other estimators become computationally intractable when applied to large choice sets. To our knowledge, none of the estimators discussed in section 4.B have been actually used in practice for choice with multiple alternatives.

5.7 Example of Estimation Results

The maximum likelihood estimation of the model presented in Section 5.5, and specified in Table 5.2, has been performed. Table 5.3 gives the estimated values for the coefficients, their asymptotic robust standard errors and t ratios, and the various summary statistics defined in chapters 4. The use of these statistics is discussed more in Chapter 6.

Estimation results are summarized in essentially the same way as for binary choice. Table 5.3 is such a summary for maximum likelihood estimation of the airline itinerary choice model. This model is very similar to the one presented in Section 4.3.2. We now consider the full choice set composed of alternatives "non stop", "one stop–same airline" and "one stop–different airline". The specification of the model is discussed in Section 5.5. In terms of explaining the summary table, only a few comments beyond those in chapter 4 are relevant.

A few specific aspects of the particular model summarized in table 5.3 are also worth noting.

The constants associated with alternatives "one stop–same airline" and "one stop–multiple airlines" are negative, meaning that the alternative "non stop" is a priori preferred, (for women who are non frequent leisure trav-

elers, because there is a modification of the constant for other individuals, as explained in the following.) The coefficients associated with the round trip fare is negative, meaning that the higher the fare, the less attractive is the alternative. The same interpretation applies to the elapsed time coefficient. Regarding the leg room, the coefficients are positive, meaning that an increase of the leg room favor the alternative. The effect is slightly more important for women, compared to men. Concerning the gap between the desired and proposed departure or arrival times, the two effects are negative, as expected. All coefficients have the expected signs and the interpretations remain the same as for the binary case.

Socio-economic characteristics have been added to the model. Specific coefficients for alternatives "one stop–same airline" and "one stop–multiple airlines" are used to capture the influence of being a frequent leisure traveler. Both parameters are negative showing that frequent travelers favor more the alternative "non stop" (non stop flight.) The effect is stronger for alternative "one stop–same airline", compared to alternative "one stop–multiple airlines" ($-0.153 > -0.349$.) The constants for frequent leisure travelers become: for "one stop–same airline", $-0.922 - 0.349 = -1.271$, and for "one stop–different airlines" $-1.31 - 0.153 = -1.463$. Finally the influence of gender is captured by the parameters β_{11} and β_{12}. Both are positively estimated, meaning that a priori, men have a tendency to favor alternatives "one stop–same airline" and "one stop–different airlines", compared to women. The effect is stronger for alternative "one stop–multiple airlines" ($0.288 > 0.188$.) For men, the constants become for "one stop–same airline": $-0.922+0.188 = -0.734$, for "one stop–multiple airlines": $-1.31+0.288 = -1.022$. Note that the alternative "non stop" is still a priori preferred by men.

Even if the coefficient estimates have the expected signs, not all the coefficient estimates are significantly different from zero at the usual 5% or 10% levels of significance. Although some researchers use this type of criterion as the sole basis for omitting a variable in later specifications, we do not recommend this as good practice. Many of the reasons for this are discussed in subsequent chapters. It suffices here to note that the inability to reject the hypothesis that some coefficient is zero at a particular significance level does not imply that the hypothesis must be accepted.

We have also estimated a model where the round trip fare is interacted with the income of the respondent, and the perception of the leg room is considered to be different between two groups of alternatives. Regarding the round trip fare, the associated term in the deterministic part of the utilities is

$$\beta_{\text{fare/income}} \times \frac{\text{Round\_trip\_fare}_{\text{in}}}{\text{income}_{\text{n}}} \tag{5.82}$$

where i is the alternative, n is the respondent and $\beta_{\text{fare/income}}$ is a *generic* parameter. Note that $\text{Round\_trip\_fare}_{in}/\text{income}_n$ has no unit. Concerning the leg room, the difference of perception between men and women is still considered, but the perception is supposed to be different between alternative "non stop" and alternatives "one stop–same airline" and "one stop–multiple airlines". The estimation results are presented in Table 5.4.

The estimation results and interpretations are mostly the same than for the previous model (see Table 5.3.) The parameters associated with the leg room are all positive, as expected, meaning that respondents have a tendency to choose an alternative with a large leg room. Regarding the interaction between the round trip fare and the income, the associated parameter is negative (-23.8), as expected. The interacted term reinforces the disutility of the round trip fare (-1.81) for low income, and its influence tends to 0 when the income increases. A significant increase of the final log likelihood is obtained, from -1652.573 to -1640.525, which is a sign that the latter specification of the utility functions is preferred. This is analyzed in more details in Section 6.5.5.

5.8 An example of transportation mode choice

We now present an example of a logit model estimated on revealed preferences data. We model the choice of transportation mode for travelers in Switzerland. The survey was conducted between 2009 and 2010 for CarPostal, the public transport branch of the Swiss Postal Service. The survey covers French and German speaking areas of Switzerland. Questionnaires were sent to people living in rural area by mail. The respondents were asked to register all the trips performed during a specified day. The collected information includes the origin, the destination, the cost, the travel time, the chosen transportation mode and the activity at the destination. Moreover, we collected socio-economic information about the respondents and their households. 1124 completed surveys were collected. For each respondent, cyclic sequences of trips (starting and ending at home) are considered and their main transport mode has been identified. Note that the values for travel time and travel cost were imputed using the websites of the Swiss railways (SBB) `www.sbb.ch` and ViaMichelin `fr.viamichelin.ch`.

We specify a logit model with three alternatives: riding public transportation (PT), driving a car, and using a slow mode such as walking or biking. The specification of the utility functions includes the following coefficients.

β_1 Alternative specific constant (PT).

β_2 Income has been coded as a continuous variable. The income of respondents who reported an income category was set to the average value of the category. The specification in the utility function is piecewise linear (see Section 5.4.1.) β_2 is the coefficient of the variable corresponding to the income range between 4 KCHF and 6 KCHF, that is

$$\max(0, \min(\text{Income}/1000 - 4, 2))),$$

and associated with the alternative PT.

β_3 Coefficient of the variable corresponding to the income range between 8 KCHF and 10 KCHF for alternative PT, that is

$$\max(0, \min(\text{Income}/1000 - 8, 2))).$$

Note that the other coefficients of the piecewise linear specification have been found to be not significantly different from zero, and the corresponding variables have been removed from the specification. A similar specification has been associated with the car alternative (see below.) The net effect of income on the difference of utility functions is illustrated in Figure 5.4 for each pair of alternatives.

β_4 Age has also been coded using a piecewise linear specification. β_4 is the coefficient of the variable corresponding to the age range 0–45 for alternative PT, that is

$$\max(0, \min(\text{Age}, 45))).$$

Note that unreported age was coded -1 in the database. The above specification provides a value of zero for these records, so that the variable does not appear in the specification when age is unknown.

β_5 Coefficient of the variable corresponding to the age range 45–65 for alternative PT, that is

$$\max(0, \min(\text{Age} - 45, 20))).$$

The coefficient of the variable for the category of travelers older than 65 was not significantly different from zero, and has not been included in the specification. A similar specification for the car alternative has been tested, but has been removed. The net effect of age on the utility function of the public transportation alternative is illustrated in Figure 5.5.

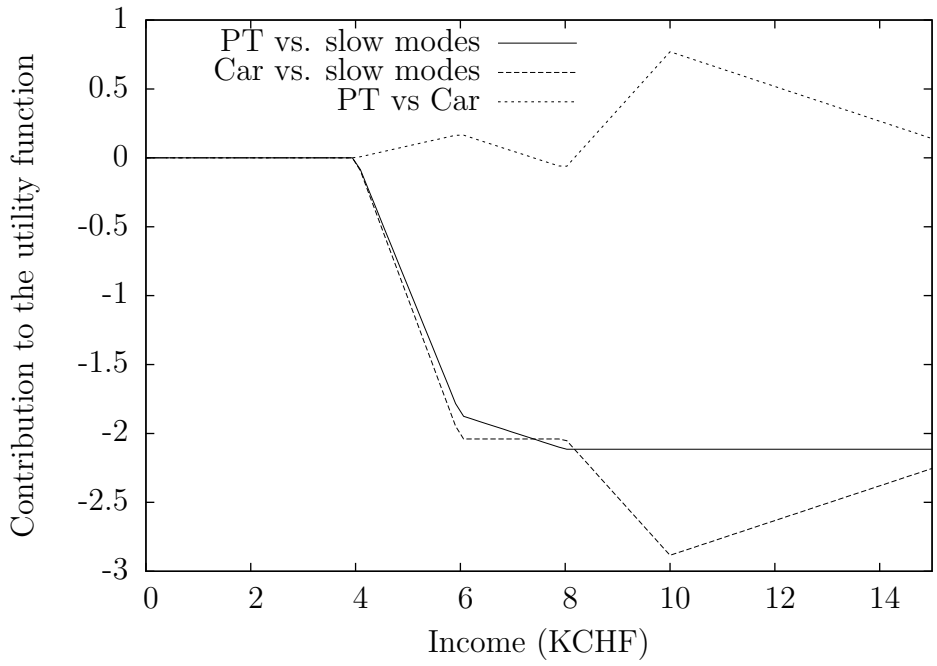


Figure 5.4: Mode choice in Switzerland: piecewise linear specification of income

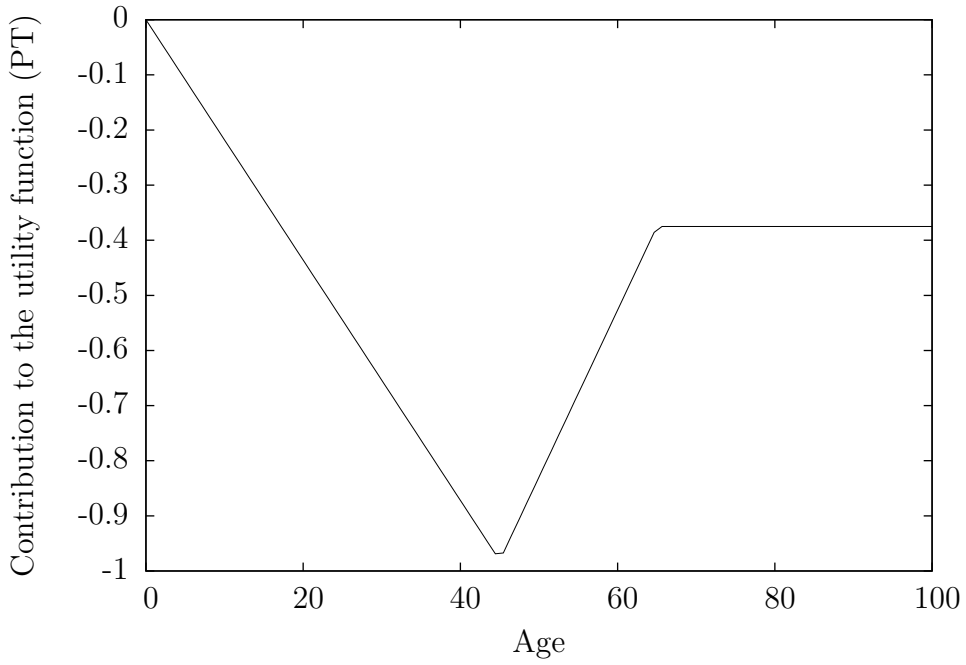


Figure 5.5: Mode choice in Switzerland: piecewise linear specification of age

β_6 Coefficient of the dummy variable representing males (PT).

β_7 Coefficient of the marginal cost (in CHF) of public transportation. It is the price of the trip, excluding the cost of travelcards. In particular, this cost is zero for holders of a season ticket.

β_8 Coefficient of the waiting time (in minutes) for travelers with a full time job (PT).

β_9 Coefficient of the waiting time (in minutes) for travelers without a full time job (part time job or other occupation) (PT).

β_{10} Coefficient of the variable interacting travel time (in minutes) and the log of distance (in kilometers) for travelers with a full time job (PT and car). The variable has been divided by 1000 for numerical reasons. The contribution of travel time to the utility function for various values of distance is illustrated in Figure 5.6.

β_{11} Coefficient of the variable interacting travel time (in minutes) and the log of distance (in kilometers) for travelers with a part time job (PT

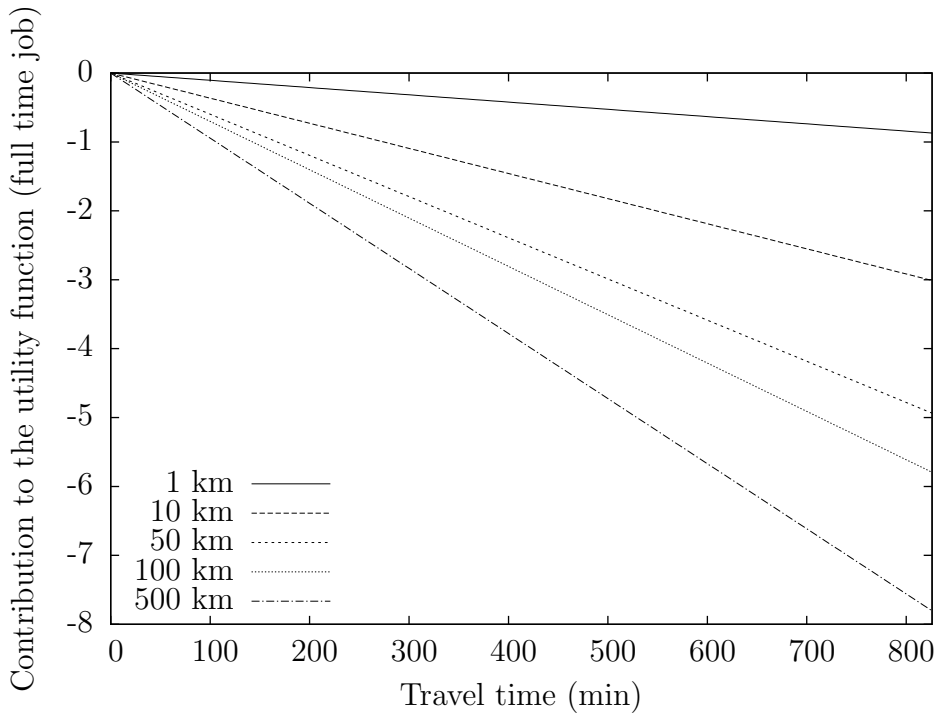


Figure 5.6: Mode choice in Switzerland: interaction between travel time and distance

and car). The variable has been divided by 1000 for numerical reasons. Note that the coefficient for travelers with another occupation has been found not significantly different from zero, and removed from the specification.

β_{12} Coefficient of the dummy variable for holders of a season ticket (PT). Note that these travelers have a marginal cost of zero for public transportation.

β_{13} Coefficient of the dummy variable for holders of a half fare travelcard (PT).

β_{14} Coefficient of the dummy variable for holders of a line related travelcard (PT).

β_{15} Coefficient of the dummy variable for holders of an area related travelcard (PT).

β_{16} Coefficient of the dummy variable for holders of any other travelcard (PT).

β_{17} Alternative specific constant (Car).

β_{18} As discussed above, a piecewise linear specification of income has been associated with the car alternative. β_{18} is the coefficient of the variable corresponding to the income range between 4 KCHF and 6 KCHF for alternative car, that is

$$\max(0, \min(\text{Income}/1000 - 4, 2))).$$

β_{19} Coefficient of the variable corresponding to the income range between 8 KCHF and 10 KCHF for alternative car, that is

$$\max(0, \min(\text{Income}/1000 - 8, 2))).$$

β_{20} Coefficient of the variable corresponding to income above 10 KCHF for alternative car, that is

$$\max(0, \text{Income}/1000 - 10).$$

β_{21} Coefficient of the dummy variable representing males (Car).

β_{22} Coefficient of the number of cars in the household (Car).

β_{23} Coefficient of the cost of the gasoline (in CHF) if the trip purpose of the tour is home–work–home (Car).

β_{24} Coefficient of the cost of the gasoline (in CHF) if the trip purpose of the tour is not home–work–home (Car).

β_{25} Coefficient of the cost of the gasoline (in CHF) if the traveler is male (Car).

β_{26} Coefficient of the dummy variable representing travelers in the French speaking part of Switzerland, where the infrastructure of public transportation is less dense in average (Car).

β_{27} Coefficient of distance (in km) for the slow mode alternative.

The estimation results are reported in Table 5.5. All coefficients have the expected sign. The use of this model for applications is commented in Chapter 10.

5.9 Other Choice Models

Logit is by far the most widely used choice model to date for multiple alternatives, but there are other models that have been developed and applied. These models fall into two distinct classes. The first might be termed logit extensions, in that they are generalizations of logit. The second class is nonlogit-based models. Only one member of this class, probit, has actually been used. In this section we briefly describe models in these two classes. In later chapters we explore some particular special cases in greater detail.

5.9.1 Continuous mixtures: Random Coefficients Logit

In the specification (5.82), the coefficient of the round trip fare varies across the population. Indeed, it is defined as $\beta_{\text{fare/income}}/\text{income}_n$, and is a function of income. Unfortunately, data does not always allow to characterize how coefficients are distributed in the population. When the variation of a coefficient across individuals cannot be explicitly characterized, the coefficient can be assumed to be random and follow a given distribution.

One of the first applications was the analysis of the demand for different types of automobiles in the United States by the Electric Power Research Institute (EPRI) (Electric Power Research Institute, 1977.) But the use of random coefficients models has become more and more important in many

recent applications due to the availability of efficient simulation-based estimation procedures (e.g. Brownstone et al., 2000, Hensher and Greene, 2003, Train, 2003, Bierlaire, 2003, Hess et al., 2005a.)

Thus each individual's coefficients, β_n, differ from the population mean, β, by some unobserved amount. This difference, termed taste variation, constitutes an additional source of randomness. For instance, it may be assumed that the absolute value of coefficients are independent and lognormally distributed, imposing the restriction that the coefficients for all members of the population have the same sign.

If each individual's coefficients were observable, the model could be expressed as follows:

$$P_n(i) = \frac{e^{\beta_n^T x_{in}}}{\sum_{j \in C_n} e^{\beta_n^T x_{jn}}}, \tag{5.83}$$

where $\beta_n=(\beta_{1n}, \beta_{2n}, \ldots, \beta_{Kn})^T$. However, we usually do not observe replications of each individual's choices, so β_n is not estimable. Instead, we use the assumption that each β_{nk} is independent and lognormally distributed with parameters (β_k, σ_k^2), and estimate both β_k and σ_k for all k.

Since β_n is unknown, the left-hand side of equation (5.83) cannot be evaluated. Instead, its expected value, taken across the population, is used. Thus the choice probability is given by

$$P_n(i) = \int_{\beta_{1n}} \cdots \int_{\beta_{Kn}} \frac{e^{\beta_n^T x_{in}}}{\sum_{j \in C_n} e^{\beta_n^T x_{jn}}} f(\beta_{Kn}) \ldots f(\beta_{1n}) d\beta_{Kn} d\beta_{1n}, \tag{5.84}$$

where $f(\beta_{kn})$ is the lognormal density function.

The estimation of this model is quite complex because no closed form solution to the integral in equation (5.84) exists. Monte-Carlo integration is required to approximate the value of $P_n(i)$. Moreover the model now has $2K$ rather than K unknown parameters because each element in β is described by a two-parameter distribution.

These models are discussed at length in Chapter 13.

5.9.2 Discrete mixtures: latent class models

5.9.3 Ordered Logistic

Amemiya (1975) describes a model that applies to ordered discrete alternatives, such as the number of trips taken or the number of automobiles owned by a household. This model represents the choice as the outcome of a sequence of binary decisions, each one consisting of the decision of whether to accept the current value or "take one more." Thus it is not based on the

assumption of global utility maximization. The decision maker stops when the first local optimum is reached. We denote U^c_{in} as the utility of proceeding to the next alternative given that the ith alternative has been reached, U^s_{in} as the utility of stopping at the ith value given that the ith alternative was reached.

The probability individual n accepts the ith alternative (and rejects the $i-1$ first alternatives) is

$$P_n(i) = \Pr(U^c_{1n} \geq U^s_{1n} \cap U^c_{2n} \geq U^s_{2n} \cap \cdots U^c_{(i-1)n} \geq U^s_{(i-1)n} \cap U^c_{in} \leq U^s_{in}).$$

Assuming that the disturbances of the utility differences $U^c_{in} - U^s_{in}$ are independently logistic distributed, this joint event can be written as the product of binary logits:

$$\begin{aligned} P_n(i) &= \left(\frac{e^{V^c_{1n}}}{e^{V^c_{1n}} + e^{V^s_{1n}}}\right)\left(\frac{e^{V^c_{2n}}}{e^{V^c_{2n}} + e^{V^s_{2n}}}\right)\cdots \\ &\quad \left(\frac{e^{V^c_{(i-1)n}}}{e^{V^c_{(i-1)n}} + e^{V^s_{(i-1)n}}}\right)\left(\frac{e^{V^s_{in}}}{e^{V^c_{in}} + e^{V^s_{in}}}\right) \qquad (5.85) \\ &= \left(1 - \frac{1}{1 + e^{-(V^c_{in} - V^s_{in})}}\right)\prod_{j=1}^{i-1}\frac{1}{1 + e^{-(V^c_{jn} - V^s_{jn})}}. \end{aligned}$$

One of the useful features of this structure is that it can be estimated via maximum likelihood using existing logit estimation computer programs. Each binary element in the product in equation (5.85) is simply treated as a separate observation, ignoring the fact that groups of observations come from the same individual's choices. *This is only possible because the disturbances are assumed to be independent at each stage of the decision sequence.*

Hendrickson and Sheffi (1978) (also Sheffi, 1979) and Hall (1980) have applied this model to represent a household's trip generation and search for a residence, respectively. Daly and Van Zwam (1981) have applied this model to predict the frequency of "tours" rather than one-way trips. Lerman and Mahmassani (1985) have explored numerous extensions to account for correlation across disturbances and various sources of observational error. Small (1981) discusses other properties of this and other "ordered response" models.

5.9.4 Probit

We can extend binary probit straightforwardly by assuming that the vector of disturbances $\varepsilon_n = (\varepsilon_{1n}, \varepsilon_{2n}, \ldots, \varepsilon_{J_n n})^T$ is multivariate normal distributed

with a vector of means 0 and a $J_n \times J_n$ variance-covariance matrix Σ_ε. This model, however, requires a solution of a $J_n - 1$ dimensional integral to evaluate the choice probabilities. Hausman and Wise (1978) use a transformation to reduce the dimensionality of the integral to $J_n - 2$.

The concept of probit appeared in mathematical psychology in writings by Thurstone (1927). Due to its computational difficulty, only a few, very limited applications have appeared in the literature (Bolduc, 1999, Stern, 1992, McFadden, 1989, Boersch-Supan and Hajivassiliou, 1993 to cite a few.) The maximum size of the choice set that can be handled without resorting to simulation is 3 or 4. Moreover, there is still no evidence to suggest in which situations the greater generality of probit is worth the additional computational burden resulting from its use.

One useful aspect of probit is the ease with which random taste variation can be incorporated into the model when the systematic component of the utility function is linear in the parameters. To show this, we define the individual's parameters, β_n, as follows:

$$\beta_n = \beta + \psi_n. \tag{5.86}$$

Then we write

$$U_{in} = \beta_n^T x_{in} + \varepsilon_{in} = \beta^T x_{in} + (\psi_n^T x_{in} + \varepsilon_{in}). \tag{5.87}$$

Here ψ_n is a vector of deviations defining the difference between individual n's parameters and the average for the population. By construction, ψ_n is a vector with K entries and mean zero. We define Σ_ψ as the $K \times K$ variance-covariance matrix of ψ_n.

Note that if we assume that ε_n and ψ_n are both normally distributed, then the composite disturbance for each utility is a linear combination of normal variates and is consequently normal. To be more specific, if

$$\varepsilon^*_{in} = \varepsilon_{in} + \psi_n^T x_{in}, \tag{5.88}$$

then the vector $\varepsilon^*_n = (\varepsilon^*_{1n}, \varepsilon^*_{2n}, \ldots, \varepsilon^*_{J_n n})^T$ is multivariate normal with mean 0. If X_n is the $K \times J_n$ matrix with columns $x_{1n}, x_{2n}, \ldots, x_{J_n n}$, then the variance-covariance matrix of ε^*_n is given by $\Sigma_\varepsilon + X_n^T \Sigma_\psi X_n$. Thus the probit model with linear-in-parameters systematic utilities allows for normally distributed taste variations. A more complete treatment of the theory of probit is given in Daganzo (1980).

5.10 Summary

This chapter discussed the analysis of choice problems with multiple alternatives. The general formulation of choice probabilities was described, and a

particular form, logit, was derived. The logit model is based on the following assumptions about the disturbance terms of the utilities:

1. they are independent,
2. identically distributed, and
3. extreme value distributed,

and is expressed as

$$P_n(i) = \frac{e^{\mu V_{in}}}{\sum_{j=1}^{J_n} e^{\mu V_{jn}}}. \tag{5.89}$$

One common form of the logit model has linear-in-parameters systematic utilities V_{in}, but various nonlinear specifications may also be considered.

The *independence from irrelevant alternative (IIA)* property of logit was described. This property holds that for any two alternatives, the ratio of their choice probabilities is independent of the systematic utility of any other alternatives in the choice set. The potential problems that can arise when two alternatives have correlated disturbances were discussed. It was shown that the use of the logit model does not imply that IIA holds at the aggregate level.

The elasticities of logit were also derived. Logit has uniform cross elasticities — that is, the cross elasticity of the choice probability of alternative i with respect to an attribute of alternative j is the same for all alternatives $i \neq j$.

Maximum likelihood estimation of the logit model's parameters was discussed for the cases of a single observation per decision maker and replicated observations. The Berkson-Theil method for least squares estimation was also developed. An example of a logit model was given for mode choice to work. Finally, some choice models other than logit were briefly summarized. The models discussed here are the random coefficients logit, the ordered logistic, and the probit model.

	Non stop flight (1)	One stop flight with the same airline (2)	One stop flight with a change of airline (3)
β_1	0	1	0
β_2	0	0	1
β_3	Round trip Fare ($100) of (1)	Round trip fare ($100) of (2)	Round trip fare ($100) of (3)
β_4	Elasped time (hours) for (1)	Elapsed time (hours) for (2)	Elapsed time (hours) for (3)
β_5	Leg room in (1) (inches), if male	Leg room in (2) (inches), if male	Leg room in (3) (inches), if male
β_6	Leg room in (1) (inches), if female	Leg room in (2) (inches), if female	Leg room in (3) (inches), if female
β_7	Being early (hours) for (1), at departure or arrival, depending on the preference of the respondent	Being early (hours) for (2), at departure or arrival, depending on the preference of the respondent	Being early (hours) for (3), at departure or arrival, depending on the preference of the respondent
β_8	Being late (hours) for (1), at departure or arrival, depending on the preference of the respondent	Being late (hours) for (2), at departure or arrival, depending on the preference of the respondent	Being late (hours) for (3), at departure or arrival, depending on the preference of the respondent
β_9	0	1 if the respondent makes more than two air trips per year	0
β_{10}	0		1 if the respondent makes more than two air trips per year
β_{11}	0	1 if male, 0 otherwise	
β_{12}	0	0	1 if male, 0 otherwise

Table 5.2: Specification table of the model for the choice of airline itinerary

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.922	0.215	-4.28	0.00
2	One stop–multiple airlines dummy	-1.31	0.222	-5.89	0.00
3	Round trip fare ($100)	-2.16	0.103	-20.92	0.00
4	Elapsed time (hours)	-0.302	0.0778	-3.88	0.00
5	Leg room (inches), if male	0.108	0.0233	4.66	0.00
6	Leg room (inches), if female	0.131	0.0219	5.99	0.00
7	Being early (hours)	-0.150	0.0188	-7.97	0.00
8	Being late (hours)	-0.0946	0.0166	-5.70	0.00
9	More than two air trips per year (one stop–same airline)	-0.349	0.138	-2.52	0.01
10	More than two air trips per year (one stop–multiple airlines)	-0.153	0.153	-1.00	0.32
11	Male dummy (one stop–same airline)	0.188	0.125	1.51	0.13
12	Male dummy (one stop–multiple airlines)	0.288	0.132	2.18	0.03

Summary statistics
Number of observations = 2544
$\mathcal{L}(0)$ = -2794.870
$\mathcal{L}(c)$ = -2203.160
$\mathcal{L}(\hat{\beta})$ = -1652.573
$-2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})]$ = 2284.594
ρ^2 = 0.409
$\bar{\rho}^2$ = 0.404

Table 5.3: Estimation results for the airline itinerary choice model: base specification

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.879	0.219	-4.02	0.00
2	One stop–multiple airlines dummy	-1.27	0.227	-5.60	0.00
3	Round trip fare ($100)	-1.81	0.151	-11.99	0.00
4	Elapsed time (hours)	-0.303	0.0778	-3.90	0.00
5	Leg room (inches), if male (non stop)	0.100	0.0330	3.04	0.00
6	Leg room (inches), if female (non stop)	0.182	0.0318	5.71	0.00
7	Leg room (inches), if male (one stop)	0.113	0.0297	3.80	0.00
8	Leg room (inches), if female (one stop)	0.0931	0.0273	3.41	0.00
9	Being early (hours)	-0.151	0.0189	-7.99	0.00
10	Being late (hours)	-0.0975	0.0167	-5.83	0.00
11	More than 2 air trips per year (one stop–same airline)	-0.300	0.141	-2.12	0.03
12	More than 2 air trips per year (one stop–multiple airlines)	-0.0847	0.157	-0.54	0.59
13	Male dummy (one stop–same airline)	0.100	0.133	0.75	0.45
14	Male dummy (one stop–multiple airlines)	0.189	0.144	1.31	0.19
15	Round trip fare / income ($100/$1000)	-23.8	8.09	-2.94	0.00

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\hat{\beta}) &= -1640.525 \\ -2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 2308.689 \\ \rho^2 &= 0.413 \\ \bar{\rho}^2 &= 0.408 \end{aligned}$$

Table 5.4: Specification of the airline itinerary choice model with an interaction between the traveling fare and the income, as well as alternative specific leg room coefficients

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	Cte. (PT)	0.977	0.605	1.61	0.11
2	Income 4-6 KCHF (PT)	-0.934	0.255	-3.67	0.00
3	Income 8-10 KCHF (PT)	-0.123	0.175	-0.70	0.48
4	Age 0-45 (PT)	-0.0218	0.00977	-2.23	0.03
5	Age 45-65 (PT)	0.0303	0.0124	2.44	0.01
6	Male dummy (PT)	-0.351	0.260	-1.35	0.18
7	Marginal cost [CHF] (PT)	-0.0105	0.0104	-1.01	0.31
8	Waiting time [min], if full time job (PT)	-0.0440	0.0117	-3.76	0.00
9	Waiting time [min], if part time job or other occupation (PT)	-0.0268	0.00742	-3.62	0.00
10	Travel time [min] $\times$log(1+ distance[km]) / 1000, if full time job	-1.52	0.510	-2.98	0.00
11	Travel time [min] $\times$log(1+ distance[km]) / 1000, if part time job	-1.14	0.671	-1.69	0.09
12	Season ticket dummy (PT)	2.89	0.346	8.33	0.00
13	Half fare travelcard dummy (PT)	0.360	0.177	2.04	0.04
14	Line related travelcard dummy (PT)	2.11	0.281	7.51	0.00
15	Area related travelcard (PT)	2.78	0.266	10.46	0.00
16	Other travel cards dummy (PT)	1.25	0.303	4.14	0.00
17	Cte. (Car)	0.792	0.512	1.55	0.12
18	Income 4-6 KCHF (Car)	-1.02	0.251	-4.05	0.00
19	Income 8-10 KCHF (Car)	-0.422	0.223	-1.90	0.06
20	Income 10 KCHF and more (Car)	0.126	0.0697	1.81	0.07
21	Male dummy (Car)	0.291	0.229	1.27	0.20
22	Number of cars in household (Car)	0.939	0.135	6.93	0.00
23	Gasoline cost [CHF], if trip purpose HWH (Car)	-0.164	0.0369	-4.45	0.00
24	Gasoline cost [CHF], if trip purpose other (Car)	-0.0727	0.0224	-3.24	0.00
25	Gasoline cost [CHF], if male (Car)	-0.0683	0.0240	-2.84	0.00
26	French speaking (Car)	0.926	0.190	4.88	0.00
27	Distance [km] (Slow modes)	-0.184	0.0473	-3.90	0.00

Summary statistics

Number of observations = 1723

Number of estimated parameters = 27

$$\begin{aligned} \mathcal{L}(\beta_0) &= -1858.039 \\ \mathcal{L}(\hat{\beta}) &= -792.931 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 2130.215 \\ \rho^2 &= 0.573 \\ \bar{\rho}^2 &= 0.559 \end{aligned}$$

Table 5.5: Specification of the transportation mode choice model

Chapter 6

Specification testing

Contents

6.1 Introduction

In the previous chapters we introduced the theory and the methods of deriving, specifying and estimating discrete choice models. In this chapter we consider the process of empirical model development. This process does not follow a definitive set of rules. It employs rigorous statistical methods but also requires many intuitive, model-building judgments. We emphasize in this chapter both the formal and informal tests that were found in many empirical studies to be the most useful aids in developing and evaluating discrete choice models.

The next section of this chapter is a discussion of the philosophy that underlies our model-building approach. In section 6.4 we present the discrete choice example that will be used throughout most of the remainder of this chapter. Section 6.5 includes a demonstration of the use of the most basic tests used in the model development process, including the t test and the likelihhood ratio test. Section 6.6 describes tests of the basic structural assumptions of discrete choice models. We focus on the assumptions inherent in the logit model. Section 6.7 describes informal test procedures that examine the predicted choice probabilities.

6.2 Background on hypothesis testing

6.3 The Art of Model Building

In general, it is impossible to determine the most appropriate specification of a model from data analysis. A "good fit" to the data does not necessarily mean an adequate model, and it is not unusual to find several alternative model specifications that fit the data equally well. Moreover a model can duplicate the data perfectly but give erroneous predictions.

Since we cannot rely exclusively on the "goodness-of-fit" criterion, we require additional procedures that will assist in determining the specification of a model. The most important roles in these procedures are played by formal theories and informal judgments that represent our best a priori knowledge of

the phenomenon being modeled. This knowledge is reflected in a set of specific assumptions about the relationships between variables. Unfortunately we do not normally have either a comprehensive understanding of a situation or a general behavioral theory that will prescribe the specification of the exact mathematical form and variables of a model.

Furthermore, even if the correct specifications of models were known, they would be empirically infeasible. There exist practical difficulties that almost always prevent us from implementing these theoretical specifications. These difficulties stem primarily from data problems: important variables must often be omitted due to lack of data or measurement problems, and included variables are often measured with errors. It is also necessary in practical studies to use a mathematical form that is computationally feasible. Thus lack of definitive theories and practical data and computational problems imply that a priori considerations virtually never lead to a unique model being selected.

Statistical tests cannot be used as the only criteria for acceptance or rejection of a model. With classical statistical inference methods, the specification of the model is never in question; the model is virtually always assumed to be correct. The purpose of a statistical estimator or of an hypothesis test is to make inferences about the values of unknown parameters of the known mathematical functions. In other words, statistical inference cannot be conclusive as to which of a large number of alternative specifications represents the true underlying process that generated the observed data.

These observations explain why the development of a model specification is not a simple, algorithmic process with clear-cut rules. It is a mixture of applications of formal behavioral theories and statistical methods with subjective judgments of the model builder. We generally begin the estimation process with an a priori theory, or set of assumptions, that is consistent with a very large number of model specifications. Then we begin a learning process that consists of a sequence of model estimations and a variety of formal and informal tests designed to help us narrow down the range of alternative specifications. At various stages of this process we may revise some aspects of our a priori assumptions that do not agree sufficiently with the statistical findings. We may discard some assumptions and devise new ones. Finally, among the specifications that are consistent with the theory, we select the one that performs best according to "goodness-of-fit" measures and statistical significance tests.

6.4 The example of the airline itinerary choice

The example of the airline itinerary choice that was introduced in Section 4.3.2 is used in this chapter to demonstrate the model specification tests.

The maximum likelihood estimation results of the model that is used as the base specification for the tests are given in Table 5.4. The table gives the estimated values of the coefficients, their asymptotic standard errors and t-ratios, and the various summary statistics defined in Chapters 4 and 5. The use of these statistics is discussed in the subsequent sections of this chapter.

6.5 Tests of Alternative Specifications of Variables

In this section we describe the most basic tests used in the model development process. We assume a given structure of the discrete choice model and test alternative specifications of the explanatory variables in the utility functions. The statistical tests that are used are the asymptotic t-test and the likelihood ratio tests. The tests of alternative model specifications are demonstrated and include comparisons of goodness-of-fit measures and tests of nonlinear utilities.

6.5.1 Informal Tests

The most basic test of the model estimation output is the examination of the values of the coefficient estimates. Usually we have a priori expectations with respect to the signs and relative values of coefficients. In the estimation results given in Table 5.4, all the coefficient estimates have the expected signs and expected relative values. For example, $\widehat{\beta}_3$, the estimate of the coefficient of the round trip fare, and $\widehat{\beta}_4$, the coefficient associated with the elapsed time, have the expected negative sign.

Another useful test consists in computing the trade-offs. A trade-off is the quantity that one variable should vary in order to compensate a small variation of another variable. In travel demand analysis, the variables "travel time" and "travel cost" are often playing an important role in the utility function. The trade-off corresponding to these two variables is called the *value of time*, and represents how much travelers would be willing to pay to decrease their travel time by one unit.

Let x_k and x_ℓ be the two variables of interest appearing in the utility

function of alternative i. Then, the corresponding trade-off is defined as

$$\frac{\partial U_{in}/\partial x_k}{\partial U_{in}/\partial x_\ell}, \tag{6.1}$$

and its unit is the unit of x_ℓ divided by the unit of x_k. For the value of time, x_k is the travel time, and x_ℓ is the travel cost.

When the utility function is linear in parameters, the trade-off simplifies to the ratio of the two corresponding coefficients in the utility function. For example, if we are interested in the trade-off between the two components of elapsed time and the round trip fare in the model presented in Table 5.4, we consider the utility function of a given alternative i:

$$U_{in} = \cdots + \beta_3 \text{round trip fare} + \beta_4 \text{elapsed time} + \cdots + \beta_{15} \frac{\text{round trip fare}}{\text{income}} + \varepsilon_{in}, \tag{6.2}$$

and compute the trade-off

$$\frac{\partial U_{in}/\partial \text{elapsed time}}{\partial U_{in}/\partial \text{round trip fare}} = \frac{\widehat{\beta}_4}{\widehat{\beta}_3 + \frac{\widehat{\beta}_{15}}{\text{income}}} \frac{\$100}{\text{hour}}, \forall i \in \mathcal{C}_n. \tag{6.3}$$

Note that all parameters are generic, so that the trade-off is the same for any alternative. Also, the unit of the trade-off depends on the units of the variables. Here the round trip fare is given in hundreds of dollars, and the elapsed time in hours, so that the trade-off is obtained in \$100 per hour.

When evaluating this expression at the sample average income of 107 (in \$1000 per year), we obtain

$$\frac{-0.303}{-1.81 + (\frac{-23.8}{107})} = 0.149 \ \$100/\text{hour} = \$14.9/\text{hour}, \forall i \in \mathcal{C}_n. \tag{6.4}$$

The interpretation is that travelers with this level of income are willing to pay \$14.91 to save one hour of elapsed time; or, symmetrically, that one additional hour of elapsed time would be compensated by a decrease of the round trip fare of \$14.91.

In addition to evaluating how well coefficients reflect our own a priori expectations, it is also desirable to compare them with analogous values from similar models calibrated for other places, times, and even other choice contexts. For example, a previous study may provide an estimate of the trade-off that can be compared with the estimate provided by the model in question. When utilizing other sources of information, one must ascertain that the two measures are in reality comparable.

6.5.2 The Use of the Asymptotic *t* Test

The asymptotic t test is used primarily to test whether a *particular* parameter in the model differs from some known constant, often zero. It is used in the same way as the t test in linear regression, except that in the case of nonlinear models this test is valid only asymptotically —that is, it is valid only for large samples. The value of the t statistic is

$$\frac{\widehat{\beta} - \beta_0}{\sigma} \tag{6.5}$$

where $\widehat{\beta}$ is the estimated value of the parameter, σ is its associated standard error, and β_0 is the reference value (usually 0 for a coefficient).

The critical values for the test statistic are percentiles of the standardized normal distribution, which for two-tailed tests at the frequently used significance levels of 0.10 and 0.05 are ± 1.65 and ± 1.96, respectively. Sometimes, the p-value is reported in estimation results, conveying the same information in another format. The p-value is the probability to get a t statistic at least as large (in absolute value) as the one reported, under the null hypothesis. In practice, the null hypothesis is rejected when the p-value is lower than the significance level (typically 0.05).

In the airline itinerary choice example, assume that we want to test the hypothesis that the fact of being early does not play a role in the choice. Equivalently, we would test the hypothesis that $\beta_9 = 0$. The t statistic given in Table 5.4 for this coefficient is -7.99. We can therefore safely reject the null hypothesis. The fact that the p-value is 0 leads to the same conclusion. We say that the estimate of the coefficient β_9 is *significant* at the 5% level. Actually, all other parameters, except β_{12}, β_{13} and β_{14} are significant at the 5% level. Note that each test focuses on one parameter. The t test cannot be used to test that several parameters are simultaneously zero, which requires an alternate test procedure described later on in this section. In particular, if a insignificant parameter is removed from the specification and the model is re-estimated, another parameter that was not significant may become significant. It is therefore not appropriate to remove several variables from the model based only on the t statistics.

The t test can also be used to test the null hypothesis that two parameters β_k and β_ℓ are equal. The value of the t statistic is

$$\frac{\widehat{\beta}_k - \widehat{\beta}_\ell}{\sqrt{\mathrm{Var}(\widehat{\beta}_k - \widehat{\beta}_\ell)}} = \frac{\widehat{\beta}_k - \widehat{\beta}_\ell}{\sqrt{\mathrm{Var}(\widehat{\beta}_k) + \mathrm{Var}(\widehat{\beta}_\ell) - 2\,\mathrm{Cov}(\widehat{\beta}_k, \widehat{\beta}_\ell)}}. \tag{6.6}$$

Before continuing the discussion, we present two specifications that we

may want to consider and test. The first model is a version of the base model presented in Table 5.4, where the leg room coefficients are generic. The estimation results are presented in Table 6.1. The second model is an extension of the first, where the elapsed time coefficient is now alternative specific.

We illustrate on the second model how to compute the statistic for an asymptotic t-test of a linear relationship between two parameters. More specifically, we would like to test if it is appropriate to use alternative specific elapsed time coefficients. We can perform three tests: $\beta_3 = \beta_4$, $\beta_3 = \beta_5$ and $\beta_4 = \beta_5$.

We use the asymptotic covariance matrix of the three parameters of interest:

	β_3	β_4	β_5
β_3	0.00729	0.00627	0.006
β_4	0.00627	0.00676	0.00553
β_5	0.006	0.00553	0.00643

We calculate the estimated variance of the difference $(\widehat{\beta}_3 - \widehat{\beta}_4)$:

$$\begin{aligned} \mathrm{Var}(\widehat{\beta}_3 - \widehat{\beta}_4) &= \mathrm{Var}(\widehat{\beta}_3) + \mathrm{Var}(\widehat{\beta}_4) - 2\,\mathrm{Cov}(\widehat{\beta}_3, \widehat{\beta}_4) \\ &= 0.00729 + 0.00676 - 2 \times 0.00627 = 0.00151 \end{aligned}$$

The test statistic for the null hypothesis $\beta_3 = \beta_4$ is then given by

$$\frac{\widehat{\beta}_3 - \widehat{\beta}_4}{\sqrt{\mathrm{Var}(\widehat{\beta}_3 - \widehat{\beta}_4)}} = \frac{-0.341 - (-0.291)}{\sqrt{0.00151}} = -1.287 \tag{6.7}$$

We cannot reject the null hypothesis that $\beta_3 = \beta_4$ at the 5% level of significance.

Similar calculations for the null hypothesis $\beta_4 = \beta_5$ show that we cannot reject the equality of these two elapsed time coefficients. The t statistic is equal to:

$$\frac{-0.291 - (-0.310)}{\sqrt{0.00676 + 0.00643 - 2 \times 0.00553}} = \frac{0.019}{0.0462} = 0.412.$$

Similarly, the null hypothesis $\beta_3 = \beta_5$ cannot be rejected at the 5% level:

$$\frac{-0.341 - (-0.310)}{\sqrt{0.00729 + 0.00643 - 2 \times 0.006}} = \frac{-0.031}{0.04153} = -0.746.$$

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.964	0.216	-4.47	0.00
2	One stop–multiple airlines dummy	-1.36	0.224	-6.09	0.00
3	Elapsed time (hours)	-0.301	0.0778	-3.87	0.00
4	Round trip fare ($100)	-1.80	0.150	-11.97	0.00
5	Leg room (inches), if female	0.132	0.0220	6.00	0.00
6	Leg room (inches), if male	0.107	0.0232	4.62	0.00
7	Being early (hours)	-0.151	0.0188	-8.04	0.00
8	Being late (hours)	-0.0958	0.0167	-5.74	0.00
9	More than 2 air trips per year (one stop–same airline)	-0.309	0.141	-2.20	0.03
10	More than 2 air trips per year (one stop–multiple airlines)	-0.0931	0.157	-0.59	0.55
11	Male dummy (one stop–same airline)	0.201	0.125	1.60	0.11
12	Male dummy (one stop–multiple airlines)	0.294	0.132	2.23	0.03
13	Round trip fare / income ($100/$1000)	-24.1	8.07	-2.98	0.00

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\widehat{\beta}) &= -1642.796 \\ -2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})] &= 2304.148 \\ \rho^2 &= 0.412 \\ \bar{\rho}^2 &= 0.408 \end{aligned}$$

Table 6.1: Estimation results for the airline itinerary choice: generic leg room coefficients

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-1.17	0.278	-4.19	0.00
2	One stop–multiple airlines dummy	-1.45	0.292	-4.98	0.00
3	Elapsed time (hours) (non stop)	-0.341	0.0854	-3.99	0.00
4	Elapsed time (hours) (one stop–same airline)	-0.291	0.0822	-3.54	0.00
5	Elapsed time (hours) (one stop–multiple airlines)	-0.310	0.0802	-3.87	0.00
6	Round trip fare ($100)	-1.78	0.151	-11.84	0.00
7	Leg room (inches), if male	0.108	0.0232	4.65	0.00
8	Leg room (inches), if female	0.132	0.0221	5.99	0.00
9	Being early (hours)	-0.151	0.0188	-8.02	0.00
10	Being late (hours)	-0.0960	0.0167	-5.73	0.00
11	More than 2 air trips per year (one stop–same airline)	-0.307	0.141	-2.18	0.03
12	More than 2 air trips per year (one stop–multiple airlines)	-0.0910	0.157	-0.58	0.56
13	Male dummy (one stop–same airline)	0.199	0.126	1.59	0.11
14	Male dummy (one stop–multiple airlines)	0.293	0.132	2.21	0.03
15	Round trip fare / income ($100/$1000)	-24.0	8.09	-2.97	0.00

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\widehat{\beta}) &= -1641.932 \\ -2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})] &= 2305.875 \\ \rho^2 &= 0.413 \\ \bar{\rho}^2 &= 0.407 \end{aligned}$$

Table 6.2: Estimation results for the airline itinerary choice: alternative specific elapsed time coefficients

We now consider the parameters associated with the leg room variable. We test if it is worth to differentiate this perception between the alternative "Non stop" and the alternatives "One stop, same airline" and "One stop, multiple airlines". For the model of Table 5.4, it corresponds to test the null hypothesis $\beta_5 = \beta_7$ and $\beta_6 = \beta_8$. The robust estimate of the covariance matrix for β_5 and β_7 is

	β_5	β_7
β_5	0.00110	0.000107
β_7	0.000107	0.000882

The t statistic is equal to: $(0.100-0.113)/\sqrt{0.00110 + 0.000882 - 2 \times 0.000107} = -0.013/0.250 = -0.310$. We cannot reject the null hypothesis that $\beta_5 = \beta_7$. For β_6 and β_8, the covariance matrix of the robust estimates is:

	β_6	β_8
β_6	0.00101	0.000111
β_8	0.000111	0.000745

The t statistic is equal to: $(0.182-0.0931)/\sqrt{0.00101 + 0.000745 - 2 \times 0.000111} = 0.0889/0.243 = 2.269$. We can reject the null hypothesis $\beta_6 = \beta_8$. It seems appropriate to keep the model where the coefficients of the leg room variable are alternative specific.

Related to the use of the t statistic for hypothesis testing is its use to calculate an asymptotic confidence interval for a single parameter. As the maximum likelihood estimates are (asymptotically) normally distributed, we have that

$$\Pr\left[-t_{\alpha/2} \leq \frac{\widehat{\beta}_k - \beta_k}{\sqrt{\mathrm{Var}(\widehat{\beta}_k)}} \leq t_{\alpha/2}\right] = 1 - \alpha. \tag{6.8}$$

where, $t_{\alpha/2}$ is the quantile of the normal distribution such that the probability is $\alpha/2$ that the t ratio will exceed $t_{\alpha/2}$. Rearranging terms, we obtain the definition of the confidence interval:

$$\Pr\left[\widehat{\beta}_k - t_{\alpha/2}\sqrt{\mathrm{Var}(\widehat{\beta}_k)} \leq \beta_k \leq \widehat{\beta}_k + t_{\alpha/2}\sqrt{\mathrm{Var}(\widehat{\beta}_k)}\right] = 1 - \alpha,$$

For example, the 95% interval ($t_{0.025} = 1.96$) for the elapsed time coefficient in the model of Table 5.4 is given by

$$\Pr\left[-0.455 \leq \beta_4 \leq -0.151\right] = 0.95. \tag{6.9}$$

6.5.3 Confidence Region for Several Parameters Simultaneously

It is possible to construct confidence regions for two or more parameters jointly. However, this is computationally more difficult and is rarely done in practice. We demonstrate its use in the two-dimensional case. (This subsection can be skipped by readers unfamiliar with matrix operations.)

The vector of estimated coefficients $\widehat{\beta}$, found by the method of maximum likelihood, is asymptotically normally distributed with expectation β and variance-covariance matrix $\Sigma_{\widehat{\beta}}$. Therefore the quadratic form

$$(\widehat{\beta} - \beta)' \Sigma_{\widehat{\beta}}^{-1} (\widehat{\beta} - \beta)$$

is asymptotically χ^2 distributed with K degrees of freedom, K being the dimension of β. This is also true for any subvector of $\widehat{\beta}$ with its corresponding submatrix of $\Sigma_{\widehat{\beta}}$. The $(1 - \alpha)$ confidence region for the vector β is of the form

$$\Pr\left[(\widehat{\beta} - \beta)' \Sigma_{\widehat{\beta}}^{-1} (\widehat{\beta} - \beta) \leq \chi^2_{K,\alpha}\right] = 1 - \alpha, \tag{6.10}$$

where $\chi^2_{K,\alpha}$ is the percentile of the χ^2 distribution with K degrees of freedom for the α level of significance. For a subvector of two coefficients the confidence region has the form of an ellipse:

$$\Pr\left[a_{kk}(\widehat{\beta}_k - \beta_k)^2 + a_{hh}(\widehat{\beta}_h - \beta_h)^2 + 2a_{kh}(\widehat{\beta}_k - \beta_k)(\widehat{\beta}_h - \beta_h) \leq \chi^2_{2,\alpha}\right] = 1 - \alpha, \tag{6.11}$$

where a_{kh} is the (k, h) element of $\Sigma^{-1}_{(\widehat{\beta}_k, \widehat{\beta}_h)}$:

$$\Sigma^{-1}_{(\widehat{\beta}_k, \widehat{\beta}_h)} = \begin{bmatrix} \mathrm{Var}(\widehat{\beta}_k) & \mathrm{Cov}(\widehat{\beta}_k, \widehat{\beta}_h) \\ \mathrm{Cov}(\widehat{\beta}_k, \widehat{\beta}_h) & \mathrm{Var}(\widehat{\beta}_h) \end{bmatrix}^{-1}. \tag{6.12}$$

Consider again the example in Table 5.4 and the round trip fare and the elapsed time coefficients. The estimated values of the coefficients and the covariance matrix are

$$\begin{bmatrix} \widehat{\beta}_3 \\ \widehat{\beta}_{15} \end{bmatrix} = \begin{bmatrix} -1.81 \\ -23.8 \end{bmatrix} \tag{6.13}$$

and

$$\Sigma_{(\widehat{\beta}_3, \widehat{\beta}_{15})} = \begin{bmatrix} 0.0228 & -0.897 \\ -0.897 & 65.4 \end{bmatrix},$$

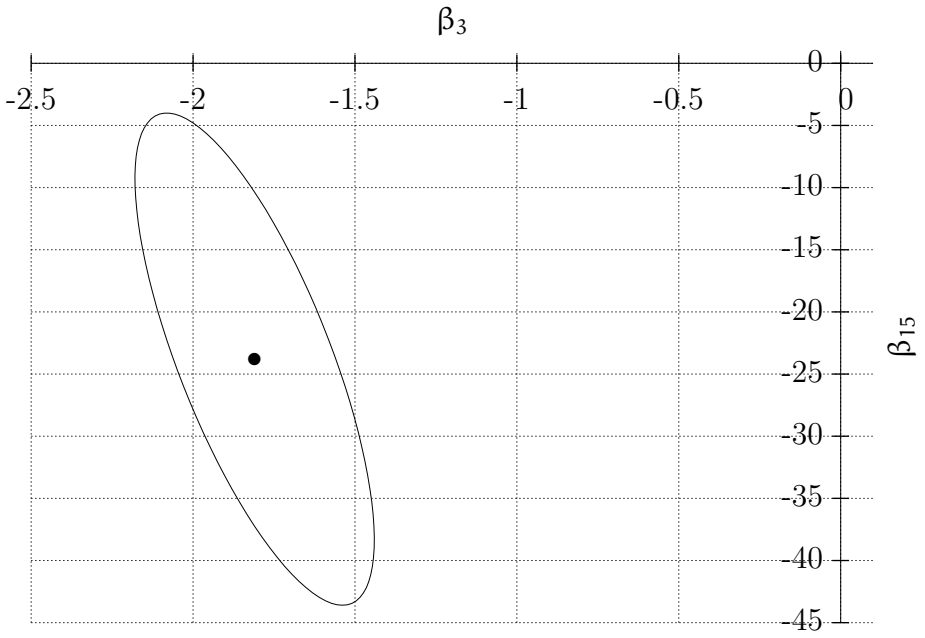


Figure 6.1: Asymptotic confidence region for two coefficients

with inverse:

$$\Sigma^{-1}_{(\widehat{\beta}_3,\widehat{\beta}_{15})} = \begin{bmatrix} 95.2 & 1.30 \\ 1.30 & 0.0332 \end{bmatrix}.$$

The 95% confidence region for these two coefficients is the solution to

$$\Pr\begin{bmatrix} 95.2(-1.81-\beta_3)^2 + 65.4(-23.8-\beta_{15})^2 + \\ 2.6(-1.81-\beta_3)(-23.8-\beta_{15}) \leq 5.99 \end{bmatrix} = 0.95. \tag{6.14}$$

It has the form of the ellipse shown in Figure 6.1 with center at $[-1.81, -23.8]$. Note that all the area of the confidence region is in the negative quadrant, which is in agreement with our expectation of negative coefficients.

6.5.4 The Use of Goodness-of-Fit Measures

The first model estimation outputs to be examined are the signs and relative values of the coefficients estimates and the significance of individual coefficients. With the estimation of more than one specification it is also useful to compare goodness-of-fit measures. Everything else being equal, a specification with a higher maximum value of the likelihood function is considered to

be better. It is more convenient to compare the value of the likelihood ratio index (rho-squared)

$$\rho^2 = 1 - \frac{\mathcal{L}(\widehat{\beta})}{\mathcal{L}(0)}, \tag{6.15}$$

which is interpreted in a fashion similar to R^2 in regression analysis. There are no general guidelines for when a ρ^2 value is sufficiently high.

For the same estimation data set, the ρ^2 of a model always increases or at least stays the same whenever new variables are added to the model, a limitation it shares with regression statistic R^2. For this reason we also use the adjusted likelihood ratio index (rho bar squared):

$$\bar{\rho}^2 = 1 - \frac{\mathcal{L}(\widehat{\beta}) - K}{\mathcal{L}(0)}. \tag{6.16}$$

where K denotes the number of unknown parameters in the model.

This measure is based on the idea of estimating the expectation of the sample log likelihood for the estimated parameter values over all samples with the log likelihood of the one sample we do have available. $\mathcal{L}(\widehat{\beta})$ is a biased estimate of the expectation over all samples; it is necessary to subtract K from $\mathcal{L}(\widehat{\beta})$ — to compensate for the fact that $\widehat{\beta}$ will not be the MLE in other samples—and to remove the effect of evaluating $\mathcal{L}(\widehat{\beta})$ at the estimated values rather than for the true parameters. The measure $[\mathcal{L}(\widehat{\beta}) - K]$ is known as the Akaike information criterion (see Akaike, 1973, Amemiya, 1980).

This $\bar{\rho}^2$ is quite similar to the one proposed by Horowitz (1982, 1983a). In Horowitz's measure $\mathcal{L}(\widehat{\beta})$ is corrected only by subtracting $K/2$. The principal practical difference between the two measures is that the correction factor of K will favor more parsimonious model specifications, unless the added explanatory power of the variable is quite significant. Later in the chapter, we show how to use $\bar{\rho}^2$ for testing a certain class of hypotheses.

Table 5.3 gives the estimation results for a restricted model for the airline itinerary choice problem. Compared to the unrestricted model (Table 5.4), the leg room coefficients are now generic, and the interaction term between fare and income has been removed. The three restrictions are therefore: $\beta_5 = \beta_7$, $\beta_6 = \beta_8$ and $\beta_{15} = 0$.

The ρ^2 are 0.409 and 0.413 for the restricted and the unrestricted models, respectively. The respective $\bar{\rho}^2$ are 0.404 and 0.408, that is a difference of 0.004. Thus, the better fit of the unrestricted specification seems to be associated with a sufficient explanatory power to compensate for the additional degrees of freedom utilized. It is possible to reach the same conclusion from lilelihood ratio test presented in the following section.

6.5.5 The Use of the Likelihood Ratio Test

The t test described in Section 6.5.2 focuses on testing individual parameters, or pair of them. We now introduce the *likelihood ratio test* which compares different specifications. The general motivation is to investigate parsimonious versions of a given specification, by introducing linear restrictions on the parameters. The null hypothesis of the test is that the parsimonious, or restricted, model is the true model. If it is rejected, the unrestricted model is preferred.

For instance, we may consider a restricted model where all coefficients are zero. In this case, the utility function contains only the error term, and the choice probability is $1/J_n$ for each alternative and each decision maker n.

Under the null hypothesis that all the coefficients are zero, that is,

$$\beta_1 = \beta_2 = ... = \beta_K = 0, \tag{6.17}$$

the statistic

$$-2(\mathcal{L}(0) - \mathcal{L}(\widehat{\beta}))$$

where

$$\mathcal{L}(0) = -\sum_{n=1}^{N} \log(J_n) \tag{6.18}$$

is χ^2 distributed with K degrees of freedom (see D.1). This statistic is given as one of the summary measures in the estimation results tables. However, it is not a very useful test because almost always we can reject this null hypothesis at a low level of significance. For example, the percentile of the χ^2 distribution with 15 degrees of freedom for a 0.005 level of significance is 23.68, and the value of the test statistic for the base specification in Table 5.4 is 2308.689.

It is more informative to test the null hypothesis that all the coefficients, except for the alternative-specific constants, are zero. In this case the test statistic is

$$-2(\mathcal{L}(c) - \mathcal{L}(\widehat{\beta}))$$

with $K - J + 1$ degrees of freedom, where J is the number of alternatives in the universal choice set and $\mathcal{L}(c)$ is the log likelihood of a model with only constants. $\mathcal{L}(c)$ can be obtained by either estimating a model with $J - 1$ alternative-specific constants or, if all observations have J available alternatives, by

$$\mathcal{L}(c) = \sum_{i=1}^{J} N_i \ln(\frac{N_i}{N}) \tag{6.19}$$

where N_i is the number of observations selecting alternative i and N is the total sample size. In our case, the value reported in Table 5.4 was obtained from 6.19 where $N_1 = 1698$, $N_2 = 445$, $N_3 = 401$ and $N = 2544$. In Table 5.4 the χ^2 test statistic with 13 degrees of freedom for the null hypothesis

$$\beta_3 = \beta_4 = ... = \beta_{15} = 0, \tag{6.20}$$

is

$$-2(-2203.160 + 1640.525) = 1125.270$$

This hypothesis can also be rejected with high confidence.

In general, the test statistic is

$$-2(\mathcal{L}(\widehat{\beta}_R) - \mathcal{L}(\widehat{\beta}_U))$$

where $\widehat{\beta}_R$ denotes the estimated coefficients of the restricted model — the model that is true under the null hypothesis — and $\widehat{\beta}_U$ denotes the coefficient estimates of the unrestricted model. This statistic is χ^2 distributed with $(K_U - K_R)$ degrees of freedom, where K_U and K_R are the numbers of estimated coefficients in the unrestricted and restricted models, respectively.

This test is used when the model under the null hypothesis is obtained by imposing linear restrictions on the more general unrestricted model. It is sometimes called a "nested" hypothesis[1]. The example would be the null hypothesis that two (or more) coefficients are jointly equal to zero or that three (or more) coefficients are equal to each other. For a null hypothesis with a single linear restriction this test serves the same purpose as the asymptotic t test. However, the χ^2 test is performed by comparing the results from two estimation runs. We compare the model of Table 5.4 with a restricted model where the coefficient for the leg room variable is generic, and the interaction between the round trip fare and the income is not accounted for. The linear restrictions are $\beta_5 = \beta_7$, $\beta_6 = \beta_8$ and $\beta_{15} = 0$. Estimation results of this restricted model are presented in Table 5.3.

The χ^2 test statistic with 3 degrees of freedom is

$$-2(-1652.573 + 1640.525+) = 24.096. \tag{6.21}$$

We can reject the null hypothesis at a 5% level of significance ($\chi^2_{3,0.05} = 7.81$). The unrestricted model (see Table 5.4) is preferred.

[1]It is unrelated to the nested logit model introduced in Chapter 7

6.5.6 Test of Generic Attributes

An important aspect of the specification of discrete choice models is the distinction between alternative-specific and generic attributes. A generic specification imposes restrictions of equality of coefficients on a more general model with alternative-specific attributes. Thus, the likelihood ratio test statistic for the null hypothesis of generic attributes is

$$-2(\mathcal{L}(\widehat{\beta}_{G}) - \mathcal{L}(\widehat{\beta}_{AS})),$$

where G and AS denote the generic and the alternative specific models, respectively. It is χ^2 distributed with the number of degrees of freedom equal to the number of restrictions, or $(K_{AS} - K_{G})$.

For the example of the airline itinerary choice, the estimation results of a specification with alternative specific elapsed time coefficients are given in Table 6.2. The restricted model with the generic specification is base specification in Table 6.1. The value of the test statistic with 2 degrees of freedom is

$$-2(-1642.796 + 1641.932) = 1.728. \tag{6.22}$$

We cannot reject the null hypothesis of the generic specification at the 0.05 significance level because $\chi^2_{2,0.05} = 5.99$.

We could also consider comparing the unrestricted model of Table 5.4 with a version with equality restrictions among the leg room coefficients. Table 6.1 corresponds to such a model. It brings down the number of elapsed time coefficients from 4 to 2 which implies 2 restrictions. The value of the test statistic is

$$-2(-1642.796 + 1640.525) = 4.542. \tag{6.23}$$

We cannot reject the null hypothesis associated with the restricted specification at the 0.05 significance level because $\chi^2_{2,0.05} = 5.99$.

6.5.7 Tests of Non-Nested Hypotheses

The classical statistical hypothesis tests that were presented in Sections 6.5.5 and 6.5.6 can only be applied with what are called *nested hypotheses*. These tests are always expressed as a comparison between restricted and unrestricted models, where the restricted model forms the null hypothesis. The restrictions are placed on the values of the parameters in such a way that the model under the null hypothesis can be obtained as a special case of the unrestricted model. The models that were rejected or not rejected up to this point were always special cases of a more general model.

There are instances where we wish to compare two models, and one is not a nested hypothesis of the other. Consider, for example, two models: the first model has its estimation results shown in Table 5.4; the second model is similar to the first, but it accounts for the square of both the scheduled early delay and the scheduled late delay, (instead of accounting for them linearly in the first model). The estimation results of this second model are presented in Table 6.3. Yet the second model cannot be obtained as a special case of the first one using linear restrictions. Also, the first model can not be obtained as a special case of the second model. Therefore,the likelihood ratio test is not applicable here.

The Adjusted Likelihood Ratio Index

The adjusted likelihood ratio index $\bar{\rho}^2$ presented earlier as a goodness-of-fit measure can be used for testing non-nested hypotheses of discrete choice models. Under the null hypothesis that model 1 is the true specification, compared to model 2, the following holds asymptotically:

$$\Pr(\bar{\rho}_2^2 - \bar{\rho}_1^2 > z) \leq \Phi\{-[-2z\mathcal{L}(0) + (K_1 - K_2)]^{\frac{1}{2}}\}, \qquad z > 0, \tag{6.24}$$

where
$\bar{\rho}_\ell^2$ = the adjusted likelihood ratio index for model $\ell = 1, 2$,
K_ℓ = the number of parameters in model $\ell = 1, 2$,
Φ = the standard normal cumulative distribution function.
In other words, the probability that the adjusted likelihood ratio index of model 2 is greater by some $z > 0$ than that of model 1, given that the latter is the true model, is asymptotically bounded above by the right-hand side of equation (6.24). If we select the model with the greater $\bar{\rho}^2$, then this bounds the probability of erroneously choosing the incorrect model over the true specification. Note that when all N observations in the sample have all J alternatives, the bound becomes

$$\Pr(\bar{\rho}_2^2 - \bar{\rho}_1^2 > z) \leq \Phi\{-\sqrt{2Nz \ln J + (K_1 - K_2)}\}, \qquad z > 0. \tag{6.25}$$

This result implies that for 250 or more observations with two or more alternatives and models having the same number of parameters, if the $\bar{\rho}^2$ of the two models differ by 0.01 or more, the model with the lower $\bar{\rho}^2$ is almost certainly incorrect.

For example, compare the estimation results in Table 5.4 and in Table 6.3. Both models involve 15 parameters, but the "being early" and "being late" variables enter the specification in two different forms: linear for the first

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.857	0.219	-3.91	0.00
2	One stop–multiple airlines dummy	-1.26	0.228	-5.52	0.00
3	Round trip fare ($100)	-1.79	0.150	-11.97	0.00
4	Elapsed time (hours)	-0.309	0.0780	-3.96	0.00
5	Leg room (inches), if male (non stop)	0.0967	0.0328	2.95	0.00
6	Leg room (inches), if female (non stop)	0.181	0.0315	5.74	0.00
7	Leg room (inches), if male (one stop)	0.113	0.0297	3.82	0.00
8	Leg room (inches), if male (one stop)	0.0918	0.0272	3.37	0.00
9	Being early^2 (hours^2)	-0.0111	0.00169	-6.58	0.00
10	Being late^2 (hours^2)	-0.00731	0.00166	-4.39	0.00
11	More than 2 air trips per year (one stop–same airline)	-0.300	0.141	-2.12	0.03
12	More than 2 air trips per year (one stop–multiple airlines)	-0.0809	0.157	-0.52	0.61
13	Male dummy (one stop–same airline)	0.114	0.133	0.86	0.39
14	Male dummy (one stop–multiple airlines)	0.194	0.143	1.36	0.18
15	Round trip fare / income ($100/$1000)	-23.8	8.12	-2.93	0.00

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\widehat{\beta}) &= -1649.407 \\ -2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})] &= 2290.925 \\ \rho^2 &= 0.410 \\ \bar{\rho}^2 &= 0.404 \end{aligned}$$

Table 6.3: Estimation results for the airline itinerary choice: quadratic specification of the scheduled delays (early and late)

model, and quadratic for the second. The $\bar{\rho}^2$ are 0.408 and 0.404, respectively. As

$$\begin{aligned}\Phi\{-\sqrt{2Nz\ln J + (K_1 - K_2)}\} &= \Phi\{-\sqrt{2 \times 2544 \times 0.004 \times \ln 3}\} \\ &= \Phi(-4.73) \\ &= 0.00000113,\end{aligned}$$

we can safely reject the second model and prefer the first one.

The Cox Test

It is possible, however, to construct a composite specification with 17 coefficients from which both the model in Table 5.4 and the one in Table 6.3 can be derived as two special cases. Then we can perform two likelihood ratio tests for each of the two restricted models against the composite model. This procedure is known as the Cox test of separate families of hypotheses (Cox 1961, 1962).

To simplify the discussion, let's assume we want to decide between the following two models:

$$M_1 : U_{in} = \cdots + \beta x_{in} + \cdots + \varepsilon_n^{(1)} \tag{6.26}$$

$$M_2 : U_{in} = \cdots + \theta x_{in}^2 + \cdots + \varepsilon_n^{(2)}. \tag{6.27}$$

The implementation of the Cox test goes as follows: Construct a composite model, which in this case is written as:

$$M_C : U_{in} = \cdots + \beta x_{in} + \theta x_{in}^2 + \cdots + \varepsilon_n. \tag{6.28}$$

Then perform the two following likelihood ratio test:

- M_1 against M_C (testing the restriction $\theta = 0$),
- M_2 against M_C (testing the restriction $\beta = 0$).

There are four possible outcomes from those two tests:

1. M_1 is preferred to M_C, and M_C is preferred to M_2. Then M_1 is preferred to M_2.
2. M_2 is preferred to M_C, and M_C is preferred to M_1. Then M_2 is preferred to M_1.
3. M_1 is preferred to M_C, and M_2 is preferred to M_C. The test is then indecisive and the model with the highest adjusted likelihood ratio index $\bar{\rho}^2$ is preferred.

4. M_C is preferred to M_1, and M_C is preferred to M_2. Both models can be improved and other specifications must be investigated. Note that, in general, M_C cannot be accepted as a valid model.

We illustrate the Cox test for the models presented in Table 5.4 and in Table 6.3. As discussed in the previous section, both models involve 15 parameters, but the "being early" and "being late" variables enter the specification in two different forms, linear and quadratic respectively. The log likelihood of each model is reported in Table 6.4.

Model	Table	$\mathcal{L}(\widehat{\beta})$	K
Linear specification	5.4	-1640.525	15
Quadratic specification	6.3	-1649.407	15
Composite		-1640.487	17

Table 6.4: Final log likelihood for the models involved in the Cox test

The statistic for the likelihood ratio test is 0.076 when comparing the linear specification with the composite model. We cannot reject the null hypothesis associated with the restricted specification at the 0.05 significance level because $\chi^2_{2,0.05} = 5.99$, and the linear specification is preferred to the composite model. The same statistics for the comparison between the quadratic specification and the composite model is 17.84, and the quadratic model is rejected. Therefore, the linear specification is preferred.

A disadvantage of this procedure is the need to estimate a model with a potentially very large number of parameters. Below, we describe the J test developed by Davidson and MacKinnon (1981) which is a general solution to the selection between two non-nested models. The J test is in general preferred to the Cox test. As we will see, it is also subject to the same four outcomes.

The Davidson and MacKinnon J Test

This is a general treatment based on generating artificial regressions that embed two competing non nested model formulation to explain a given dependent variable. Consider two specifications:

$$M_1 : U_{in} = V^{(1)}_{in}(x_{in}; \beta) + \varepsilon^{(1)}_{in} \tag{6.29}$$

$$M_2 : U_{in} = V^{(2)}_{in}(x_{in}; \gamma) + \varepsilon^{(2)}_{in} \tag{6.30}$$

To choose between model 1 in equation 6.29 and model 2 in equation 6.30, we consider the following composite specification:

$$M_C : U_{in} = (1-\alpha)V_{in}^{(1)}(x_{in};\beta) + \alpha V_{in}^{(2)}(x_{in};\gamma) + \varepsilon_{in}. \tag{6.31}$$

Intuitively, the idea is to test the competing models against the composite model in equation (6.31). Note that if $\alpha = 0$, the model collapses to the model M_1 while with $\alpha = 1$, the composite model collapses to the model M_2. The major problem is that very often, the composite model cannot be estimated. For instance, consider that the two non nested models we want to compare are those displayed in Tables 5.4 and 6.3. The composite model is definitively not estimable given that there will be exact multicollinearity among the explanatory variables. A second problem comes from the fact that the α coefficient would not be identified.

The J test solution to this problem is to replace the unknown parameters not being tested by consistent estimates. In order to test M_1, one could consider the following composite model:

$$M_C : U_{in} = (1-\alpha)V_{in}^{(1)}(x_{in};\beta) + \alpha V_{in}^{(2)}(x_{in};\widehat{\gamma}) + \varepsilon_{in}, \tag{6.32}$$

where model 2 in equation (6.30) has been previously estimated. Thus, $V_{in}^{(2)}(x_{in};\widehat{\gamma})$ corresponds to the fitted systematic utility of model 2 and represents in this artificial model a single variable associated with the parameter α. Under the null hypothesis that model 1 is correct, the true value of α in the composite model is 0. The objective is then to test if $\alpha = 0$ using a t test. This would involve estimating model 1 with the additional variable computed as $V_{in}^{(2)}(x_{in};\widehat{\gamma})$.

In order to test M_2, one could instead consider the following composite model:

$$M_C : U_{in} = (1-\alpha)V_{in}^{(1)}(x_{in};\widehat{\beta}) + \alpha V_{in}^{(2)}(x_{in};\gamma) + \varepsilon_{in}. \tag{6.33}$$

where model 1 in equation (6.29) has been previously estimated. Thus, $V_{in}^{(1)}(x_{in};\widehat{\beta})$ corresponds to the fitted systematic utility of model 1 and represents in this artificial model a single variable associated with the parameter $(1-\alpha)$. Under the null hypothesis that model 2 is correct, the true value of α is 1. The objective is then to test if $\alpha = 1$ using a t test.

The J test has the same four outcomes presented in the previous section for the Cox test, that is

- M_1 is rejected and M_2 is not rejected. Then, it is reasonable to prefer model 2.
- M_1 is not rejected and M_2 is rejected. Then it is reasonable to prefer model 1.

- M_1 and M_2 are rejected. This indicates that better models should be developed.
- Neither M_1 nor M_2 can be rejected. Then, in this case, the data does not seem to be informative enough to distinguish between the two competing models, and the $\bar{\rho}^2$ should be used.

It should now be clear that one of the two models does not have to represent the truth. Both models could be unsatisfactory.

We now apply this test to decide between the specifications considered in Tables 5.4 and 6.3 which are called model 1 and model 2 respectively. We first test M_1 adding to the specification of model 1 the fitted systematic utility of model 2 as a single generic variable. Estimation results are presented in Table 6.5. Then we test M_2 adding to specification of model 2 the fitted systematic utility of model 1 as a single generic variable. Estimation results are provided in Table 6.6.

The t statistic of the coefficient associated with the artificial variable added to the specification of model 1 in Table 6.5 is very low, showing that the coefficient is not significantly different from 0. Model 1 is preferred. In addition, the t statistic of the coefficient associated with the artificial variable added to the specification of model 2 in Table 6.6 shows that the coefficient is not significantly different from 1:

$$\frac{1.06 - 1.00}{0.272} = 0.221 < 1.96. \tag{6.34}$$

Consequently, model 1 is preferred.

6.5.8 Tests of Nonlinear Specifications

The models that we have considered so far are based on linear-in-parameters utility functions. As in linear regression models this assumption is made for practical reasons and is not too restrictive because it allows nonlinear specifications of variables.

Often we do not have well-founded prior knowledge about the functional forms of these nonlinear transformation of variables, and we want to test empirically a wide range of nonlinear functions of the variables. Two useful approaches that involve estimating models that are linear in the parameters are the piecewise linear approximation and the power series expansion (see Section 5.4.1).

With a piecewise linear approximation we test the hypothesis that a coefficient may have different values for different ranges of the corresponding

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.878	0.205	-4.29	0.00
2	One stop–multiple airlines dummy	-1.27	0.213	-5.98	0.00
3	Round trip fare ($100)	-1.81	0.141	-12.82	0.00
4	Elapsed time (hours)	-0.304	0.0728	-4.17	0.00
5	Leg room (inches), if male (non stop)	0.100	0.0308	3.25	0.00
6	Leg room (inches), if female (non stop)	0.182	0.0298	6.10	0.00
7	Leg room (inches), if male (one stop)	0.113	0.0278	4.07	0.00
8	Leg room (inches), if female (one stop)	0.0930	0.0256	3.64	0.00
9	Being early (hours)	-0.149	0.0189	-7.88	0.00
10	Being late (hours)	-0.0964	0.0163	-5.93	0.00
11	More than 2 air trips per year (one stop–same airline)	-0.300	0.132	-2.27	0.02
12	More than 2 air trips per year (one stop–multiple airlines)	-0.0849	0.147	-0.58	0.56
13	Male dummy (one stop–same airline)	0.100	0.125	0.81	0.42
14	Male dummy (one stop–multiple airlines)	0.190	0.135	1.41	0.16
15	Round trip fare / income ($100/$1000)	-23.8	7.57	-3.14	0.00
16	α	-0.0698	0.301	-0.23	0.82

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\widehat{\beta}) &= -1640.493 \\ -2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})] &= 2308.754 \\ \rho^2 &= 0.413 \\ \bar{\rho}^2 &= 0.407 \end{aligned}$$

Table 6.5: Estimation results for the J test on the airline itinerary choice model: test of the linear specification

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.868	3.77	-0.23	0.82
2	One stop–multiple airlines dummy	-1.27	3.92	-0.32	0.75
3	Round trip fare ($100)	-1.80	2.60	-0.69	0.49
4	Elapsed time (hours)	-0.301	1.34	-0.22	0.82
5	Leg room (inches), if male (non stop)	0.0972	0.569	0.17	0.86
6	Leg room (inches), if female (non stop)	0.184	0.550	0.34	0.74
7	Leg room (inches), if male (one stop)	0.115	0.513	0.22	0.82
8	Leg room (inches), if female (one stop)	0.0919	0.471	0.20	0.85
9	Being early^2 (hours^2)	-0.0126	0.0283	-0.44	0.66
10	Being late^2 (hours^2)	-0.00982	0.0294	-0.33	0.74
11	More than 2 air trips per year (one stop–same airline)	-0.303	2.43	-0.12	0.90
12	More than 2 air trips per year (one stop–multiple airlines)	-0.0759	2.70	-0.03	0.98
13	Male dummy (one stop–same airline)	0.113	2.30	0.05	0.96
14	Male dummy (one stop–multiple airlines)	0.189	2.48	0.08	0.94
15	Round trip fare / income ($100/$1000)	-23.8	140.	-0.17	0.86
16	α	1.06	0.272	3.89	0.00

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\widehat{\beta}) &= -1640.492 \\ -2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})] &= 2308.756 \\ \rho^2 &= 0.413 \\ \bar{\rho}^2 &= 0.407 \end{aligned}$$

Table 6.6: Estimation results for the J test on the airline itinerary choice model: test of the quadratic specification

variables. This approach is demonstrated for the elapsed time variable in the model given in Table 6.7. It corresponds to a modification of the base specification of Table 5.4. The elapsed time variable in the base specification (with a coefficient estimate of -0.303) has been replaced with the following four variables. For simplicity we denote the value of the elapsed time variable as T.

$$T_{(0-2)} = \min(T, 2) = \begin{cases} T & \text{if } T < 2 \\ 2 & \text{otherwise,} \end{cases}$$

$$T_{(2--4)} = \max(0, \min(T-2, 2)) = \begin{cases} 0 & \text{if } T < 2 \\ T-2 & \text{if } 2 \leq T < 4 \\ 2 & \text{otherwise,} \end{cases}$$

$$T_{(4--8)} = \max(0, \min(T-4, 4)) = \begin{cases} 0 & \text{if } T < 4 \\ T-4 & \text{if } 4 \leq T < 8 \\ 4 & \text{otherwise,} \end{cases}$$

$$T_{(8+)} = \max(0, T-8) = \begin{cases} 0 & \text{if } T < 8 \\ T-8 & \text{otherwise,} \end{cases}$$

The systematic part of the utility function, with the four coefficient estimates $\widehat{\beta}_4$, $\widehat{\beta}_5$, $\widehat{\beta}_6$ and $\widehat{\beta}_7$, follows the pattern shown in Figure 6.2. The sensitivity to changes in elapsed time decreases as it gets larger, and increases again when it becomes large.

The χ^2 test statistic for the null hypothesis

$$\beta_4 = \beta_5 = \beta_6 = \beta_7$$

is obtained by comparison with the restricted model of Table 5.4, yielding a χ^2 statistic of

$$-2(-1640.525 + 1634.131) = 12.788. \tag{6.35}$$

Since $\chi^2_{3,0.05} = 7.81$, we can reject the hypothesis of a linear elapsed time variable.

Another specification can be investigated, as the elapsed time coefficient for the ranges 2–4 hours, and 4–8 hours have similar values (-0.268 and -0.231). In Figure 6.2, the slope barely changes after $T = 4$. Therefore, we estimate a model where only one elapsed time coefficient is considered for the range 2–8 hours. Results are presented in Table 6.8.

The model presented in Table 6.7 and the model presented in Table 6.8 can be compared using a likelihood ratio test as the model of Table 6.8 is a

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.933	0.225	-4.14	0.00
2	One stop–multiple airlines dummy	-1.32	0.232	-5.71	0.00
3	Round trip fare ($100)	-1.80	0.153	-11.82	0.00
4	Elapsed time (0 - 2 hours)	-0.802	0.241	-3.32	0.00
5	Elapsed time (2 - 4 hours)	-0.268	0.100	-2.67	0.01
6	Elapsed time (4 - 8 hours)	-0.231	0.0834	-2.77	0.01
7	Elapsed time ($>$ 8 hours)	-0.962	0.319	-3.02	0.00
8	Leg room (inches), if male (non stop)	0.104	0.0331	3.13	0.00
9	Leg room (inches), if female (non stop)	0.185	0.0320	5.79	0.00
10	Leg room (inches), if male (one stop)	0.118	0.0297	3.98	0.00
11	Leg room (inches), if female (one stop)	0.0939	0.0274	3.42	0.00
12	Being early (hours)	-0.150	0.0190	-7.87	0.00
13	Being late (hours)	-0.0988	0.0167	-5.90	0.00
14	More than 2 air trips per year (one stop–same airline)	-0.283	0.141	-2.00	0.05
15	More than 2 air trips per year (one stop–multiple airlines)	-0.0791	0.158	-0.50	0.62
16	Male dummy (one stop–same airline)	0.0838	0.134	0.63	0.53
17	Male dummy (one stop–multiple airlines)	0.181	0.144	1.26	0.21
18	Round trip fare / income ($100/$1000)	-23.1	8.17	-2.82	0.00

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\widehat{\beta}) &= -1634.131 \\ -2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})] &= 2321.478 \\ \rho^2 &= 0.415 \\ \bar{\rho}^2 &= 0.409 \end{aligned}$$

Table 6.7: Estimation results for the airline itinerary choice: piecewise linear elapsed time, with 4 intervals

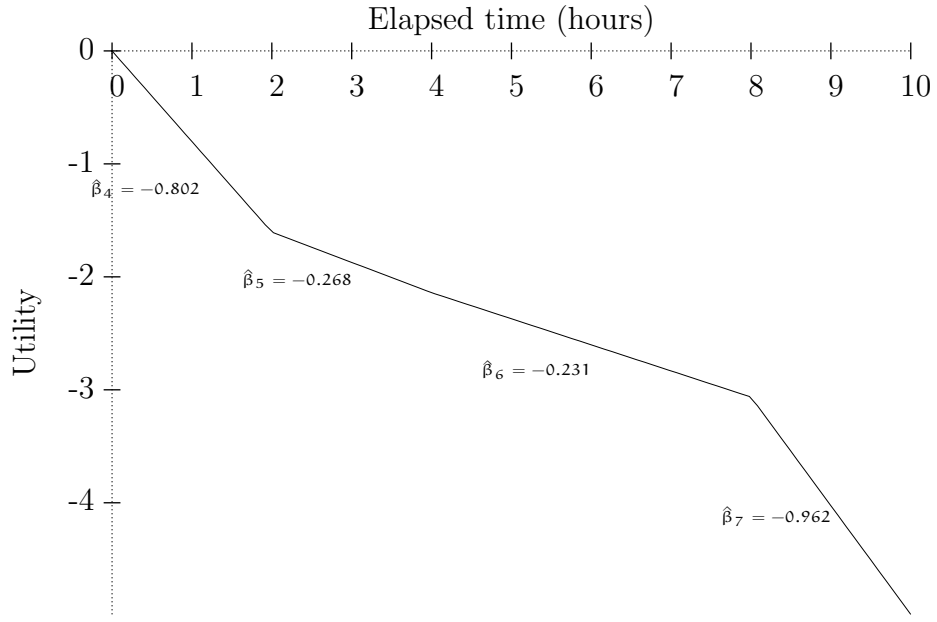


Figure 6.2: Piecewise linear approximation of the disutility of elapsed time

restriction of the model of Table 6.7. It yields a χ^2 statistic of

$$-2(-1634.266 + 1634.131) = 0.27. \tag{6.36}$$

Since $\chi^2_{1,0.05} = 3.84$, we cannot reject the hypothesis that the two models are equivalent. We perform also a likelihood ratio test to compare this last model with the base model in Table 5.4, testing the hypothesis that $\beta_4 = \beta_5 = \beta_6$. The statistics is

$$-2(-1634.266 + 1640.525) = 12.518.$$

Since $\chi^2_{2,0.05} = 5.99$, the model of Table 6.8 is preferred.

The major disadvantage of the piecewise linear approximation approach is related to the degrees of freedom. A large number of ranges causes a large increase of degrees of freedom and potentially a very small number of observations in some of the ranges. On the other hand, significant nonlinearities may be concealed with a small number of ranges. Another minor problem with this approach is that usually the endpoints of the ranges must be decided on arbitrarily because resources are often unavailable to test different range specifications.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.938	0.225	-4.17	0.00
2	One stop–multiple airlines dummy	-1.33	0.231	-5.74	0.00
3	Round trip fare ($100)	-1.81	0.152	-11.86	0.00
4	Elapsed time (0 - 2 hours)	-0.849	0.226	-3.75	0.00
5	Elapsed time (2 - 8 hours)	-0.239	0.0820	-2.91	0.00
6	Elapsed time (> 8 hours)	-0.931	0.313	-2.97	0.00
7	Leg room (inches), if male (non stop)	0.104	0.0331	3.13	0.00
8	Leg room (inches), if female (non stop)	0.186	0.0320	5.80	0.00
9	Leg room (inches), if male (one stop)	0.118	0.0297	3.97	0.00
10	Leg room (inches), if female (one stop)	0.0933	0.0274	3.41	0.00
11	Being early (hours)	-0.150	0.0190	-7.88	0.00
12	Being late (hours)	-0.0988	0.0167	-5.91	0.00
13	More than 2 air trips per year (one stop–same airline)	-0.283	0.141	-2.00	0.05
14	More than 2 air trips per year (one stop–multiple airlines)	-0.0791	0.158	-0.50	0.62
15	Male dummy (one stop–same airline)	0.0821	0.133	0.62	0.54
16	Male dummy (one stop–multiple airlines)	0.181	0.144	1.25	0.21
17	Round trip fare / income ($100/$1000)	-23.1	8.15	-2.83	0.00

Summary statistics

Number of observations = 2544

$$\mathcal{L}(0) = -2794.870$$
$$\mathcal{L}(c) = -2203.160$$
$$\mathcal{L}(\widehat{\beta}) = -1634.266$$
$$-2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})] = 2321.207$$
$$\rho^2 = 0.415$$
$$\bar{\rho}^2 = 0.409$$

Table 6.8: Estimation results for the airline itinerary choice: piecewise linear elapsed time, with 3 intervals

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.912	0.224	-4.08	0.00
2	One stop–multiple airlines dummy	-1.30	0.230	-5.64	0.00
3	Round trip fare ($100)	-1.80	0.153	-11.80	0.00
4	Elapsed time (hours)	-1.00	0.235	-4.27	0.00
5	Elapsed time (hours^2)	0.160	0.0507	3.14	0.00
6	Elapsed time (hours^3)	-0.0105	0.00347	-3.03	0.00
7	Leg room (inches), if male (non stop)	0.104	0.0332	3.14	0.00
8	Leg room (inches), if female (non stop)	0.185	0.0320	5.78	0.00
9	Leg room (inches), if male (one stop)	0.118	0.0298	3.94	0.00
10	Leg room (inches), if female (one stop)	0.0932	0.0274	3.40	0.00
11	Being early (hours)	-0.150	0.0191	-7.88	0.00
12	Being late (hours)	-0.0986	0.0167	-5.90	0.00
13	More than 2 air trips per year (one stop–same airline)	-0.279	0.142	-1.97	0.05
14	More than 2 air trips per year (one stop–multiple airlines)	-0.0727	0.157	-0.46	0.64
15	Male dummy (one stop–same airline)	0.0879	0.134	0.66	0.51
16	Male dummy (one stop–multiple airlines)	0.184	0.144	1.27	0.20
17	Round trip fare / income ($100/$1000)	-23.2	8.22	-2.82	0.00

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\widehat{\beta}) &= -1635.347 \\ -2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})] &= 2319.046 \\ \rho^2 &= 0.415 \\ \bar{\rho}^2 &= 0.409 \end{aligned}$$

Table 6.9: Estimation results for the airline itinerary choice: polynomial function of degree 3 of elapsed time

The second approach often used in practice is to represent a nonlinear function by a power series expansion that includes the linear specification as a special case. The elapsed time variable, for example, can be introduced as a polynomial by including in the model the three variables of elapsed time, elapsed time squared and elapsed time cubed. In principle, we can use a polynomial of a higher degree as long as we do not exhaust the available degrees of freedom. In practice, the elements of the polynomial are highly correlated, and the series must be truncated at a low degree. Note also that a polynomial of degree 2 would impose the function to be either concave or convex. It should not be used when no a priori assumption about the concavity or convexity of the function can be made. A model with elapsed time, elapsed time squared and elapsed time cubed is presented in Table 6.9. Testing this model against the base specification presented in Table 5.4, we have the following statistic:

$$-2(-1640.525 + 1635.347) = 10.356,$$

and we can reject the base model at a 5% level of significance ($\chi^2_{2,0.05} = 5.99$).

The two approaches to nonlinear utilities that we considered earlier are simple to perform because they involve estimations of models that are linear in the unknown parameters. However, it is also possible, but computationally more difficult, to test nonlinear transformations of variables that are not linear in the unknown parameters. One useful transformation for non-negative variables if the Box-Cox transform, described in Section 5.4.1:

$$\frac{x^\lambda - 1}{\lambda}, \qquad x \geq 0, \tag{6.37}$$

where λ is an unknown parameter. A model based on the model presented in Table 5.4, with a Box-Cox transformation of the elapsed time is presented in Table 6.10. The model of Table 5.4 is obtained from the linear restriction $\lambda = 1$ from the model of Table 6.10, so that a likelihood ratio test can be used to compare the two. Equivalently, the hypothesis $\lambda = 1$ can also be performed with a t test.

The χ^2 test statistic for the null hypothesis

$$\beta_{16} = 1$$

yields a χ^2 statistic of

$$-2(-1640.525 + 1639.317) = 2.416. \tag{6.38}$$

Since $\chi^2_{1,0.05} = 3.84$, we cannot reject the hypothesis of a linear elapsed time variable at the 5% level. The t test for $\lambda = 1$ is

$$\frac{0.690 - 1}{0.213} = -1.46,$$

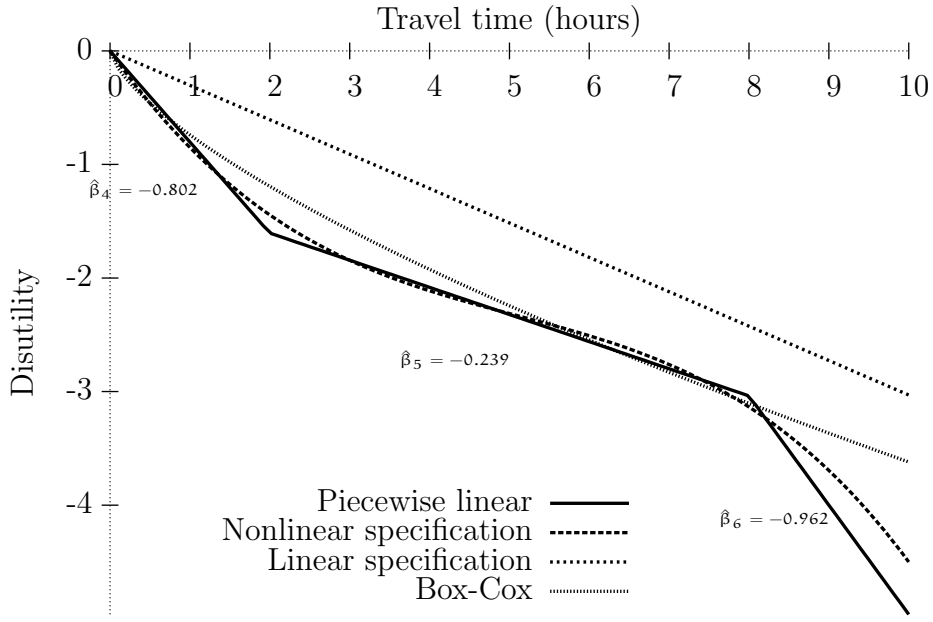


Figure 6.3: Piecewise linear approximation of the disutility of elapsed time

and the hypothesis cannot be rejected at the 5% level.

In Figure 6.3, we compare the linear, the piecewise linear, the polynomial and the Box-Cox specification.

As a conclusion, we prefer the model with a piecewise linear specification of the elapsed time (see Table 6.8).

6.5.9 Constrained Estimation

At the beginning of this section we discussed the role of prior information in the initial examination of the estimated coefficients. There are also situations where we find it useful to incorporate prior information directly in the estimation procedures. The justification for this approach is based on the assumption that the prior information is correct. By constraining parameters to what we believe to be their correct values, we improve the statistical efficiency of the model relative to the unconstrained estimates. The need for such a procedure usually arises in studies with inadequate data: either the sample is too small or one or more of the key variables have limited variability in the data. Three types of constraints are usually applied:

1. inequality constraints,

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.832	0.224	-3.72	0.00
2	One stop–multiple airlines dummy	-1.23	0.231	-5.31	0.00
3	Round trip fare ($100)	-1.79	0.151	-11.79	0.00
4	Elapsed time (hours)	-0.510	0.174	-2.93	0.00
5	Leg room (inches), if male (non stop)	0.101	0.0331	3.06	0.00
6	Leg room (inches), if female (non stop)	0.181	0.0319	5.69	0.00
7	Leg room (inches), if male (one stop)	0.114	0.0297	3.84	0.00
8	Leg room (inches), if female (one stop)	0.0948	0.0275	3.45	0.00
9	Being early (hours)	-0.151	0.0190	-7.95	0.00
10	Being late (hours)	-0.0977	0.0168	-5.82	0.00
11	More than 2 air trips per year (one stop–same airline)	-0.295	0.141	-2.09	0.04
12	More than 2 air trips per year (one stop–multiple airlines)	-0.0790	0.157	-0.50	0.62
13	Male dummy (one stop–same airline)	0.0993	0.133	0.74	0.46
14	Male dummy (one stop–multiple airlines)	0.188	0.144	1.31	0.19
15	Round trip fare / income ($100/$1000)	-23.7	8.10	-2.92	0.00
16	λ	0.690	0.213	3.24	0.00

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\widehat{\beta}) &= -1639.317 \\ -2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})] &= 2311.106 \\ \rho^2 &= 0.413 \\ \bar{\rho}^2 &= 0.408 \end{aligned}$$

Table 6.10: Estimation results for the airline itinerary choice: Box-Cox transform of elapsed time

2. fixed value constraints,

3. linear equality constraints.

The strongest a priori information is usually about the sign of a coefficient. Thus it is possible, for example, to incorporate constraints of the form

$$\beta_k \leq 0 \tag{6.39}$$

for the coefficients of elapsed time and round trip fare. Inequality constraints require a more complex estimation procedure and raise the following practical issue: Should we use a constrained estimation procedure if unconstrained estimation produces, for example, a significantly positive coefficient when we expect a negative sign? It is possible that the wrong sign is caused by a model specification error or by erroneous data, or that the prior information is wrong. If we impose the negativity constraint in such a case, we ignore a more fundamental problem that may affect all the other estimation results. Thus an estimation procedure with inequality constraints is usually of limited practical value.

The other two types of constraints are more useful and are easier to implement computationally. Fixed value constraints are used when we have a priori information about the values of individual parameters.

Constraints in the form of linear relationships between parameters are also straightforward to implement computationally. They are useful in cases with prior information in the form of equality of parameters, values of trade-offs, ratios of parameters, and other linear relationships.

Consider, for example, the trade-off between the traveling time and the round trip fare, for the choice of airline itinerary. Garrow (2010) has calculated the value of time in the same context of application. The value of time is equal to \$19.64 per hour. In section 6.5, we found it equal to \$14.91 per hour. We may wish to impose the value of time to be equal to the value of reference, \$19.64 per hour. We use the model presented in Table 5.4. The value of time for this model is given by (6.1):

$$\frac{\beta_4}{(\beta_3 + \frac{\beta_{15}}{\text{income}})\frac{1}{100}}, \tag{6.40}$$

so that the constraint can be written as

$$\beta_4 = 19.64(\beta_3 + \frac{\beta_{15}}{\text{income}})\frac{1}{100}. \tag{6.41}$$

Consequently, the parameter of elapsed time β_4 is replaced by this formulation, in all the utilities. This implies that the linear relationship

$$\beta_4 \text{Elapsed time} + \beta_3 \text{Round trip fare} + \beta_{15} \frac{\text{Round trip fare}}{\text{Income}}, \tag{6.42}$$

present in the base specification, after substituting equation 6.41 into equation 6.42 becomes:

$$\beta_3[\text{Round trip fare} + 0.1964 \times \text{Elapsed time}] + \beta_{15}[\frac{\text{Round trip fare}}{\text{Income}} + 0.1964 \times \frac{\text{Elapsed time}}{\text{Income}}],$$

We thus impose this constraint by constructing the two variables displayed in the last equation. The estimation results for this constrained estimation is given in Table 6.11. Since this model is a special case of the base specification, we can compute the value of the likelihood ratio test statistic with one degree of freedom as:

$$-2(-1647.748 + 1640.525) = 14.446.$$

Since $\chi^2_{1,0.05} = 3.84$, we can reject the constrained model.

6.6 Tests of the Model Structure

Up to this point the overall structure of the model was taken as given, and we explored statistical tests and informal procedures to develop acceptable specifications of the utility functions. In this section we consider the basic assumptions of the model structure itself.

In discrete choice models those assumptions are on either distributional properties of random utilities or the properties of choice probabilities. For example, the computational advantages of the logit model can be realized if we accept the property of the independence from irrelevant alternatives. The logit model also assumes that there are no random taste variations—that is, that the differences in tastes among individuals are captured by the socioeconomic variables in the model specification. We know a priori that these basic assumptions of the logit model can only be considered as reasonable approximations of more complex relationships. We are interested in finding out if significant violations occur and, if so, how to remedy the problems to obtain an acceptable model.

The direct approach to detect violations is a comparison of estimation results with a generalized model that relaxes the basic assumptions that we

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.546	0.137	-4.00	0.00
2	One stop–multiple airlines	-0.945	0.153	-6.18	0.00
3	Round trip fare + 0.1964 Elapsed time ($100)	-1.92	0.146	-13.16	0.00
4	Elapsed time (hours)	.	.		
5	Leg room (inches), if male (non stop)	0.103	0.0330	3.13	0.00
6	Leg room (inches), if female (non stop)	0.185	0.0317	5.82	0.00
7	Leg room (inches), if male (one stop)	0.113	0.0295	3.82	0.00
8	Leg room (inches), if female (one stop)	0.0898	0.0272	3.30	0.00
9	Being early (hours)	-0.151	0.0190	-7.94	0.00
10	Being late (hours)	-0.0950	0.0165	-5.74	0.00
11	More than two air trips per year (one stop–same airline)	-0.355	0.140	-2.53	0.01
12	More than two air trips per year (one stop–multiple airlines)	-0.142	0.154	-0.92	0.36
13	Male dummy (one stop–same airline)	0.0735	0.133	0.55	0.58
14	Male dummy (one stop–multiple airlines)	0.171	0.143	1.19	0.23
15	(Round trip fare + 0.1964 Elapsed time) / income ($100/$1000)	-14.6	6.99	-2.09	0.04

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\hat{\beta}) &= -1647.748 \\ -2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 2294.243 \\ \rho^2 &= 0.410 \\ \bar{\rho}^2 &= 0.405 \end{aligned}$$

Table 6.11: Estimation results for the airline itinerary choice: constrained specification

want to test. The practical difficulty of implementing this approach arises from the prohibitive computational costs of estimating a less restrictive model structure. The probit model with unrestricted covariance matrix and random taste variations, discussed in section 5.9, is the most general form available. However, it can be reasonably estimated only for a small number of alternatives and parameters, eliminating its use in most practical situations. The same level of generality can also be obtained with the logit mixture model with random taste variations and unrestricted error covariance structure. Later chapters cover in detail this model. It is a lot more demanding computationally than the logit. Thus we concentrate on tests that do not require estimation of probit or logit mixture models. The tests that are described are of lagrange multiplier type. Here, to keep the discussion simple, we focus on the implementation which requires only the estimation of a logit model. Lagrange multiplier tests are convenient since they only require the use of the restricted version of the model. Moreover, when a serious violation is detected, we attempt to find improvements to the specification of the model that give more satisfactory results and so avoid the need to use a more complex model structure.

6.6.1 Tests of the IIA Assumption

McFadden et al. (1977) investigated a wide range of computationally feasible tests to detect violations of the IIA assumption. (This assumption is discussed in sections 3.7 and 5.3). Two obvious cases in which IIA violations can occur:

1. Alternatives share unobserved attributes, so the error terms are correlated.

2. Error terms of the alternatives have different variances violating the assumption that they are "identically distributed".

We describe here the most useful of their tests which proved to be the most powerful. The test involves comparisons of logit models estimated with subsets of alternatives from the universal choice set. If the IIA assumption (i.e., the logit model structure) holds for the full choice set, then the logit model also applies to a choice from any subset of alternatives.

We discuss two IIA tests: the Hausman-McFadden test and McFadden's omitted variables test. The first test is based on the identification of a subset of alternatives suspected to share unobserved attributes. IIA is rejected if the parameter estimates differ from the full choice set estimates. The second test is based on estimating a more general model and testing the restriction

that the logit model is valid. Note that these tests all require the modeler to choose the subset $\widetilde{\mathcal{C}}_n$ to perform the tests on.

The logit model for the full choice set is written as:

$$P(i|\mathcal{C}_n) = \frac{e^{V_{in}}}{\sum_{j\in\mathcal{C}_n} e^{V_{jn}}}, \tag{6.43}$$

and for a restricted choice set $\widetilde{\mathcal{C}}_n \subseteq \mathcal{C}_n$ we obtain the following logit model:

$$P(i|\widetilde{\mathcal{C}}_n) = \frac{e^{V_{in}}}{\sum_{j\in\widetilde{\mathcal{C}}_n} e^{V_{jn}}}, \qquad i \in \widetilde{\mathcal{C}}_n \subseteq \mathcal{C}_n. \tag{6.44}$$

With the absence of some alternatives from $\widetilde{\mathcal{C}}_n$, it is in general possible to identify only a subvector of β. *This subvector does not include the parameters specific to alternatives not included in the restricted choice set* $\widetilde{\mathcal{C}}_n$. The estimation data used for the model with a restricted set of alternatives is a subset of the full data set, *omitting observations with chosen alternatives not in the restricted choice set.*

Thus, if the logit model is correctly specified, we can obtain consistent coefficient estimates of the same subvector of parameters from a logit model estimated with a full choice set and from a logit model estimated with a restricted choice set. Denote the estimated coefficients from the restricted set of alternatives as $\widehat{\beta}_{\widetilde{\mathcal{C}}_n}$ and the estimated values for the same subvector of coefficients from a model with a full choice set as $\widehat{\beta}_{\mathcal{C}_n}$. Denote analogously the covariance matrices as $\Sigma_{\widehat{\beta}_{\widetilde{\mathcal{C}}_n}}$ and $\Sigma_{\widehat{\beta}_{\mathcal{C}_n}}$, where the latter matrix is the appropriate submatrix from the estimation with a full choice set. Under the hypothesis that IIA holds, Hausman and McFadden (1984) have shown that

$$(\widehat{\beta}_{\widetilde{\mathcal{C}}_n} - \widehat{\beta}_{\mathcal{C}_n})'(\Sigma_{\widehat{\beta}_{\widetilde{\mathcal{C}}_n}} - \Sigma_{\widehat{\beta}_{\mathcal{C}_n}})^{-1}(\widehat{\beta}_{\widetilde{\mathcal{C}}_n} - \widehat{\beta}_{\mathcal{C}_n}) \sim \chi^2_{\tilde{K}}$$

where $\mathcal{C}_n$ and $\widetilde{\mathcal{C}}_n$ denote the full choice set and sub-choice set, respectively, and $\Sigma_{\widehat{\beta}_{\mathcal{C}_n}}$ and $\Sigma_{\widehat{\beta}_{\widetilde{\mathcal{C}}_n}}$ are the variance-covariance matrices of the (common) estimated coefficients under each model. $\tilde{K}$ is the number of coefficients that are identifiable from the restricted choice set model (i.e., the dimension of $\widehat{\beta}_{\widetilde{\mathcal{C}}_n}$). Therefore, IIA is rejected if the statistics is larger than the quantile of the $\chi^2_{\tilde{K}}$ at the desired level.

McFadden Omitted Variables Test

McFadden's test checks if cross-alternative variables enter the model. If so, IIA is violated.

The idea here is to estimate a more general model and test the validity of the logit restriction. The general model is a "logit" which includes an auxiliary variable in the utility function of each alternative in the subset of alternatives that are suspected to be correlated. This variable captures the influence on the utility of i of the attributes of the alternatives that are correlated with alternative i. The logit restriction is that the coefficients of the auxiliary variables are zero.
The test includes the following steps:

1. Estimate the basic logit model, using all the observations.
2. Suppose $\widetilde{\mathcal{C}}$ is a specified subset of alternatives. Create new variables in one of the following two forms:
   (a) If the utility function is linear in parameters, and x_{in} are the variables in the basic logit model, define new variables:

$$z_{in}^{\widetilde{\mathcal{C}}_n} = \begin{cases} x_{in} - \dfrac{\sum_{j\in\widetilde{\mathcal{C}}_n} \widehat{P}(j|\mathcal{C}_n)x_{jn}}{\sum_{j\in\widetilde{\mathcal{C}}_n} \widehat{P}(j|\mathcal{C}_n)} & \text{if } i \in \widetilde{\mathcal{C}}_n \\ 0 & \text{if } i \notin \widetilde{\mathcal{C}}_n \end{cases}, \tag{6.45}$$

   where $\widehat{P}(j|\mathcal{C}_n)$ is calculated from the estimated model. The variables $z_{in}^{\widetilde{\mathcal{C}}_n}$ can be written in abbreviated form as

$$z_{in}^{\widetilde{\mathcal{C}}_n} = \delta_{i\widetilde{\mathcal{C}}_n}\left(x_{in} - \widehat{x}_{\widetilde{\mathcal{C}}_n}\right),$$

   where $\delta_{i\widetilde{\mathcal{C}}_n} = 1$ $i \in \widetilde{\mathcal{C}}_n$, 0 otherwise, and $\widehat{x}_{\widetilde{\mathcal{C}}_n} = \dfrac{\sum_{j\in\widetilde{\mathcal{C}}_n} \widehat{P}(j|\mathcal{C}_n)x_{jn}}{\sum_{j\in\widetilde{\mathcal{C}}_n} \widehat{P}(j|\mathcal{C}_n)}$.
   (b) If $\widehat{V}_{in} = \widehat{\beta}'x_{in}$ is the systematic utility from the basic model calculated at the basic model estimated parameters, define the new variable

$$z_{in}^{\widetilde{\mathcal{C}}_n} = \begin{cases} \widehat{V}_{in} - \dfrac{\sum_{j\in\widetilde{\mathcal{C}}_n} \widehat{P}(j|\mathcal{C}_n)\widehat{V}_{jn}}{\sum_{j\in\widetilde{\mathcal{C}}_n} \widehat{P}(j|\mathcal{C}_n)} & \text{if } i \in \widetilde{\mathcal{C}}_n \\ 0 & \text{if } i \notin \widetilde{\mathcal{C}}_n \end{cases}, \tag{6.46}$$

   or more compactly, $z_{in}^{\widetilde{\mathcal{C}}_n} = \delta_{i\widetilde{\mathcal{C}}_n}\left(\widehat{V}_{in} - \widehat{V}_{\widetilde{\mathcal{C}}_n}\right)$, where $\delta_{i\widetilde{\mathcal{C}}_n} = 1$ if $i \in \widetilde{\mathcal{C}}_n$, 0 otherwise, and $\widehat{V}_{\widetilde{\mathcal{C}}_n} = \dfrac{\sum_{j\in\widetilde{\mathcal{C}}_n} \widehat{P}(j|\mathcal{C}_n)\widehat{V}_{jn}}{\sum_{j\in\widetilde{\mathcal{C}}_n} \widehat{P}(j|\mathcal{C}_n)}$.

Since $z_{in}^{\widetilde{\mathcal{C}}_n}$ is non zero only for the alternatives in set $\widetilde{\mathcal{C}}_n$, this is really only about the set of alternatives in $\widetilde{\mathcal{C}}_n$. It contains information specific to the other alternatives in $\widetilde{\mathcal{C}}_n$. This is really the spirit of the proposed test: test the presence of cross-alternative variables.

3. Estimate an expanded model that contains the basic model variables plus the new variables $z_{in}^{\widetilde{\mathcal{C}}_n}$, and test whether the added variables are significant. One way is to carry out a likelihood ratio test that the coefficients of $z_{in}^{\widetilde{\mathcal{C}}_n}$ are zero, using the following statistic:

$$= -2[(\text{Log likelihood without } z\text{'s}) - (\text{Log likelihood with } z\text{'s})]$$

If IIA holds, this statistic has a χ^2 distribution with one degree of freedom if form (b) is used or with degrees of freedom equal to the number of added $z_{in}^{\widetilde{\mathcal{C}}_n}$ variables (after eliminating any that are linearly dependent) if form (a) is used.

Properties

- The test using variables of type (a) is equivalent to the Hausman-McFadden test for the subset of alternatives $\widetilde{\mathcal{C}}$.
- The test using variables of type (b) is equivalent to a one-degree-of-freedom Hausman-McFadden test focused in the direction determined by the parameters β. It is likely to have greater power than the previous test if there is substantial variation in the $V's$ across $\widetilde{\mathcal{C}}$.
- The test of type (b) is equivalent to a test of the basic logit model against a nested logit model in which subjects discriminate more sharply between alternatives within $\widetilde{\mathcal{C}}$ than they do between alternatives that are not both in $\widetilde{\mathcal{C}}$. One plus the coefficient of the variable can be interpreted as a preliminary estimate of the inclusive value coefficient for the nest $\widetilde{\mathcal{C}}$.
- The tests described above are for a single specified subset $\widetilde{\mathcal{C}}$. However, it is trivial to test the logit model against several subsets of alternatives at once, simply by introducing an omitted variable for each suspected nest, and testing jointly that the coefficients of these omitted variables are zero. Alternative nests in the test can be overlapping. The coefficients on the omitted variables provide some guide to the choice of nesting structure if the IIA hypothesis fails.

- If there are $\widetilde{\mathcal{C}}$-specific dummy variables in the basic model, then some of the omitted type (a) variables duplicate these variables, and cannot be used in the testing procedure.
- One may get a rejection of the null hypothesis either if IIA is false, or if there is some other problem with the model specification, such as omitted variables or a failure of the logit form due, say, to asymmetry or to fat tails in the disturbances.

Application of McFadden Omitted Variables Test

We now apply the McFadden omitted variables test to the choice of airline itinerary. We suspect that the two alternatives which consider one stop may share unobserved variables related to the presence of a transfer, which would violate the i.i.d. assumption, and generate a model not complying with the IIA property. To reflect this, we define the set $\widetilde{\mathcal{C}}$ as being composed of the alternatives 2 and 3. The implementation steps were presented above. We first estimate the base model on the full set of alternatives. Then, we compute the $\widehat{V}$'s and the choice probabilities $\widehat{P}(j|\mathcal{C}_n)$. Then, the auxiliary variables $z_{in}^{\widetilde{\mathcal{C}}}$ defined in equation (6.46) were computed for each of the two alternatives in choice set $\widetilde{\mathcal{C}}$. Those variables were then added to the database and the model is re-estimated adding the variables $z_{2n}^{\widetilde{\mathcal{C}}}$ in alternative 2, $z_{3n}^{\widetilde{\mathcal{C}}}$ in alternative 3, and a value of 0 for alternative 1, with a single generic coefficient. Results are presented in Table 6.12. The coefficient of the auxiliary variable parameter, β_{18}, is significantly different from zero as indicated by the value of its t statistic. Performing a likelihood ratio test for the null hypothesis $\beta_{18} = 0$, the test statistic is:

$$-2(-1634.266 + 1611.816) = 44.9, \tag{6.47}$$

where the restricted model is the model without the auxiliary variables and the unrestricted model is the model with the auxiliary variables. The test statistic is asymptotically χ^2 distributed with 1 degree of freedom. Since 44.9 $>$ 3.841, (the critical value of the χ^2 distribution with 1 degree of freedom at the 5% level), we reject the null hypothesis and conclude that the IIA property does not hold for the alternatives containing one stop. Note that these two tests are equivalent, as there is only one degree of freedom here.

6.6.2 Test of Taste Variations

Choice theory and the discrete choice models derived from it are disaggregate relationships, by nature. They describe the behavior of a single decision

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-1.06	0.228	-4.66	0.00
2	One stop–multiple airlines dummy	-1.29	0.234	-5.51	0.00
3	Round trip fare ($100)	-1.56	0.148	-10.53	0.00
4	Elapsed time (0 - 2 hours)	-0.826	0.223	-3.70	0.00
5	Elapsed time (2 - 8 hours)	-0.177	0.0843	-2.10	0.04
6	Elapsed time (> 8 hours)	-0.767	0.313	-2.45	0.01
7	Leg room (inches), if male (non stop)	0.100	0.0313	3.20	0.00
8	Leg room (inches), if female (non stop)	0.174	0.0303	5.74	0.00
9	Leg room (inches), if male (one stop)	0.0885	0.0304	2.91	0.00
10	Leg room (inches), if female (one stop)	0.0710	0.0277	2.57	0.01
11	Being early (hours)	-0.128	0.0189	-6.76	0.00
12	Being late (hours)	-0.0809	0.0165	-4.89	0.00
13	More than 2 air trips per year (one stop–same airline)	-0.243	0.139	-1.75	0.08
14	More than 2 air trips per year (one stop–multiple airlines)	-0.127	0.158	-0.80	0.42
15	Male dummy (one stop–same airline)	0.123	0.133	0.92	0.36
16	Male dummy (one stop–multiple airlines)	0.153	0.145	1.06	0.29
17	Round trip fare / income ($100/$1000)	-21.0	7.49	-2.80	0.01
18	McFadden's auxiliary variable for alternatives (one stop)	0.713	0.150	4.74	0.00

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\widehat{\beta}) &= -1611.816 \\ -2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})] &= 2366.106 \\ \rho^2 &= 0.423 \\ \bar{\rho}^2 &= 0.417 \end{aligned}$$

Table 6.12: Estimation results for the airline itinerary choice: specification for McFadden's omitted variables test

maker. Yet in estimated models we assume the same model structure and the same values of the unknown parameters for all members of the population represented by the sample. Two approaches have been applied to account for the differences in the values of taste parameters among individuals. In the first approach the socioeconomic variables that describe the decision maker are included in various forms in the specification of the utility functions. The second, and more general approach, captures unobservable taste variations by using model structures with random taste variations, also known as random coefficients models. (See the discussion of the random coefficients logit and probit models in later chapters.)

Unfortunately random taste variation models are complex and expensive computationally. Moreover, even if we can estimate a random coefficients model, it is still desirable to capture systematic taste variations in the utility specification. It may be erroneous to assume constant parameters of the distribution of the uncaptured taste differences because they may change over time with demographic shifts. Therefore we describe a procedure, called *market segmentation*, to search for systematic variations of taste parameters among population subgroups in order to include them explicitly in the specification of the variables.

To perform the taste variation test, we first classify the estimation data into socioeconomic groups. The simplest market segmentation scheme is based on ranges of the value of a single socioeconomic characteristic, such as low, medium, and high income ranges. We assume the same specification across market segments and apply the estimation procedure to the first subset of data, to the second subset, and so on, and finally estimate the pooled model with the full data set.

Denote by N_g the sample size of market segment $g = 1, \cdots, G$, where G is the number of market segments and

$$\sum_{g=1}^{G} N_g = N, \tag{6.48}$$

where N is the full sample size. The null hypothesis of no taste variations across the market segments is

$$\beta^1 = \beta^2 = ... = \beta^G \tag{6.49}$$

where β^g is the vector of coefficients of market segment g. The likelihood ratio test statistic is given by

$$-2\left[\mathcal{L}_N(\widehat{\beta}) - \sum_{g=1}^{G} \mathcal{L}_{N_g}(\widehat{\beta}^g)\right],$$

where $\mathcal{L}_N(\widehat{\beta})$ is the log likelihood for the restricted model that is estimated on the pooled data set with a single vector of coefficients $\widehat{\beta}$; $\mathcal{L}_{N_g}(\widehat{\beta}^g)$ is the maximum likelihood of the model estimated with the gth subset of the data. This test statistic is χ^2 distributed with the degrees of freedom equal to the number of restrictions,

$$\sum_{g=1}^{G} K_g - K,$$

where K_g is the number of coefficients in the gth market segment model. K_g is equal to K except when one or more of the pooled model coefficients are not identifiable with the gth subset of the data.

For the example of the airline itinerary choice, we test the market segmentation by trip purposes. Two purposes are considered: leisure and non leisure. The base specification is the same than in Table 5.4, but estimated on the full data set, meaning that now we also consider the non leisure trips in the estimation. Estimation results are presented in Table 6.13.

The value of the log likelihood is -2300.453 for the model estimated on the entire population, -1640.525 for the model estimated on the "leisure" segment, and -629.080 for the model estimated on the "non leisure" segment. The statistic of the likelihood ratio tests against the base specification is:

$$-2(-2300.453-(-1640.525-629.080)) = -2(-2300.453+2269.605) = 61.696.$$

With this evidence, we can reject at the 5% level the joint specification for the leisure and non leisure trips, as the test statistic is higher than $\chi^2_{15,0.05}$, as 61.693 is larger than 25.00.

A rejection of the hypothesis of equal vectors of coefficients across market segments suggests further exploration of the importance of and the reasons for the statistically significant differences. It is useful to know if the rejection of the joint hypothesis can be attributed to a single or a subset of coefficients. This can be done by comparing individual coefficients between market segments. To simplify the discussion, assume that the two segments are labeled (1) and (2), and that the specification is linear-in-parameters:

$$U_{in}^{(1)} = \sum_{k=1}^{K} \beta_k^{(1)} x_{ink} + \varepsilon_{in}^{(1)} \tag{6.50}$$

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.942	0.190	-4.95	0.00
2	One stop–multiple airlines dummy	-1.29	0.198	-6.53	0.00
3	Round trip fare ($100)	-1.60	0.124	-12.83	0.00
4	Elapsed time (hours)	-0.299	0.0672	-4.45	0.00
5	Leg room (inches), if male (non stop)	0.108	0.0268	4.03	0.00
6	Leg room (inches), if female (non stop)	0.141	0.0272	5.18	0.00
7	Leg room (inches), if male (one stop)	0.125	0.0250	4.99	0.00
8	Leg room (inches), if female (one stop)	0.0850	0.0233	3.64	0.00
9	Being early (hours)	-0.140	0.0162	-8.64	0.00
10	Being late (hours)	-0.105	0.0138	-7.61	0.00
11	More than 2 air trips per year (one stop–same airline)	0.0263	0.114	0.23	0.82
12	More than 2 air trips per year (one stop–multiple airlines)	0.0144	0.123	0.12	0.91
13	Male dummy (one stop–same airline)	0.100	0.133	0.75	0.45
14	Male dummy (one stop–multiple airlines)	0.189	0.144	1.31	0.19
15	Round trip fare / income ($100/$1000)	-24.8	7.57	-3.27	0.00

Summary statistics

Number of observations = 3609

$$\begin{aligned} \mathcal{L}(0) &= -3964.892 \\ \mathcal{L}(\widehat{\beta}) &= -2300.453 \\ -2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})] &= 3328.878 \\ \rho^2 &= 0.420 \\ \bar{\rho}^2 &= 0.416 \end{aligned}$$

Table 6.13: Estimation results for the airline itinerary choice: base specification (same as Table 5.4) estimated on the full data set including both leisure and non leisure trips

Parameter number	Description	Coefficient estimate (Rob. asympt. std error) Leisure	Non Leisure
1	One stop–same airline dummy	-0.879	-1.37
		(-4.02)	(-3.36)
2	One stop–multiple airlines dummy	-1.27	-1.58
		(-5.60)	(-3.62)
3	Round trip fare ($100)	-1.81	-1.29
		(-11.99)	(-6.32)
4	Elapsed time (hours)	-0.303	-0.300
		(-3.90)	(-2.24)
5	Leg room (inches), if male (non stop)	0.100	0.110
		(3.04)	(2.38)
6	Leg room (inches), if female (non stop)	0.182	0.0212
		(5.71)	(0.39)
7	Leg room (inches), if male (one stop)	0.113	0.166
		(3.80)	(3.58)
8	Leg room (inches), if female (one stop)	0.0931	0.0661
		(3.41)	(1.37)
9	Being early (hours)	-0.151	-0.118
		(-7.99)	(-3.43)
10	Being late (hours)	-0.0975	-0.126
		(-5.83)	(-4.86)
11	More than 2 air trips per year (one stop–same airline)	-0.300	0.0308
		(-2.12)	(0.11)
12	More than 2 air trips per year (one stop–multiple airlines)	-0.0847	0.0611
		(-0.54)	(0.19)
13	Male dummy (one stop–same airline)	0.100	-0.0446
		(0.75)	(-0.19)
14	Male dummy (one stop–multiple airlines)	0.189	-0.349
		(1.31)	(-1.39)
15	Round trip fare / income ($100/$1000)	-23.8	-17.6
		(-2.94)	(-1.24)
Summary statistics			
Number of observations by market segment (total: 3609)		2544	1065
$\mathcal{L}_{N_g}(\widehat{\beta})$		-1640.525	-629.08
$\mathcal{L}(0)$ = -3964.892			
$\mathcal{L}(\widehat{\beta})$= -2269.605			
$-2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})]$= 3390.574			
$\rho^2 = 0.428$			
$\overline{\rho}^2 = 0.420$			

Table 6.14: Estimation results for the airline itinerary choice: market segmentation by trip purpose (leisure and non leisure)

for segment (1) and

$$U_{in}^{(2)} = \sum_{k=1}^{K} \beta_k^{(2)} x_{ink} + \varepsilon_{in}^{(2)} \tag{6.51}$$

for segment (2). For each k, we want to test the hypothesis that $\beta_k^{(1)} = \beta_k^{(2)}$. As described in Section 6.5.2, the statistic for the asymptotic t test is given by (6.6), which involves the term $\text{Cov}(\widehat{\beta}_k^{(1)}, \widehat{\beta}_k^{(2)})$.

In order to obtain this quantity, all parameters have to be estimated jointly. This is performed with the following specification:

$$U_{in} = \sum_{k=1}^{K} \beta_k^{(1)} x_{ink} \delta_n^{(1)} + \sum_{k=1}^{K} \beta_k^{(2)} x_{ink} \delta_n^{(2)} + \varepsilon_{in}, \tag{6.52}$$

where $\delta_n^{(1)} = 1$ if observation n belongs to segment (1), 0 otherwise, and $\delta_n^{(2)}$ is defined in the same way. If, for some reasons, the joint model cannot be estimated, the statistic (6.6) can be approximated using the independence assumption, that is $\text{Cov}(\widehat{\beta}_k^{(1)}, \widehat{\beta}_k^{(2)}) = 0$. However, this latter approach is not recommended.

The application of this test for the models presented in Table 6.14 is given in Table 6.15. It is possible that all the t tests are insignificant despite the fact that the joint likelihood ratio test is significant. It is also possible that the joint test will not reject the hypothesis, though a few individual coefficients are significantly different. From Table 6.15 we find that the coefficients associated with the round trip fare β_3, to the leg room for females and alternative "Non stop" (β_6), are significantly different between the two types of trip purposes. For β_3, it seems obvious, as the sensitivity to fare is not the same in case of leisure and non leisure trips.

A limited market segmentation test can be applied to a subset of the coefficients. We use the likelihood ratio test for the null hypothesis of equality of subsets of coefficients across market segments. This test is performed by comparing the pooled model with an unrestricted model that is estimated with the full data set but with a longer vector of coefficients in which a subset of the original coefficients is replaced with two or more market segment specific subsets of coefficients. A typical application is the test of the equality of the coefficients of attributes (e.g., the level-of-service variables in the mode choice model) across market segments.

In the example, we employed simple univariate market segmentation schemes. However, the tests that were presented can be applied to all types of multivariate segmentation schemes. In some of these market segment models it may be impossible to incorporate the different tastes in the specifications

Parameter number	Description	t statistic
1	"One stop–same airline" dummy	1.06
2	"One stop–multiple airlines" dummy	0.61
3	Round trip fare ($100)	*-2.07*
4	Elapsed time (hours)	-0.02
5	Leg room (inches), if male (non stop)	-0.18
6	Leg room (inches), if female (non stop)	*2.54*
7	Leg room (inches), if male (one stop)	-0.97
8	Leg room (inches), if female (one stop)	0.49
9	Being early (hours)	-0.82
10	Being late (hours)	0.92
11	More than 2 air trips per year "one stop–same airline"	-1.05
12	More than 2 air trips per year "one stop–multiple airlines"	-0.40
13	Male dummy "one stop–same airline"	0.54
14	Male dummy "one stop–multiple airlines"	1.86
15	Round trip fare / income ($100/$1000)	-0.38

Table 6.15: Asymptotic t tests for coefficient differences between market segments of trip purposes

of the utilities, so it may be better to proceed with separate market segment models.

The market segmentation likelihood ratio tests for joint hypotheses and the individual t tests are also used in tests of spatial and temporal stability, or transferability, of models. In a test of transferability the samples from different locations or from different points in time are treated in the same way as samples from different socioeconomic market segments. There is, however, a major difference between market segmentation and transferability tests. The first test is useful during the model development process and is concerned with the entire set of coefficients. A transferability test, on the other hand, is focused on the stability of the policy-relevant coefficients. For a mode choice model we would be interested in the stability of the travel time and cost coefficients between different locations, but we assume a priori that alternative specific constants and coefficients of socioeconomic variables may differ. Thus the unrestricted model will be the collection of models estimated for the separate data sets. In the restricted model the coefficients of interest, such as elapsed time and cost coefficients, will be constrained to have the same value across data sets, and all other coefficients will be allowed to vary.

6.6.3 Test of Heteroscedasticity

Since the scale of the utilities of the logit model is inversely proportional to the standard error of the random utility components (see section 5.2), the basic assumption of a constant scale for all the observations is the same as the assumption of homoscedastic (or equal variance) random utilities. As discussed in Section 5.4.3, there are circumstances where this assumption may be inappropriate, as we may expect the variance of the error term to have different variances in different segments of the population.

It is explained in Section 5.4.3 how to estimate a scale parameter λ_g for each segment g of the population, except for one segment where the scale is normalized to one. Therefore, in order to test if two segments are associated with the same scale, a t test must be performed to test if the two parameters are equal. If one of the two segments to be tested corresponds to the segment where the scale parameter has been normalized to one, a t test must be performed to test that the scale parameter of the other segment is equal to one.

For the choice of airline itinerary, suppose that we want to test the hypothesis that two market segments have different scale parameters. We consider two examples: a segmentation by the gender and a segmentation by the trip frequency. For the latter, we consider a traveler to be a "frequent" flyer if the number of trips traveled per year is 3 or more. Note that we use only observations related to leisure trips in the remaining analysis.

For the gender, the scale parameter of the male group is fixed to 1, whereas for the female it is estimated. Results are shown in Table 6.16. The scale parameter of the female group is not significantly different from 1 at the 95% level (t statistic: -1.15). This means that it is not significantly different from the scale parameter of the male group. In addition, we perform a likelihood ratio test between this model and the restricted model of Table 5.4, where there is a unique scale parameter:

$$-2(-1640.525 + 1639.693) = 1.664 \tag{6.53}$$

This is under $\chi^2_{1,0.95} = 3.84$, which means that we cannot reject the hypothesis that the two models are equivalent. Consequently we keep the model of Table 5.4, which contains a unique scale parameter.

For the trip frequency, the scale parameter of the non frequent flyer group is fixed to 1, whereas for the frequent flyer it is estimated. Results are shown in Table 6.17. The scale parameter of the frequent flyer group is not significantly different from 1 at the 95% level (-1.42). This means that it is not significantly different from the scale parameter of the non frequent flyer

Parameter number	Description[a]	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.885	0.228	-3.88	0.00
2	One stop–multiple airlines dummy	-1.29	0.237	-5.43	0.00
3	Round trip fare ($100)	-1.90	0.183	-10.39	0.00
4	Elapsed time (hours)	-0.312	0.0816	-3.82	0.00
5	Leg room (inches), if male (non stop)	0.110	0.0370	2.97	0.00
6	Leg room (inches), if female (non stop)	0.186	0.0328	5.67	0.00
7	Leg room (inches), if male (one stop)	0.123	0.0334	3.68	0.00
8	Leg room (inches), if female (one stop)	0.0947	0.0280	3.38	0.00
9	Being early (hours)	-0.156	0.0201	-7.78	0.00
10	Being late (hours)	-0.103	0.0180	-5.71	0.00
11	More than 2 air trips per year (one stop–same airline)	-0.313	0.148	-2.11	0.03
12	More than 2 air trips per year (one stop–multiple airlines)	-0.0924	0.164	-0.56	0.57
13	Male dummy (one stop–same airline)	-0.00688	0.184	-0.04	0.97
14	Male dummy (one stop–multiple airlines)	0.0762	0.198	0.38	0.70
15	Round trip fare / income ($100/$1000)	-24.7	8.40	-2.94	0.00
16	Scale parameter, if male	0.905	0.0830	-1.15[1]	0.25

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\widehat{\beta}) &= -1639.693 \\ -2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})] &= 2310.354 \\ \rho^2 &= 0.413 \\ \bar{\rho}^2 &= 0.408 \end{aligned}$$

[1] t-test against 1

Table 6.16: Estimation results for the airline itinerary choice: different scale parameters for males and females

group. In addition, we perform a likelihood ratio test between this model and the restricted model of Table 5.4, where there is a unique scale parameter:

$$-2(-1640.525 + 1639.387) = 2.276 \tag{6.54}$$

This is under $\chi^2_{1,0.95} = 3.84$, which means that we cannot reject the hypothesis that the two models are equivalent. Consequently we keep the model of Table 5.4, which is simpler.

6.7 Prediction Tests

The tests described in the previous sections dealt with the coefficients of the utility functions. In this section we consider tests in which we examine the predicted choice probabilities.

We distinguish between two types of data that are used in prediction tests: internal and external. Internal data are derived from the same source as the estimation sample. In many cases the prediction and estimation samples are identical. With large data sets, however, it is possible to avoid an overlap between the two samples. In general, external data are not necessarily of the same type as the estimation data and may, for example, be aggregate data. We would therefore rely on one of the aggregate prediction procedures described in chapter 10. As a consequence, external data are most useful for the joint test of a system of disaggregate models combined with a particular aggregation procedure. In our discussion of tests of the disaggregate models we consider only disaggregate prediction tests in which the prediction sample may or may not overlap the estimation sample; the application of disaggregate prediction tests is essentially the same as that of the sample enumeration aggregation technique described in section 10.2.

6.7.1 Outlier Analysis

An outlier analysis is an important prediction test that should be performed at an early stage of the model development process.

We use the model with estimation results presented in Table 6.18. It is a refined version of the model presented in Table 6.8, where the male dummies have been removed. Indeed, their associated t stats (0.62 and 1.25) show that there are not significant at the 95% level. In addition, a likelihood ratio test can be performed between the models of Tables 6.8 and 6.18.

$$-2(-1635.068 + 1634.266) = 1.604 \tag{6.55}$$

Parameter number	Description[a]	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.891	0.238	-3.75	0.00
2	One stop–multiple airlines dummy	-1.31	0.248	-5.27	0.00
3	Round trip fare ($100)	-2.04	0.246	-8.30	0.00
4	Elapsed time (hours)	-0.331	0.0877	-3.77	0.00
5	Leg room (inches), if male (non stop)	0.116	0.0393	2.97	0.00
6	Leg room (inches), if female (non stop)	0.202	0.0373	5.42	0.00
7	Leg room (inches), if male (one stop)	0.128	0.0351	3.63	0.00
8	Leg room (inches), if female (one stop)	0.103	0.0313	3.29	0.00
9	Being early (hours)	-0.169	0.0257	-6.56	0.00
10	Being late (hours)	-0.109	0.0208	-5.25	0.00
11	More than 2 air trips per year (one stop–same airline)	-0.474	0.223	-2.12	0.03
12	More than 2 air trips per year (one stop–multiple airlines)	-0.258	0.231	-1.12	0.26
13	Male dummy (one stop–same airline)	0.112	0.148	0.75	0.45
14	Male dummy (one stop–multiple airlines)	0.206	0.160	1.28	0.20
15	Round trip fare / income ($100/$1000)	-24.3	8.84	-2.75	0.01
16	Scale parameter if frequent flyer	0.872	0.0907	-1.42[1]	0.16

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\widehat{\beta}) &= -1639.387 \\ -2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})] &= 2310.965 \\ \rho^2 &= 0.413 \\ \bar{\rho}^2 &= 0.408 \end{aligned}$$

Table 6.17: Estimation results for the airline itinerary choice: different scale parameters for frequent and non frequent flyers

[1] t-test against 1

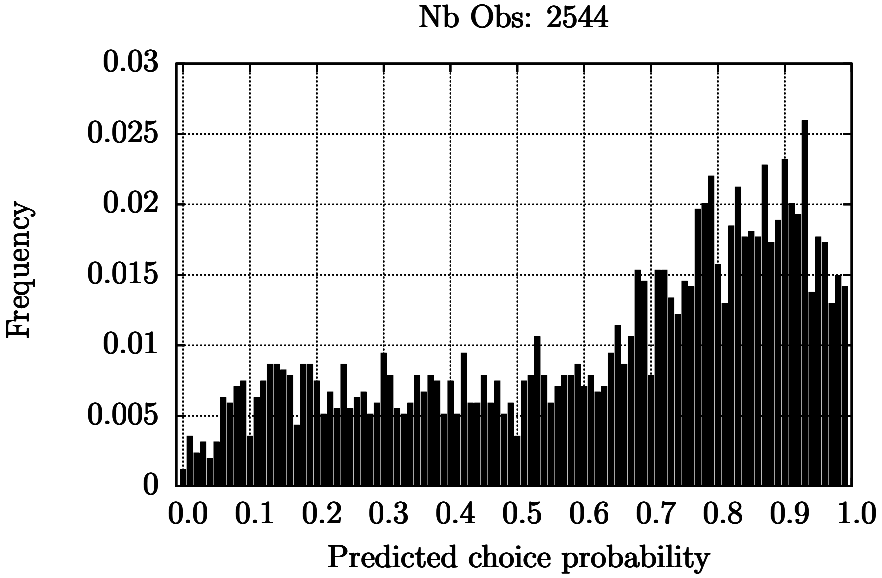


Figure 6.4: Distribution of the predicted choice probabilities for the final logit model

which is under $\chi^2_{2,0.05} = 5.99$, so we prefer the model of Table 6.18.

For this model and for all the observations in the estimation sample, we calculate the predicted choice probability of the chosen alternative. The distribution of these values is plotted in Figure 6.4. The distribution is shifted on the right, which is a good sign. The model predicts lots of high probabilities for the chosen alternatives. For checking outliers, we focus on the left part of the graph, which correspond to low predicted choice probabilities.

We check for unusually large deviations by inspecting all the observations with predicted probabilities less than some arbitrary small value. We begin this analysis at a low limit of, say, 0.01 (or even 0.001 for models with very large choice sets). If we decide to continue the search for less serious outliers we may increase this limit to 0.05, and so on. *Under no circumstances should one simply throw out of the estimation sample these observations without further analysis.* We first check these observations for data errors. We may find, for example, that the chosen alternative was simply miscoded or uncover other coding and measurement errors. If errors are not found, the observation is then classified as an outlier.

The next step, after all the uncovered errors are corrected, is to test the sensitivity of the estimation results to the presence of outliers. If one finds a high level of sensitivity, one can conclude that the outliers contain important information about the model. One could search for improvements to the

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.898	0.218	-4.13	0.00
2	One stop–multiple airlines dummy	-1.24	0.223	-5.58	0.00
3	Round trip fare ($100)	-1.81	0.152	-11.90	0.00
4	Elapsed time (0 - 2 hours)	-0.856	0.226	-3.79	0.00
5	Elapsed time (2 - 8 hours)	-0.241	0.0820	-2.93	0.00
6	Elapsed time (> 8 hours)	-0.936	0.314	-2.99	0.00
7	Leg room (inches), if male (non stop)	0.0972	0.0330	2.94	0.00
8	Leg room (inches), if female (non stop)	0.193	0.0315	6.15	0.00
9	Leg room (inches), if male (one stop)	0.128	0.0290	4.42	0.00
10	Leg room (inches), if female (one stop)	0.0845	0.0259	3.26	0.00
11	Being early (hours)	-0.150	0.0190	-7.89	0.00
12	Being late (hours)	-0.0993	0.0167	-5.94	0.00
13	More than 2 air trips per year (one stop–same airline)	-0.279	0.141	-1.98	0.05
14	More than 2 air trips per year (one stop–multiple airlines)	-0.0670	0.157	-0.43	0.67
15	Round trip fare / income ($100/$1000)	-23.0	8.11	-2.83	0.00

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\widehat{\beta}) &= -1635.068 \\ -2[\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})] &= 2319.603 \\ \rho^2 &= 0.415 \\ \bar{\rho}^2 &= 0.410 \end{aligned}$$

Table 6.18: Estimation results for the airline itinerary choice: preferred model

model specification that will increase the predicted probability of the chosen alternative for some of the outlier observations. It will often be the case that outlier analysis suggests specification improvements that not only result in a better fit for the outliers but also for all observations. Note that because the model is probabilistic some outliers should exist in the data and the goal is not to eliminate them entirely.

6.7.2 Market Segment Prediction Tests

The likelihood ratio index is a useful measure to compare the goodness of fit of alternative specifications. It does not, however, convey direct information about the differences between predicted probabilities and sample frequencies. The purpose of market segment prediction tests is to examine the goodness of fit of a model by its ability to replicate observed shares of alternatives for a market segment.

Denote the number of observations in market segment g choosing alternative i as N_{gi}, where

$$\sum_{i \in \mathcal{C}} N_{gi} = N_g. \tag{6.56}$$

We compare N_{gi}, or the share N_{gi}/N_g, with the prediction given by

$$\sum_{n=1}^{N_g} P_n(i) \qquad \text{or} \qquad \frac{1}{N_g} \sum_{n=1}^{N_g} P_n(i),$$

respectively.

Recall from section 5.6 that for the estimation sample in a logit model with a full set of alternative-specific constants, the predicted shares are equal to the observed shares, as follows:

$$\sum_{n=1}^{N_g} P_n(i) = N_i, \qquad i \in \mathcal{C}, \tag{6.57}$$

where N_i is the number of observations in the estimation sample choosing alternative i. However, this condition does not apply in general to subsets of the estimation sample. It does hold for market segments that have a full set of alternative-specific market segment dummy variables.

The output of a market segment prediction test is in the format shown in Table 6.19. The cell for alternative i (or alternative group i) and market segment g compares the number of observations in the subsample for that

market segment choosing alternative i, N_{gi}, with its predicted value,

$$\sum_{n=1}^{N_g} P_n(i).$$

To make these comparisons more meaningful, it is useful to compute the standard errors of the observed and predicted values. The variance of a predicted choice probability depends on the covariance matrix of the coefficient estimates and on the predicted choice probabilities (see the more detailed derivations in Horowitz 1979). It requires complex calculations and therefore is usually not performed. Horowitz (1983b) develops and shows how to use a formal statistical test to compare predicted and observed market segment choice frequencies, but again its use is not straightforward.

Using the binomial distribution, however, it is straightforward to evaluate an approximate variance of the observed values due to sampling errors as follows:

$$\text{Var}(N_{gi}) \cong N_g \left(\frac{N_{gi}}{N_g}\right)\left(1 - \frac{N_{gi}}{N_g}\right) = N_{gi}\left(1 - \frac{N_{gi}}{N_g}\right), \tag{6.58}$$

or, for the observed share

$$\text{Var}(N_{gi}) \cong \frac{1}{N_g}\left(\frac{N_{gi}}{N_g}\right)\left(1 - \frac{N_{gi}}{N_g}\right) = \frac{N_{gi}}{N_g^2}\left(1 - \frac{N_{gi}}{N_g}\right), \tag{6.59}$$

The disaggregate prediction table is examined to detect systematic trends of over and under predictions. As a simple rule of thumb we can begin the evaluation of this table by focusing on the cells in which the deviations exceed two standard errors. Significant deviations are indicative of specification errors. Therefore we attempt to correct the specification of the model to eliminate these prediction errors. This is similar to the procedure that was discussed in the market segmentation test of taste variations. Note, however, that in the prediction test we can employ a larger number of market segments and the market segment subsamples can be smaller because they are not used to estimate separate models. A small subsample will be reflected in large standard errors of the observed shares, which are inversely proportional to the square root of the sample size. In many studies this prediction test has proved to be a very valuable aid in improving the specification of discrete choice models.

To demonstrate a specific prediction test, we use the example of the airline itinerary choice. We focus on the model of Table 6.18, and we consider two

Alternative or groups of alternatives		Market segment 1	2	⋯	g	⋯	G	Total
1	Predicted	$\sum_{n=1}^{N_1} P_n(1)$	$\sum_{n=1}^{N_2} P_n(1)$	…	$\sum_{n=1}^{N_g} P_n(1)$	…	$\sum_{n=1}^{N_G} P_n(1)$	$\sum_{n=1}^{N} P_n(1)$
	Observed	N_{11}	N_{21}		N_{g1}		N_{G1}	N_1
2	Predicted	$\sum_{n=1}^{N_1} P_n(2)$	$\sum_{n=1}^{N_2} P_n(2)$	…	$\sum_{n=1}^{N_g} P_n(2)$	…	$\sum_{n=1}^{N_G} P_n(2)$	$\sum_{n=1}^{N} P_n(2)$
	Observed	N_{12}	N_{22}		N_{g2}		N_{G2}	N_2
	⋮	⋮	⋮		⋮		⋮	⋮
i	Predicted	$\sum_{n=1}^{N_1} P_n(i)$	$\sum_{n=1}^{N_2} P_n(i)$	…	$\sum_{n=1}^{N_g} P_n(i)$	…	$\sum_{n=1}^{N_G} P_n(i)$	$\sum_{n=1}^{N} P_n(i)$
	Observed	N_{1i}	N_{2i}		N_{gi}		N_{Gi}	N_i
	⋮	⋮	⋮		⋮		⋮	⋮
J	Predicted	$\sum_{n=1}^{N_1} P_n(J)$	$\sum_{n=1}^{N_2} P_n(J)$	…	$\sum_{n=1}^{N_g} P_n(J)$	…	$\sum_{n=1}^{N_G} P_n(J)$	$\sum_{n=1}^{N} P_n(J)$
	Observed	N_{1J}	N_{2J}		N_{gJ}		N_{GJ}	N_J

Note: In a policy prediction table, "predicted" is replaced with "predicted under the policy" and "observed" is replaced with "predicted base."

Table 6.19: The format of a disaggregate prediction table

market segments: males and females. Table 6.20 shows the number of observations per market segment and alternatives. Table 6.21 displays the model predictions. The approximate standard errors of the observed values due to sampling errors are calculated using the equation (6.58) and are presented in Table 6.22. The intervals of acceptability for the predictions are presented in Table 6.23 (between -2 and 2 standard errors around N_{gi}). Regarding Tables 6.21 and 6.23, we conclude that the predictions are acceptable, as no prediction is out of the acceptability interval.

	Male	Female
Non stop	757	941
One stop, same airline	214	231
One stop multiple airlines	205	196

Table 6.20: Number of observations per gender and per alternative, for the airline itinerary choice

	Male	Female
Non stop	766.45	931.55
One stop, same airline	212.95	232.05
One stop multiple airlines	196.6	204.4

Table 6.21: Prediction of the model per gender and per alternative, for the airline itinerary choice

	Male	Female
Non stop	22.06	25.43
One stop, same airline	6.23	6.23
One stop multiple airlines	5.96	5.29

Table 6.22: Approximate standard errors per gender of the observed values due to sampling errors, for the airline itinerary choice

Another test is performed, where we focus on the market segmentation by flight direction. Three directions are considered in the data, and they constitute the three market segments: East-West, West-East and North-South. Table 6.24 shows the number of observations per market segments. Table 6.25 displays the model predictions. The approximate standard errors of

	Male	Female
Non stop	712.88 - 801.12	890.14 - 991.86
One stop, same airline	201.55 - 226.45	218.54 - 243.46
One stop multiple airlines	193.07 - 216.93	185.43 - 206.57

Table 6.23: Intervals of acceptability for the predictions per gender, for the airline itinerary choice

the observed values due to sampling errors are presented in Table 6.26. The intervals of acceptability for the predictions are presented in Table 6.27. Regarding Tables 6.25 and 6.27, the predictions are satisfactory, except for the cells "One stop, same airline"- "East-West", and "One stop, same airline"- "West- East", because the predictions are out of the acceptability interval.

	East-West	West-East	North-South
Non stop	698	453	547
One stop, same airline	229	75	141
One stop multiple airlines	173	75	153

Table 6.24: Number of observations per flight direction and per alternative, for the airline itinerary choice

	East-West	West-East	North-South
Non stop	708.38	444.39	545.23
One stop, same airline	207.64	86.49	150.86
One stop multiple airlines	183.98	72.12	144.91

Table 6.25: Prediction of the model per flight direction and per alternative, for the airline itinerary choice

6.7.3 Validation sample

The validation consists in studying the prediction of a model on external data. Depending on the application, external data exists or not. Most of the time, the analyst has access to only one data set, and he uses it to develop the model. The artificial way for creating external data, is to divide the data set into two parts: one internal, and one external. The final specification

	East-West	West-East	North-South
Non stop	21.34	18.08	18.78
One stop, same airline	6.25	3.50	5.18
One stop multiple airlines	5.53	2.92	4.98

Table 6.26: Approximate standard errors of the observed values due to sampling errors (flight direction), for the airline itinerary choice

	East-West	West-East	North-South
Non stop	655.31 - 740.69	416.85 - 489.15	509.43 - 584.57
One stop, same airline	216.51 - 241.49	68.00 - 82.00	130.63 - 151.37
One stop multiple airlines	161.94 - 184.06	69.17 - 80.83	143.04 - 162.96

Table 6.27: Intervals of acceptability for the predictions (flight direction), for the airline itinerary choice

(developed on the entire data) is estimated on the internal part, and applied on the external part. Predictions of the model on the external part are checked carefully as described in the section dealing with the outliers analysis.

We propose a method where the original data are divided into two parts, each containing the same number of observations. The final model specification is estimated on one part and applied on the remaining part. Consequently, two experiences are conducted. For each data set, estimation results of one experience, and simulation results of the other experience are compared using the adjusted likelihood ratio index, presented in Section 6.5.7. If estimation and prediction results are equivalent by data sets, the model is considered valid.

We illustrate the method with the example of the airline itinerary choice. The logit model of Table 5.3 is used. We propose to divide the original data in two data sets, which contain half of the observations. The repartition of the observations across the data sets is done arbitrarily. Table 6.28 shows the number of observations by data sets.

	Original data set	data set 1	data set 2
Obs. Nb.	3609	1804	1805

Table 6.28: Repartition of the observations across the data sets, for validation

	data set 1	data set 2
M_1	-819.897 / 0.399	-836.978 / 0.398
M_2	-829.760 / 0.392	-828.001 / 0.405

Table 6.29: Estimation and simulation results of the validation (log likelihood / $\bar{\rho}^2$)

6.7.4 Policy Forecasting Tests

The purpose of developing a model is of course forecasting. We are principally interested in the use of a model for incremental forecasts of the effects of policy changes. This is usually done using the aggregate point elasticities described in Section 5.3. The validity of point elasticities, however, is limited to small changes in the variables. Therefore, to test the response predicted by the model to large changes, we perform disaggregate predictions. The difference from the previous section is that in a forecasting test we first modify the values of one or more explanatory variables in accordance with the policy being tested.

The disaggregate forecasting test is an application of the sample enumeration forecasting procedure described in Section 10.2. Its use during the model development process is to check the policy sensitivity of the model. We examine the forecasts and consider whether or not they are reasonable. If not, we consider the potential underlying causes of an unexpected forecast. If an unreasonable forecast can be explained by model specification or data errors, we attempt to correct them and repeat this test.

The usefulness of a forecasting test heavily depends on the availability of prior information that can be used to determine the ranges of "reasonable forecasts." The most useful information is derived from "before and after" studies which are conducted on occasions of real changes in policy or environmental factors and which collect data on the system variables before and after changes occur. An in-depth study may also collect two samples, one before and one after, which could be used to test for stability of the coefficients.

We present a prediction example with the choice of airline itinerary. We consider the logit model of Table 6.18. The influence of an increase of the travel fare is tested. A simple hypothesis is tested, where the fare of the "Non stop" alternative is increased by 10%. The estimated model of Table 6.18 is applied on such modified data. The predicted shares by alternatives are presented in Table 6.30. The "Predicted base" row refers to the counting made in the original estimation data. The "Prediction made on the original

data" row is obtained by applying the model of Table 6.18 on the original estimation data. The last row is obtained by applying the model of Table 6.18 on the modified data where the round trip fare of the alternative "Non stop" has been multiplied by 1.1. Note that values of the first and second rows are mostly the same. This is logical because the model contains alternative specific constants (β_1 and β_2). The model is able to reproduce exactly the shares of the estimation data. When we look at the last row, we see a decrease of the number of people choosing the alternative "Non stop", which is logical as its fare has increased. The decrease is translated in an increase of the shares associated with the alternatives "One stop"

6.8 Summary

The process of model building, which involves a great amount of judgment, is nevertheless partially susceptible to rigorous statistical procedures. We have outlined a number of procedures (statistical or otherwise) that we have found most useful for model development.

Three major categories of tests are covered. First, the model structure is taken as given, and we present formal and informal specification tests for the utility functions. Second, we no longer assume the model structure is given, and we show how to test for violation of the IIA assumption of the logit model for the presence of taste variation in the population, as well as for heteroscedasticity (unequal variances) in the utility functions. Finally, we describe tests of model predictions and outliers that can be used once a model specification seems reasonably satisfactory.

It is important to realize that the model-building process is an iterative one. The process is begun with a set of a priori assumptions by the analyst, which may be subsequently revised as the analyst learns about the choice process in the model's development. The tests and procedures we have presented here represent a practical approach to systematizing, to the extent possible, the model-building process.

	Non stop	One stop, same airline	One stop, multiple airlines
Predicted base	1698	445	401
Prediction made on the original data	1697.99	444.01	401.00
Predicted under the policy (1.1 travel fare)	1403.79	598.73	541.48

Table 6.30: Observed shares; prediction of the model on the original data;

Table	Description	K	$\mathcal{L}(\widehat{\beta})$	$\bar{\rho}^2$
5.4	Base model	15	-1640.525	0.408
6.2	Alternative specific elapsed time coefficients, generic leg room coefficients	15	-1641.932	0.407
6.1	Generic elapsed time coefficients	13	-1642.796	0.408
5.3	Generic leg room coefficient, no interaction between cost and income	12	-1652.573	0.404
6.3	Square of scheduled delays	15	-1649.407	0.404
6.5	J-test, first model	16	-1640.493	0.407
6.6	J-test, second model	16	-1640.492	0.407
6.7	Piecewise linear elapsed time	18	-1634.131	0.409
6.8	Piecewise linear elapsed time, parsimonious specification	17	-1634.266	0.409
6.9	Polynomial function of elapsed time	17	-1635.347	0.409
6.10	Box-Cox transform of elapsed time	16	-1639.317	0.408
6.12	McFadden's omitted variables test	18	-1611.816	0.417
6.13	Larger sample: leisure and non leisure trips	15	-2300.453	0.416
6.14	Larger sample: segment specific model	30	-2269.605	0.420
6.16	Gender specific variance	16	-1639.693	0.408
6.17	Frequent flyer specific variance	16	-1639.387	0.408
6.18	Final model	15	-1635.068	0.410

Table 6.31: Summary of the models discussed in the chapter

Chapter 7

The Nested Logit model

Contents

The logit model has been derived based on the assumption that the error terms ε_{in}, associated with alternative i and individual n, are i.i.d. across i and n. Although mathematically convenient, this assumption is invalid in some situations. In this chapter, we relax this assumption, while keeping as much as possible the convenient properties of logit models.

The red bus/blue bus example described in Section 3.7 illustrates the potential counter-intuitive results obtained from a logit model when the error terms are not independent across alternatives. In practice, the correlation between the error terms is usually not as strong as in this simplistic example, but in many cases, independence may not be assumed due to alternatives sharing unobserved attributes. A more realistic transportation mode choice example would be a choice between 'train', 'bus' and 'car', where alternatives 'train' and 'bus' share attributes related to public transportation alternatives, some of which being likely to be unobserved.

In the airline itinerary example developed in previous chapters, the alternatives "One stop, same airline" and "One stop, multiple airlines" share

all attributes associated with the fact that there is one stop, irrespectively of the airline. If some of these attributes are unobserved, the error terms of these two alternatives are correlated, and the logit model may also generate counter-intuitive results.

The *nested logit* (NL) model is designed to explicitly capture the presence of shared unobserved attributes. The basic idea is to partition the universal choice set $\mathcal{C}$ into M mutually exclusive and collectively exhaustive subsets, called *nests*, denoted by $\mathcal{C}_1, \ldots, \mathcal{C}_M$. Each alternative belongs to one and only one nest. That is

$$\mathcal{C} = \bigcup_{m=1}^{M} \mathcal{C}_m,$$

and

$$\mathcal{C}_m \cap \mathcal{C}_\ell = \emptyset, \quad \forall m \neq \ell.$$

For each individual n, the choice set $\mathcal{C}_n$ is partitioned into nests $\mathcal{C}_{1n}, \ldots, \mathcal{C}_{Mn}$, where $\mathcal{C}_{mn}$ is the intersection between the nest $\mathcal{C}_m$ and the individual specific choice set $\mathcal{C}_n$.

The partitioning must be designed such that alternatives sharing unobserved attributes belong to the same nest. Sometimes, there is no satisfactory partitioning of the choice set, and more complex models are required. Some of those models are described in Chapters 8 and 13.

Once the nesting structure has been defined, the utility of alternative i in nest $\mathcal{C}_m$ can be written as

$$U_{in} = V_{in} + \varepsilon_{in} = V_{in} + \varepsilon_{mn} + \varepsilon_{imn} \tag{7.1}$$

where the error term ε_{in} is explicitly written as the sum of two terms or *components*. The first, ε_{mn}, is nest specific, and is the same for all alternatives in the nest. The second, ε_{imn}, is alternative specific. ε_{mn} captures the unobserved attributes shared by alternatives in nest m, while ε_{imn} captures unobserved attributes specific to alternative i.

In the next section, we present an intuitive derivation of the nested logit model for the airline itinerary example. The formal derivation using specification (7.1) is provided in Section 7.2. Section 7.3 discusses the estimation of the model parameters. The example is revisited in Section 7.4. Section 7.5 presents a generalization of the nested logit model involving several layers of nesting.

7.1 Illustration

We introduce the nested logit model with the airline itinerary example. We consider the same specification as the logit model presented in Table 6.18. In

order to capture the possible correlation between the error terms of alternatives "One stop, same airline" and "One stop, multiple airlines", we partition the choice set into two nests. The first nest, labeled "Non stop", contains only the "Non stop" (NS) alternative. The second nest, labeled "One stop", contains the two alternatives involving one stop: "One stop, same airline" (SAME) and "One stop, multiple airlines" (MULT). The nesting structure is pictured in Figure 7.1.

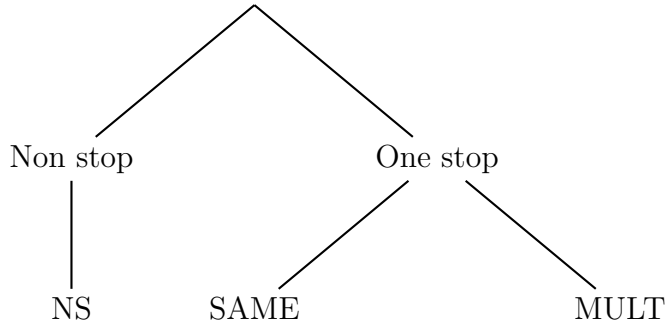


Figure 7.1: Tree structure for the airline itinerary example

We exploit the nesting structure by decomposing the analysis into two parts:

- at the upper level, the *marginal* probabilities

  $$\Pr(\text{Non stop}|\{\text{Non stop}, \text{One stop}\}) \text{ and } \Pr(\text{One stop}|\{\text{Non stop}, \text{One stop}\})$$

  of choosing a nest that is, of choosing any alternative within this nest, and,

- at the lower level, the *conditional* probabilities

  $$\Pr(\text{NS}|\text{Non stop}), \Pr(\text{SAME}|\text{One stop}) \text{ and } \Pr(\text{MULT}|\text{One stop}),$$

  of choosing the alternative, given that the nest is chosen. Note that, in this case, alternative NS is the only one in the nest "Non stop", and $\Pr(\text{NS}|\text{Non stop})$ is trivially equal to 1.

The probability of interest, that is the probability that a given alternative is chosen, will then be obtained by the product of the marginal and the conditional probabilities, that is

$$\begin{aligned} \Pr(\text{NS}) &= \Pr(\text{NS}|\text{Non stop}) \Pr(\text{Non stop}|\{\text{Non stop}, \text{One stop}\}), \\ \Pr(\text{SAME}) &= \Pr(\text{SAME}|\text{One stop}) \Pr(\text{One stop}|\{\text{Non stop}, \text{One stop}\}), \\ \Pr(\text{MULT}) &= \Pr(\text{MULT}|\text{One stop}) \Pr(\text{One stop}|\{\text{Non stop}, \text{One stop}\}). \end{aligned} \tag{7.2}$$

We first consider the conditional probabilities at the lower level. We focus on the nest "One stop", and model the choice between the "One stop, same airline" (SAME) and the "One stop, multiple airlines" (MULT) alternatives. It is a binary choice. To estimate the parameters of this model, we consider only the 846 observations in the sample where the chosen alternative is one of the "One stop" alternatives. We assign a random utility function to each of the two alternatives, composed of a systematic part and a random part. The systematic parts are the same as for the logit model (see Table 6.18), with minor modifications due to normalization. First, it is not possible to estimate both constants associated with the "One stop, same airline" and "One stop, multiple airlines" dummies. The latter has been normalized to zero. For the same reason, the coefficient of the variable "More than two air trips per year (one stop, multiple airlines)" has also been normalized to 0. Finally, looking at the data, we note that the elapsed time for "One stop" alternative exceeds two hours for each of the 846 observations. Therefore, the coefficient of the variable "Elapsed time (0–2 hours)" cannot be identified.

In binary choice model, all attributes, observed or not, shared by the two "One stop" alternatives affect the two utility functions in the exact same way and, consequently, cancel out and are not included. As a consequence, it is appropriate to estimate a binary logit model, based on the assumptions that the error terms capturing the other unobserved attributes are i.i.d. extreme value. Therefore, we assume error terms to be i.i.d. extreme value, with scale parameter $\mu_{\text{One stop}}$, and we obtain a binary logit model:

$$\Pr(\text{SAME}|\text{One stop}) = \frac{e^{\mu_{\text{One stop}} V_{\text{SAME}}}}{e^{\mu_{\text{One stop}} V_{\text{SAME}}} + e^{\mu_{\text{One stop}} V_{\text{MULT}}}}. \tag{7.3}$$

The results are reported in Table 7.1, where the numbering of the parameters has been maintained to be consistent with Table 6.18.

We now focus on the upper level of the tree represented in Figure 7.1, and model the choice between the "Non stop" alternative and the aggregate alternative "One stop" corresponding to the nest. It is again a binary choice. We use the same framework as before and assign a random utility function to each of the two alternatives (that is, for each of the two nests), composed of a systematic part and a random part. The systematic part associated with the "Non stop" alternative is the same as for the logit model (see Table 6.18), with minor modifications due to normalization. Indeed, the normalization of the alternative specific constant and the "More than 2 air trips per year" interactions has been adjusted to accommodate the alternatives involved in the binary choice. The coefficients 0 and 12' are replacing coefficients 2 and 14 in the logit specification. Also, the coefficients that have already been estimated at the lower level are not re-estimated. The estimated values (see

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop, same airline dummy	0.469	0.188	2.50	0.01
3	Round trip fare ($100)	-2.87	0.624	-4.61	0.00
5	Elapsed time (2–8 hours)	-0.387	0.136	-2.84	0.00
6	Elapsed time (> 8 hours)	-2.33	0.759	-3.07	0.00
9	Leg room (inches), if male (one stop)	0.170	0.0464	3.65	0.00
10	Leg room (inches), if female (one stop)	0.104	0.0421	2.46	0.01
11	Being early (hours)	-0.250	0.0422	-5.91	0.00
12	Being late (hours)	-0.0942	0.0286	-3.29	0.00
13	More than two air trips per year (one stop, same airline)	-0.220	0.218	-1.01	0.31
15	Round trip fare / income ($100/$1000)	-37.8	40.8	-0.93	0.35

Summary statistics

Number of observations = 846

$$\mathcal{L}(0) = -586.403$$
$$\mathcal{L}(c) = -585.258$$
$$\mathcal{L}(\hat{\beta}) = -318.994$$
$$-2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] = 534.816$$
$$\rho^2 = 0.456$$
$$\bar{\rho}^2 = 0.439$$

Table 7.1: Conditional choice of “One stop” alternatives

Table 7.1) are used instead. Therefore, only 5 coefficients are estimated at the upper level.

For the "One stop" alternative, in order to be consistent with utility maximization, we should use the utility of "One stop, same airline", or the utility of "One stop, multiple airlines", depending on which one is the largest. Note that it would not be valid to consider the largest of the two deterministic parts, as it would ignore the error terms containing unobserved attributes. Instead, we consider the expected value of the random variable representing the largest utility, that is

$$\tilde{V}_{\text{One stop}} = \text{E}[\max(U_{\text{SAME}}, U_{\text{MULT}})]. \tag{7.4}$$

As described in Section 5.3.3, $\tilde{V}_{\text{One stop}}$ is a scalar summary of the expected "worth" of the nest containing the two alternatives. Because we have assumed a logit model within the nest, the utility functions of these two alternatives are extreme value distributed, and the expected maximum utility is given by the logsum formula (4.83). We obtain that

$$\begin{aligned} \tilde{V}_{\text{One stop}} &= \text{E}[\max(U_{\text{SAME}}, U_{\text{MULT}})] \\ &= \frac{1}{\mu_{\text{One stop}}} \log(e^{\mu_{\text{One stop}} V_{\text{SAME}}} + e^{\mu_{\text{One stop}} V_{\text{MULT}}}). \end{aligned} \tag{7.5}$$

The scale parameter $\mu_{\text{One stop}}$ has been normalized to 1, and

$$\tilde{V}_{\text{One stop}} = \log(e^{V_{\text{SAME}}} + e^{V_{\text{MULT}}}). \tag{7.6}$$

Note that no unknown parameter is involved in this specification.

With respect to the error terms of the two alternatives at the upper level, we assume that they are i.i.d. extreme value, with scale parameter μ. Therefore the choice between "Non stop" and "One stop" is modeled by a binary logit model:

$$\Pr(\text{One stop}|\{\text{Non stop}, \text{One stop}\}) = \frac{e^{\mu \tilde{V}_{\text{One stop}}}}{e^{\mu V_{\text{NS}}} + e^{\mu \tilde{V}_{\text{One stop}}}}. \tag{7.7}$$

As the scale parameter $\mu_{\text{One stop}}$ has been normalized to 1 at the lower level, the scale parameter μ is identified at the upper level. The estimation results are reported in Table 7.2.

The results of the estimation at the lower level (Table 7.1) and at the upper level (Table 7.2) have been gathered in Table 7.3. The log likelihood of the joint model is simply the sum of the log likelihood of each model. In comparison with the logit model, it can be seen that one additional parameter

(the scale parameter μ) has been estimated. The log likelihood function has increased from -1635.068 up to -1617.129. Using a likelihood ratio test, the hypothesis that the logit model is correct can be rejected at the 95% level, as the statistics $-2(-1635.07 + 1617.13) = 35.378$ is above $\chi^2_{1,0.95} = 3.84$.

This procedure to estimate the nested logit model is called *sequential estimation*. The estimates can be shown to be consistent, but they are not efficient. Another procedure, called *simultaneous estimation* or *full information maximum likelihood estimation*, is described later in this chapter, and is both consistent and efficient. The parameter estimates for the simultaneous estimation are reported in Table 7.4.

7.2 Derivation

The example in the previous section introduced the nested logit model and its sequential and joint estimation. We now provide a formal derivation of the model.

Given a partition of $\mathcal{C}$ into M nests $\mathcal{C}_1, \ldots, \mathcal{C}_M$, and an individual n with choice set $\mathcal{C}_n \subseteq \mathcal{C}$, we denote by $\mathcal{C}_{1n}, \ldots, \mathcal{C}_{Mn}$ the nests for individual n, defined as $\mathcal{C}_{mn} = \mathcal{C}_m \cap \mathcal{C}_n$. The derivation of the nested logit model is obtained from the following definition of marginal probability:

$$P_n(i|\mathcal{C}_n) = \sum_{m=1}^{M} \Pr(i|\mathcal{C}_{mn}, \mathcal{C}_n) \Pr(\mathcal{C}_{mn}|\mathcal{C}_n), \tag{7.8}$$

where $\Pr(i|\mathcal{C}_{mn}, \mathcal{C}_n)$ is the probability for individual n to select alternative i within nest $\mathcal{C}_{mn}$, and $\Pr(\mathcal{C}_{mn}|\mathcal{C}_n)$ is the probability to select an alternative in the nest $\mathcal{C}_{mn}$. The definition of the nest structure guarantees that only one term in (7.8) is nonzero. Indeed, each alternative belongs to exactly one nest. Consequently, $\Pr(i|\mathcal{C}_{mn}, \mathcal{C}_n) = 0$ if alternative i does not belong to nest $\mathcal{C}_{mn}$. In addition, we assume that the conditional choice in a nest is independent of the remainder of $\mathcal{C}_n$, as follows: $\Pr(i|\mathcal{C}_{mn}, \mathcal{C}_n) = \Pr(i|\mathcal{C}_{mn})$. Therefore, similarly to (7.2) in the previous example, we can write[1]

$$P_n(i|\mathcal{C}_n) = \Pr(i|\mathcal{C}_{mn}) \Pr(\mathcal{C}_{mn}|\mathcal{C}_n), \tag{7.9}$$

[1]Because of the decomposition (7.9), the nested logit model has been sometimes interpreted as capturing a *sequence* of choices, where the nest is chosen first, and the alternative within the nest is chosen afterward. We emphasize that no additional behavioral assumption, such as a chronology in the choices, is made in specifying the nested logit model. The derivation of the model from the random utility model is purely based on the properties of the error terms of the random utilities.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
0	Non stop dummy	4.90	0.817	6.00	0.00
4	Elapsed time (0–2 hours)	-1.60	0.405	-3.95	0.00
7	Leg room (inches), if male (non stop)	0.170	0.0584	2.91	0.00
8	Leg room (inches), if female (non stop)	0.338	0.0565	5.98	0.00
12'	More than 2 air trips per year (non stop)	0.219	0.215	1.02	0.31
16	μ	0.526	0.0307	-15.42[1]	0.00

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -1763.366 \\ \mathcal{L}(c) &= -1617.902 \\ \mathcal{L}(\hat{\beta}) &= -1298.135 \\ -2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 930.463 \\ \rho^2 &= 0.264 \\ \bar{\rho}^2 &= 0.260 \end{aligned}$$

[1] t-test against 1

Table 7.2: Nested logit model: binary choice between "Non stop" and "One stop

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
0	Non stop dummy	4.90	0.817	6.00	0.00
1	One stop, same airline dummy	0.469	0.188	2.50	0.01
3	Round trip fare ($100)	-2.87	0.624	-4.61	0.00
4	Elapsed time (0–2 hours)	-1.60	0.405	-3.95	0.00
5	Elapsed time (2–8 hours)	-0.387	0.136	-2.84	0.00
6	Elapsed time (> 8 hours)	-2.33	0.759	-3.07	0.00
7	Leg room (inches), if male (non stop)	0.170	0.0584	2.91	0.00
8	Leg room (inches), if female (non stop)	0.338	0.0565	5.98	0.00
9	Leg room (inches), if male (one stop)	0.170	0.0464	3.65	0.00
10	Leg room (inches), if female (one stop)	0.104	0.0421	2.46	0.01
11	Being early (hours)	-0.250	0.0422	-5.91	0.00
12	Being late (hours)	-0.0942	0.0286	-3.29	0.00
12'	More than 2 air trips per year (non stop)	0.219	0.215	1.02	0.31
13	More than two air trips per year (one stop, same airline)	-0.220	0.218	-1.01	0.31
15	Round trip fare / income ($100/$1000)	-37.8	40.8	-0.93	0.35
16	μ	0.526	0.0307	-15.42[1]	0.00

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2349.769 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\hat{\beta}) &= -1617.129 \\ -2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 1465.28 \\ \rho^2 &= 0.312 \\ \bar{\rho}^2 &= 0.305 \end{aligned}$$

Table 7.3: Nested logit model: sequential estimation

[1] t-test against 1

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
0	Non stop dummy	1.74	0.337	5.16	0.00
1	One stop, same airline dummy	0.437	0.183	2.39	0.02
3	Round trip fare ($100)	-2.81	0.315	-8.91	0.00
4	Elapsed time (0–2 hours)	-1.49	0.417	-3.57	0.00
5	Elapsed time (2–8 hours)	-0.348	0.112	-3.10	0.00
6	Elapsed time (> 8 hours)	-1.62	0.506	-3.21	0.00
7	Leg room (inches), if male (non stop)	0.168	0.0587	2.86	0.00
8	Leg room (inches), if female (non stop)	0.330	0.0624	5.28	0.00
9	Leg room (inches), if male (one stop)	0.175	0.0396	4.41	0.00
10	Leg room (inches), if female (one stop)	0.112	0.0344	3.25	0.00
11	Being early (hours)	-0.234	0.0338	-6.92	0.00
12	Being late (hours)	-0.135	0.0241	-5.61	0.00
12'	More than two air trips per year (non stop)	0.199	0.243	0.82	0.41
13	More than two air trips per year (one stop, same airline)	-0.237	0.210	-1.13	0.26
15	Round trip fare / income ($100/$1000)	-36.4	14.3	-2.55	0.01
16	μ	0.546	0.0595	-7.62[1]	0.00

Summary statistics

Number of observations = 2544

$$\mathcal{L}(0) = -2794.870$$
$$\mathcal{L}(c) = -2203.160$$
$$\mathcal{L}(\hat{\beta}) = -1613.858$$
$$-2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] = 2362.022$$
$$\rho^2 = 0.423$$
$$\bar{\rho}^2 = 0.417$$

Table 7.4: Nested logit model: simultaneous estimation

[1] t-test against 1

where m is the index of the (unique) nest containing alternative i.

We assume first that the nest $\mathcal{C}_{mn}$ is given and derive the model $\Pr(i|\mathcal{C}_{mn})$. For the sake of notational simplicity, we omit the index n of the individual in the following. According to the principle of utility maximization, the probability that an alternative i in nest $\mathcal{C}_m$ is chosen is

$$\begin{aligned} \Pr(i|\mathcal{C}_m) &= \Pr(U_i \geq U_j, j \in \mathcal{C}_m) && (7.10)\\ &= \Pr(V_i + \varepsilon_m + \varepsilon_{im} \geq V_j + \varepsilon_m + \varepsilon_{jm}, j \in \mathcal{C}_m) && (7.11)\\ &= \Pr(V_i + \varepsilon_{im} \geq V_j + \varepsilon_{jm}, j \in \mathcal{C}_m) && (7.12) \end{aligned}$$

where U_i is defined by (7.1). The error term of an alternative i in nest $\mathcal{C}_m$ is assumed to consist of two components: ε_m that is nest specific and ε_{im} that is specific to the alternative. The terms ε_m cancel out in (7.11) as they are identical for all alternatives in nest m.

We next assume that the error terms ε_{im} are i.i.d. EV distributed, with scale parameter μ_m. Consequently, we obtain a logit model, that is:

$$\Pr(i|\mathcal{C}_m) = \frac{e^{\mu_m V_i}}{\sum_{j\in\mathcal{C}_m} e^{\mu_m V_j}}. \tag{7.13}$$

In order to obtain $P_n(i|\mathcal{C}_n)$ as in (7.9), we next derive $\Pr(\mathcal{C}_{mn}|\mathcal{C}_n)$ (again omitting index n). According to the principle of utility maximization, the probability of selecting a nest m is the probability that the utility of the best alternative in the nest m is larger than the utility of the best alternative in any other nest, that is

$$\Pr(\mathcal{C}_m|\mathcal{C}) = \Pr(\max_{i\in\mathcal{C}_m} U_i \geq \max_{j\in\mathcal{C}_\ell} U_j, \forall \ell \neq m). \tag{7.14}$$

From (7.1), we obtain

$$\Pr(\mathcal{C}_m|\mathcal{C}) = \Pr(\varepsilon_m + \max_{i\in\mathcal{C}_m}(V_i + \varepsilon_{im}) \geq \varepsilon_\ell + \max_{j\in\mathcal{C}_\ell}(V_j + \varepsilon_{j\ell}), \forall \ell \neq m), \tag{7.15}$$

where we have again exploited the fact that ε_m is identical for all alternatives in nest m. From property 6 of extreme value distributions (see Section 4.2.2), we have

$$\max_{i\in\mathcal{C}_m}(V_i + \varepsilon_{im}) \sim \mathrm{EV}(\tilde{V}_m, \mu_m), \tag{7.16}$$

where

$$\tilde{V}_m = \frac{1}{\mu_m} \ln \sum_{i\in\mathcal{C}_m} e^{\mu_m V_i}. \tag{7.17}$$

The quantity defined by (7.17) corresponds to the systematic composite utility of nest m. Because of (7.16), it is called the *expected maximum utility* of

the nest. Given the form of (7.17), it is also sometimes called the *logsum* and is discussed in details in Section 5.3.3. The random variable $\max_{i\in\mathcal{C}_m}(V_i+\varepsilon_{im})$ is decomposed into its systematic component $\tilde{V}_m$, and an error term ε'_m, so that

$$\max_{i\in\mathcal{C}_m}(V_i+\varepsilon_{im}) = \tilde{V}_m + \varepsilon'_m, \tag{7.18}$$

where

$$\varepsilon'_m \sim \text{EV}(0,\mu_m). \tag{7.19}$$

and (7.15) becomes

$$\Pr(\mathcal{C}_m|\mathcal{C}) = \Pr(\tilde{V}_m + \varepsilon'_m + \varepsilon_m \geq \tilde{V}_\ell + \varepsilon'_\ell + \varepsilon_\ell, \forall \ell \neq m). \tag{7.20}$$

We gather the random terms and define

$$\tilde{\varepsilon}_m = \varepsilon'_m + \varepsilon_m, \tag{7.21}$$

to obtain

$$\Pr(\mathcal{C}_m|\mathcal{C}) = \Pr(\tilde{V}_m + \tilde{\varepsilon}_m \geq \tilde{V}_\ell + \tilde{\varepsilon}_\ell, \forall \ell \neq m). \tag{7.22}$$

Finally, we assume that the error terms ε_m are also independent across m and that the $\tilde{\varepsilon}_m$ are i.i.d. EV distributed with scale parameter μ to obtain:

$$\Pr(\mathcal{C}_m|\mathcal{C}) = \frac{e^{\mu\tilde{V}_m}}{\sum_{p=1}^{M} e^{\mu\tilde{V}_p}}. \tag{7.23}$$

Note that it is an indirect assumption about the distribution of ε_m. Using (7.9), (7.13) and (7.23), we obtain the nested logit choice probabilities:

$$\begin{aligned}
P(i|\mathcal{C}) &= \frac{e^{\mu_m V_i}}{\sum_{j\in\mathcal{C}_m} e^{\mu_m V_j}} \frac{e^{\mu\tilde{V}_m}}{\sum_{p=1}^{M} e^{\mu\tilde{V}_p}} \\
&= \frac{e^{\mu_m V_i}}{\sum_{j\in\mathcal{C}_m} e^{\mu_m V_j}} \frac{\exp\left(\frac{\mu}{\mu_m}\ln\sum_{\ell\in\mathcal{C}_m} e^{\mu_m V_\ell}\right)}{\sum_{p=1}^{M}\exp\left(\frac{\mu}{\mu_p}\ln\sum_{\ell\in\mathcal{C}_p} e^{\mu_p V_{\ell p}}\right)} \\
&= \frac{e^{\mu_m V_i}}{\sum_{j\in\mathcal{C}_m} e^{\mu_m V_j}} \frac{\left(\sum_{\ell\in\mathcal{C}_m} e^{\mu_m V_\ell}\right)^{\frac{\mu}{\mu_m}}}{\sum_{p=1}^{M}\left(\sum_{\ell\in\mathcal{C}_p} e^{\mu_p V_\ell}\right)^{\frac{\mu}{\mu_p}}},
\end{aligned}$$

where m is the nest containing alternative i, and the term $\tilde{V}_m$ is defined by (7.17). Incorporating back the index n of the decision-maker, we obtain the

nested logit model:

$$P_n(i|\mathcal{C}_n) = \frac{e^{\mu_m V_{in}}}{\sum_{j\in\mathcal{C}_{mn}} e^{\mu_m V_{jn}}} \frac{\left(\sum_{\ell\in\mathcal{C}_{mn}} e^{\mu_m V_{\ell n}}\right)^{\frac{\mu}{\mu_m}}}{\sum_{p=1}^{M}\left(\sum_{\ell\in\mathcal{C}_{pn}} e^{\mu_p V_{\ell n}}\right)^{\frac{\mu}{\mu_p}}}. \tag{7.24}$$

The assumptions that have been used to derive this result are

- the error term ε_{in} is decomposed into two components ε_{mn} and ε_{imn};
- for each nest m, the terms ε_{imn} are i.i.d. Extreme Value distributed with scale parameter μ_m;
- the error terms ε_{in} and ε_{jn} of alternatives belonging to different nests are independent;
- the terms ε_{mn} are independent across m;
- the terms ε_{mn} are distributed such that $\tilde{\varepsilon}_{mn} = \varepsilon_{mn} + \varepsilon'_{mn} = \varepsilon_{mn} + \max_{i\in\mathcal{C}_{mn}}(V_{in} + \varepsilon_{imn}) - \tilde{V}_{mn}$ is Extreme Value distributed with scale parameter μ.

Note that the last assumption is an indirect one, as nothing explicit is said about the distribution of ε_{mn}. A rigorous derivation starting from a distributional assumption about the combined error $\varepsilon_{imn} + \varepsilon_{mn}$ will be presented in Chapter 8.

It is sometimes useful to identify explicitly the part of the utility function which is common to all alternatives in a nest. In the red bus/blue bus example mentioned in the introduction of the chapter, most observed attributes, such as fare or travel time, would be common to the two alternatives. In this case, the systematic utility of alternative i in nest m is decomposed as

$$V_i = V_{im} + V_m, \tag{7.25}$$

where V_m is the part of the utility which is common to all alternatives in nest m. The conditional choice probability (7.13) writes

$$\Pr(i|m) = \frac{e^{\mu_m(V_{im}+V_m)}}{\sum_{j\in\mathcal{C}_m} e^{\mu_m(V_{jm}+V_m)}} = \frac{e^{\mu_m V_{im}}}{\sum_{j\in\mathcal{C}_m} e^{\mu_m V_{jm}}}, \tag{7.26}$$

as the V_m are identical across alternatives in nest m and, consequently, cancel out. For the same reason, the expected maximum utility (7.17) writes

$$\tilde{V}_m = V_m + \frac{1}{\mu_m} \ln \sum_{i\in\mathcal{C}_m} e^{\mu_m V_{im}}. \tag{7.27}$$

As a consequence, the marginal choice probability (7.23) writes

$$\Pr(m|\mathcal{C}) = \frac{e^{\mu V_m + \frac{\mu}{\mu_m} \ln \sum_{i \in \mathcal{C}_m} e^{\mu_m V_{im}}}}{\sum_{p=1}^{M} e^{\mu V_p + \frac{\mu}{\mu_p} \ln \sum_{i \in \mathcal{C}_p} e^{\mu_p V_{ip}}}}. \tag{7.28}$$

Note that the terms V_m and V_{im} are scaled differently in this expression. Putting everything together, and including back the index n of the decision-maker, we obtain the nested logit model with nest specific utility functions:

$$P_n(i|\mathcal{C}_n) = \frac{e^{\mu_m V_{imn}}}{\sum_{j \in \mathcal{C}_{mn}} e^{\mu_m V_{jmn}}} \frac{e^{\mu V_{mn} + \frac{\mu}{\mu_m} \ln \sum_{i \in \mathcal{C}_{mn}} e^{\mu_m V_{imn}}}}{\sum_{p=1}^{M} e^{\mu V_{pn} + \frac{\mu}{\mu_p} \ln \sum_{i \in \mathcal{C}_{pn}} e^{\mu_p V_{ipn}}}}, \tag{7.29}$$

where m is the (only) nest that contains alternative i, and $\mathcal{C}_{mn} = \mathcal{C}_m \cap \mathcal{C}_n$ is the set of available alternatives in $\mathcal{C}_n$ belonging to nest m.

As shown in Section 8.6, if

$$\mu \leq \mu_m, \tag{7.30}$$

or equivalently,

$$\mathrm{Var}(\tilde{\varepsilon}_m) \geq \mathrm{Var}(\varepsilon_{im}), \tag{7.31}$$

then the model is consistent with random utility theory. Also the correlation between two alternatives is

$$\mathrm{Corr}(U_i, U_j) = \begin{cases} 1 & \text{if } i = j, \\ 1 - \dfrac{\mu^2}{\mu_m^2} & \text{if } i \neq j,\ i \text{ and } j \text{ are in the same nest } m, \\ 0 & \text{otherwise,} \end{cases} \tag{7.32}$$

Note that, as a consequence of (7.30), the correlation is always non negative. Also, from (7.32), we see that the correlation matrix associated with the nested logit has a special structure. The correlation matrix is a matrix such that the entry (i, j) is $\mathrm{Corr}(U_i, U_j)$. Due to the nest structure, the matrix is block diagonal. And, within a given block, corresponding to a given nest m, all entries are equal to the same value $1 - \mu^2/\mu_m^2$ (except, of course, the elements on the diagonal that are always 1). As the variance of the error term is the same for any alternative, that is $\mathrm{Var}(\varepsilon_m + \varepsilon_{im}) = \pi^2/6\mu^2$, the

covariance is simply

$$\mathrm{Cov}(U_i, U_j) = \begin{cases} \dfrac{\pi^2}{6\mu^2} & \text{if } i = j, \\ \dfrac{\pi^2}{6\mu^2} - \dfrac{\pi^2}{6\mu_m^2} & \text{if } i \neq j,\ i \text{ and } j \text{ are in the same nest } m, \\ 0 & \text{otherwise.} \end{cases} \tag{7.33}$$

When $\mu = \mu_m$, the correlation and the covariance are equal to 0, and the two utilities are independent random variables. If it is the case for all nests, the nested logit model collapses into a logit model. It shows that the logit model is a restriction of the nested logit model. Therefore, likelihood ratio tests can be applied to test these restrictions.

As an example, the correlation matrix associated with the airline itinerary example (Figure 7.1) is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \gamma \\ 0 & \gamma & 1 \end{pmatrix},$$

where $\gamma = 1 - \mu^2/\mu_m^2$.

The parameters μ and μ_m are closely related in the model. Actually, only their ratios μ/μ_m are meaningful. Therefore one of the scale parameters must be normalized to an arbitrary value, usually one.

A model where the scale parameter μ is normalized to 1 is said to be "normalized from the top". In this case, (7.30) writes $\mu_m \geq 1$, $\forall m$. A model where one of the parameters μ_m is normalized to 1 is said to be "normalized from the bottom." Actually, as only the ratio μ/μ_m is meaningful, some model formulations involve $\theta_m = \mu/\mu_m$. Note that, in this case, (7.30) writes $\theta_m \leq 1$.

For the nested logit model, the impact of the selected normalization is not as straightforward as for the logit model. Indeed, it clearly appears from (7.24) that the utilities of alternatives in different nests may be scaled differently. If it happens that the utility functions of these alternatives share some coefficients, the estimation may be problematic if the scale is not explicitly present in the formulation. For instance, we may rewrite (7.24) by defining $\theta_m = \mu/\mu_m$ and $\tilde{\beta}_{\ell m} = \mu_m \beta_\ell$, where β_ℓ are the coefficients of the

linear-in-parameter utility function $V_i = \sum_\ell \beta_\ell x_{i\ell}$:

$$P(i|\mathcal{C}) = \frac{\exp(\sum_\ell \tilde{\beta}_{\ell m} x_{i\ell})}{\sum_{j\in\mathcal{C}_m} \exp(\sum_\ell \tilde{\beta}_{\ell m} x_{j\ell})} \frac{\exp\left(\theta_m \ln \sum_{i\in\mathcal{C}_m} \exp(\sum_\ell \tilde{\beta}_{\ell m} x_{i\ell})\right)}{\sum_{p=1}^{M} \exp\left(\theta_p \ln \sum_{i\in\mathcal{C}_p} \exp(\sum_\ell \tilde{\beta}_{\ell p} x_{i\ell})\right)}. \tag{7.34}$$

If the set of β coefficients is different for each nest, the above formulation can be simplified by defining $\bar{V}_i = \mu_m V_i = \sum_\ell \tilde{\beta}_{\ell m} x_{i\ell}$, so that

$$P(i|\mathcal{C}) = \frac{e^{\bar{V}_i}}{\sum_{j\in\mathcal{C}_m} e^{\bar{V}_j}} \frac{\exp\left(\theta_m \ln \sum_{j\in\mathcal{C}_m} e^{\bar{V}_j}\right)}{\sum_\ell \exp\left(\theta_\ell \ln \sum_{j\in\mathcal{C}_\ell} e^{\bar{V}_j}\right)}. \tag{7.35}$$

This formulation simplifies the computation of the derivatives, needed by parameter estimation procedures (see Section 7.3). For this reason, it has been adopted in various estimation procedures (e.g. Daly, 1987). We emphasize here that this simplification is not valid when the same parameters are involved in the utility function of alternatives belonging to different nests. Indeed, if $\beta_\ell = \beta_p$, β_ℓ appears in the utility of an alternative in nest m, and β_p in the utility of an alternative in nest k, we have $\tilde{\beta}_{\ell m} = \tilde{\beta}_{pk}$ only if $\mu_m = \mu_k$. Therefore, if this latter condition is not met and formulation (7.35) is adopted, it is erroneous to include the same coefficient $\tilde{\beta}_{\ell m}$ in the utility of alternatives belonging to different nests. Actually, the constraint $\mu_m = \mu_k$ is sometimes imposed in practice in order to use formulation (7.35) of the nested logit model. It is important to note that it does not correspond to the most general specification.

In order to avoid these issues, we recommend the normalization $\mu = 1$, as it is done for the logit model. Each μ_m is then estimated, imposing the condition that $\mu_m \geq 1$. This normalization is less convenient for a sequential estimator, where utility functions of alternatives in different nests share some parameters. In this case, the normalization $\mu_1 = \mu_2 = \ldots = \mu_M$ should be preferred.

7.3 Estimation

The maximum likelihood estimation of the nested logit model is similar to the estimation of the logit model. It consists in solving the optimization problem

$$\max_{\beta_1,\ldots,\beta_K,\mu_1,\ldots,\mu_M} \mathcal{L}(\beta_1,\ldots,\beta_K,\mu_1,\ldots,\mu_M), \tag{7.36}$$

where

$$\mathcal{L}(\beta_1, \ldots, \beta_K, \mu_1, \ldots, \mu_M) = \sum_n \sum_{i \in \mathcal{C}_n} y_{in} \ln P_n(i|\mathcal{C}_n), \tag{7.37}$$

assuming that the model has been normalized from the top.

It is slightly more difficult than for the logit model, as the probability model $P_n(i|\mathcal{C}_n)$ defined by (7.24) is a lot more involved than the expression in (5.14). A first consequence is that the formula of the derivatives (provided in Section 5.6.1 for the logit model), although obtained through straightforward calculus, become more tedious to handle and to implement (see appendix 7.A). A second consequence is that the log likelihood function is no longer concave with respect to the parameters even when the utility functions are linear-in-parameters. Therefore, numerical optimization methods such as those described in Section B.7 may get trapped in a local maximum. The loss of concavity is attributable to the presence of the scale parameters. Daganzo and Kusnic (1993) have shown that if the scale parameters are fixed (not estimated), the log likelihood function is concave if the V's are linear-in-parameters.

It may thus be tempting to exploit the special structure of the model to estimate the within-nest model and the across-nest model separately. This technique, called *sequential estimation* and illustrated in Section 7.1, consists in three steps:

1. Estimate the within-nest logit model (7.13) for each nest using maximum likelihood in order to obtain the estimates $\widehat{\beta}$. If some parameters are common to several models, a joint estimation is necessary.

2. Compute the logsum's (7.17) for each nest using the values of $\widehat{\beta}$ computed at the previous step, and call them $\widehat{V}_m$.

3. Estimate the across-nest model (7.23).

In this case, each estimation is based on a logit specification, involving a concave log likelihood function (if the V's are linear in parameters.) Except in the presence of a large sample, simultaneous estimation produces more efficient estimators of the parameters. In addition to various technical issues associated with keeping the consistency of parameters' scale across various models, Brownstone and Small (1989) have shown that the sequential estimator can be much less efficient (that is, with a much larger variance) than the simultaneous estimator, and its uncorrected second-stage standard-error estimates may be strongly downward biased. They have experienced cases in which the sequential estimator does not exist, but simultaneous estimation still performs well.

7.4 Airline itinerary choice

We consider again the airline itinerary choice example, and estimate a nested logit specification, with the exact same specification of the utility functions as the logit model presented in Table 6.18, where the nest parameters have been normalized from the top, that is μ has been normalized to 1, and the nest parameter μ_m has been estimated. Estimation results are presented in Table 7.5. Note that the estimated value of the μ_m parameter is larger than 1. It is interesting to compare the estimation results in Tables 7.4 and 7.5. Indeed, the two models are equivalent, up to the normalization of their parameters.

The parameter estimates presented in Table 7.5 can be obtained from the estimates in Table 7.4 using the following processing:

1. rescale all parameters by multiplying them by $\mu = 0.546$,
2. shift the three alternative specific constants by $-\mu\beta_0 = -0.949$,
3. shift the three alternative specific coefficients of the variable "More than two air trips per year" by $-\mu\beta_{12'} = -0.109$.

In order to test if the nested logit is better than the logit model, we have used a likelihood ratio test in section 7.1. Here, we test the hypothesis that the nest parameter μ_m is equal to one using a t-test. Indeed, if it was the case, the nested logit model would collapse to a logit. The value of the test is (see Eq. (6.5))

$$\frac{1.83 - 1}{0.199} = 4.17 \tag{7.38}$$

and is reported in the estimation results. The hypothesis can be rejected at the 95% level. Finally, the correlation between the error terms of the two alternatives in the nest is equal to

$$1 - \frac{1}{1.83^2} = 0.701,$$

so that the correlation matrix associated with the nested logit model is

$$\Sigma_{NL} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.701 \\ 0 & 0.701 & 1 \end{pmatrix}.$$

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.710	0.169	-4.20	0.00
2	One stop–multiple airlines	-0.949	0.173	-5.47	0.00
3	Round trip fare ($100)	-1.54	0.149	-10.29	0.00
4	Elapsed time (0–2 hours)	-0.815	0.215	-3.80	0.00
5	Elapsed time (2–8 hours)	-0.190	0.0610	-3.12	0.00
6	Elapsed time (> 8 hours)	-0.887	0.267	-3.32	0.00
7	Leg room (inches), if male (non stop)	0.0919	0.0310	2.96	0.00
8	Leg room (inches), if female (non stop)	0.180	0.0296	6.08	0.00
9	Leg room (inches), if male (one stop)	0.0954	0.0219	4.35	0.00
10	Leg room (inches), if female (one stop)	0.0610	0.0193	3.16	0.00
11	Being early (hours)	-0.128	0.0160	-7.97	0.00
12	Being late (hours)	-0.0739	0.0141	-5.23	0.00
13	More than two air trips per year (one stop–same airline)	-0.239	0.124	-1.93	0.05
14	More than two air trips per year (one stop–multiple airlines)	-0.109	0.132	-0.82	0.41
15	Round trip fare / income ($100/$1000)	-19.9	7.47	-2.66	0.01
16	μ_m	1.83	0.199	4.17[1]	0.00

Summary statistics

Number of observations = 2544

$$\mathcal{L}(0) = -2794.870$$
$$\mathcal{L}(c) = -2203.160$$
$$\mathcal{L}(\hat{\beta}) = -1613.858$$
$$-2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] = 2362.022$$
$$\rho^2 = 0.423$$
$$\bar{\rho}^2 = 0.417$$

Table 7.5: Nested logit model: simultaneous estimation normalized from the top

[1] t-test against 1

7.5 Multiple levels

The nest structure can be extended to multiple levels, using the same arguments as above. In this case, each nest can be partitioned into sub-nests, which in turn can be partitioned as well.

As for the nested logit model, we partition the choice set into M mutually exclusive subsets, called *nests*, such that each alternative belongs to one and only one nest. That is

$$\mathcal{C} = \bigcup_{m=1}^{M} \mathcal{C}_m.$$

Then, each nest m is partitioned into M_m sub-nests, that is

$$\mathcal{C}_m = \bigcup_{p=1}^{M_m} \mathcal{C}_{pm},$$

and the process can be repeated recursively. We derive here the 3-level model. Generalization to more levels is straightforward. It is usually convenient to represent the model using a tree structure. Each leaf-node corresponds to an alternative, and is associated with its utility. Each node in the tree which is not a leaf corresponds to a logit model among aggregate alternatives, with a specific scale parameter, denoted μ_m at level 1, and μ_{pm} at level 2 (see Figure 7.2). These nodes are associated with an aggregate utility, which is the expected maximum utility of the associated logit model.

The utility of alternative $i \in \mathcal{C}_{pm}$ is decomposed as

$$U_i = V_i + \varepsilon_m + \varepsilon_{pm} + \varepsilon_{ipm}. \tag{7.39}$$

The probability model (7.9) now writes

$$P(i|\mathcal{C}) = \Pr(i|\mathcal{C}_{pm}) \Pr(\mathcal{C}_{pm}|\mathcal{C}_m) \Pr(\mathcal{C}_m|\mathcal{C}), \tag{7.40}$$

where $\mathcal{C}_{pm}$ and $\mathcal{C}_m$ are the only nests, at their respective level, containing alternative i. The first factor is derived as (7.12), leading to

$$\Pr(i|\mathcal{C}_{pm}) = \Pr(V_i + \varepsilon_{ipm} \geq V_j + \varepsilon_{jpm}, j \in \mathcal{C}_{pm}). \tag{7.41}$$

Assuming that the error terms ε_{ipm} are i.i.d. EV distributed, with scale parameter μ_{pm}, we have

$$\Pr(i|\mathcal{C}_{pm}) = \frac{e^{\mu_{pm} V_i}}{\sum_{j \in \mathcal{C}_{pm}} e^{\mu_{pm} V_j}}. \tag{7.42}$$

The second factor is derived similarly to (7.15) to obtain

$$\Pr(\mathcal{C}_{pm}|\mathcal{C}m) = \Pr(\varepsilon_{pm} + \max_{i\in\mathcal{C}_{pm}}(V_i + \varepsilon_{ipm}) \geq \varepsilon_{qm} + \max_{j\in\mathcal{C}_{qm}}(V_j + \varepsilon_{jqm})\ \forall q). \quad (7.43)$$

We define

$$\tilde{V}_{pm} + \varepsilon'_{pm} = \max_{i\in\mathcal{C}_{pm}}(V_i + \varepsilon_{ipm}), \quad (7.44)$$

where

$$\tilde{V}_{pm} = \frac{1}{\mu_{pm}} \ln \sum_{i\in\mathcal{C}_{pm}} e^{\mu_{pm}V_i}, \quad (7.45)$$

as well as

$$\tilde{\varepsilon}_{pm} = \varepsilon_{pm} + \varepsilon'_{pm} \quad (7.46)$$

so that

$$\Pr(\mathcal{C}_{pm}|\mathcal{C}_m) = \Pr(\tilde{V}_{pm} + \tilde{\varepsilon}_{pm} \geq \tilde{V}_{qm} + \tilde{\varepsilon}_{qm}\ \forall q). \quad (7.47)$$

Assuming that the $\tilde{\varepsilon}_{pm}$ are i.i.d. EV distributed with scale parameter μ_m, we obtain

$$\Pr(\mathcal{C}_{pm}|\mathcal{C}_m) = \frac{e^{\mu_m \tilde{V}_m}}{\sum_{k\in\mathcal{C}_m} e^{\mu_k \tilde{V}_k}}. \quad (7.48)$$

Note that the i.i.d. assumption on $\tilde{\varepsilon}_{pm}$ implies i.i.d. for ε_{pm} as well. The last factor in (7.40) is obtained in a similar way: $\Pr(\mathcal{C}_m|\mathcal{C}) =$

$$\Pr\left(\max_{p\in\mathcal{C}_m}\max_{i\in\mathcal{C}_{pm}}(V_i + \varepsilon_m + \varepsilon_{pm} + \varepsilon_{ipm}) \geq \max_{q\in\mathcal{C}_k}\max_{j\in\mathcal{C}_{qk}}(V_j + \varepsilon_k + \varepsilon_{qk} + \varepsilon_{jqk}), \forall k\right), \quad (7.49)$$

that is, using (7.44)–(7.46),

$$\Pr(m|\mathcal{C}) = \Pr(\varepsilon_m + \max_{p\in\mathcal{C}_m}(\tilde{V}_{pm} + \tilde{\varepsilon}_{pm}) \geq \varepsilon_k + \max_{p\in\mathcal{C}_k}(\tilde{V}_{qk} + \tilde{\varepsilon}_{qk}), \forall k). \quad (7.50)$$

We define

$$\widehat{V}_m + \varepsilon''_m = \max_{p\in\mathcal{C}_m}(\tilde{V}_{pm} + \tilde{\varepsilon}_{pm}), \quad (7.51)$$

where

$$\widehat{V}_m = \frac{1}{\mu_m} \ln \sum_{p\in\mathcal{C}_m} e^{\mu_m \tilde{V}_{pm}}, \quad (7.52)$$

and

$$\widehat{\varepsilon}_m = \varepsilon_m + \varepsilon''_m. \quad (7.53)$$

Assuming that the $\widehat{\varepsilon}_m$ are i.i.d. EV distributed with scale parameter μ, we obtain

$$\Pr(\mathcal{C}_m|\mathcal{C}) = \frac{e^{\mu\widehat{\varepsilon}_m}}{\sum_k e^{\mu\widehat{\varepsilon}_k}}. \quad (7.54)$$

Similarly to what has been discussed for the nested-logit model, not all μ parameters are identified. Although any of them may be normalized to one, it is recommended to normalize from the top, that is $\mu = 1$. Also, the condition (7.30) generalizes here to

$$0 \leq \mu \leq \mu_m \leq \mu_{pm}, \text{ for all } m, p, \tag{7.55}$$

which can also be written

$$0 \leq \frac{\mu}{\mu_{pm}} \leq \frac{\mu_m}{\mu_{pm}} \leq 1, \text{ for all } m, p. \tag{7.56}$$

Therefore, the μ parameters must increase as we go down the tree.

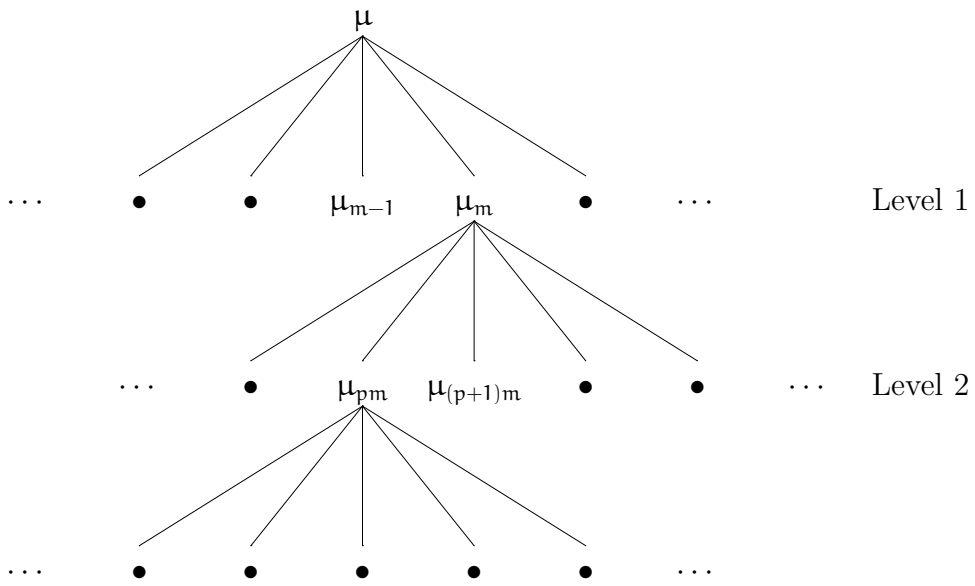


Figure 7.2: Tree representation of a multiple levels nested logit model

The correlation structure of the model is a straightforward generalization of the single level nested logit model. Consider two alternatives i and j within the same lower level nest, that is $i, j \in \mathcal{C}_{pm}$. We have

$$\text{Corr}(U_i, U_j) = 1 - \frac{\mu^2}{\mu_{pm}^2}. \tag{7.57}$$

Consider now alternatives i and j, sharing the same upper level nest m, such that $i \in \mathcal{C}_{pm}$ and $j \in \mathcal{C}_{qm}$, with $p \neq q$. In this case,

$$\text{Corr}(U_i, U_j) = 1 - \frac{\mu^2}{\mu_m^2}. \tag{7.58}$$

Consider an example with 8 alternatives, 4 lower level nests with two alternatives each, and 2 upper level nests containing two lower level nests each, represented by the tree in Figure 7.3.

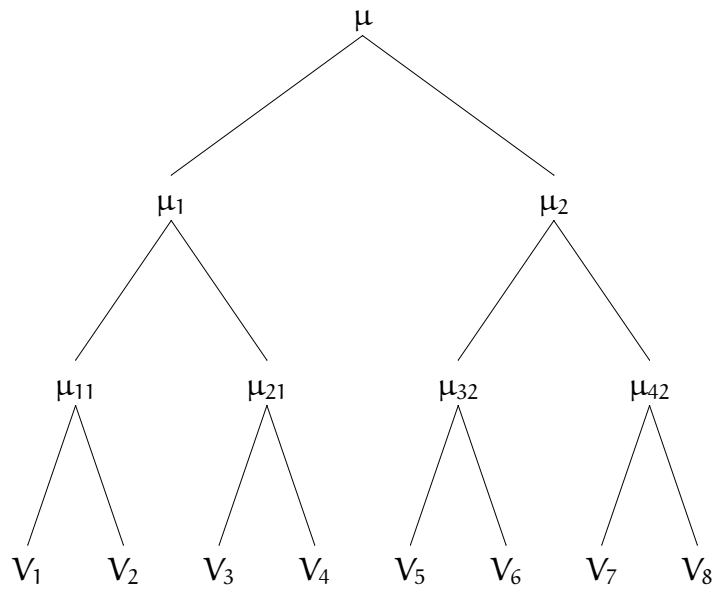


Figure 7.3: Example of multiple level nested logit model

Denoting $\gamma_{pm} = 1 - \mu^2/\mu_{pm}^2$ and $\gamma_m = 1 - \mu^2/\mu_m^2$, we have the following correlation matrix:

$$\begin{pmatrix} 1 & \gamma_{11} & \gamma_1 & \gamma_1 & 0 & 0 & 0 & 0 \\ \gamma_{11} & 1 & \gamma_1 & \gamma_1 & 0 & 0 & 0 & 0 \\ \gamma_1 & \gamma_1 & 1 & \gamma_{21} & 0 & 0 & 0 & 0 \\ \gamma_1 & \gamma_1 & \gamma_{21} & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \gamma_{32} & \gamma_2 & \gamma_2 \\ 0 & 0 & 0 & 0 & \gamma_{32} & 1 & \gamma_2 & \gamma_2 \\ 0 & 0 & 0 & 0 & \gamma_2 & \gamma_2 & 1 & \gamma_{42} \\ 0 & 0 & 0 & 0 & \gamma_2 & \gamma_2 & \gamma_{42} & 1 \end{pmatrix} \tag{7.59}$$

7.6 Summary

The derivation of the logit model is founded on the assumption that the error terms are independent and identically distributed across the alternatives and the individuals. The nested logit model is motivated by the need to relax this assumption in some circumstances for the sake of realism, while trying to maintain the nice properties of logit for the sake of simplicity. Designed to

explicitly capture the presence of shared unobserved attributes, the nested logit model relies on a partitioning of the universal choice set into mutually exclusive nests, such that alternatives sharing unobserved attributes lie in the same nest.

Given

- a partition $\mathcal{C} = \bigcup_{m=1}^{M} \mathcal{C}_m$ of the universal choice set,
- an individual specific choice set $\mathcal{C}_n$,

the nested logit model is given by

$$P_n(i|\mathcal{C}_n) = \frac{e^{\mu_m V_{in}}}{\sum_{j\in\mathcal{C}_{mn}} e^{\mu_m V_{jn}}} \frac{\left(\sum_{\ell\in\mathcal{C}_{mn}} e^{\mu_m V_{\ell n}}\right)^{\frac{\mu}{\mu_m}}}{\sum_{p=1}^{M}\left(\sum_{\ell\in\mathcal{C}_{pn}} e^{\mu_p V_{\ell n}}\right)^{\frac{\mu}{\mu_p}}},$$

where $\mathcal{C}_{mn} = \mathcal{C}_m \cap \mathcal{C}_n$ is the intersection between the nest C_m and the choice set $\mathcal{C}_n$, μ, μ_1, ..., μ_M are unknown parameters to be estimated, verifying $\mu \leq \mu_m$, for all m. For identification purposes, one of these parameters must be normalized to 1. It is convenient to normalize $\mu = 1$, similarly to the normalization of the logit model, so that $\mu_m \geq 1$, for all m. An approximation of the correlation between two alternatives in the same nest is given by

$$\mathrm{Corr}(U_i, U_j) = 1 - \frac{\mu^2}{\mu_m^2}.$$

It is possible to extend the model by partitioning each nest into subnests, in order to obtain a more complex correlation structure. The derivation of the multiple level nested logit model follows the same principles as the single level model.

Chapter Appendix

Like the logit model, the nested logit model is defined by a closed form formula, and its derivatives can be obtained from straightforward calculus. They are useful for the numerical algorithms that optimize the log likelihood function. They are also required to compute elasticities. In Section 7.A, the first derivatives of the log likelihood function with respect to the unknown parameters are computed. The model elasticities are provided in Section 7.B.

7.A Derivatives of the log likelihood function

Consider the log likelihood function (7.37):

$$\mathcal{L}(\beta_1, \dots, \beta_K, \mu_1, \dots, \mu_M) = \sum_n \sum_{i \in \mathcal{C}_n} y_{in} \ln P_n(i|\mathcal{C}_n), \tag{7.60}$$

where $P_n(i|\mathcal{C}_n)$ is the nested logit model (7.24). Optimization algorithms designed to maximize such a function rely on its derivatives with respect to each unknown parameter. We have, for $k = 1, \dots, K$,

$$\frac{\partial \mathcal{L}}{\partial \beta_k}(\beta_1, \dots, \beta_K, \mu_1, \dots, \mu_M) = \sum_{j \in \mathcal{C}_n} \frac{\partial \mathcal{L}}{\partial V_{jn}} \frac{\partial V_{jn}}{\partial \beta_k}, \tag{7.61}$$

where

$$\frac{\partial \mathcal{L}}{\partial V_{jn}} = \sum_n \sum_{i \in \mathcal{C}_n} y_{in} \frac{\partial \ln P_n(i|\mathcal{C}_n)}{\partial V_{jn}} = \sum_n \sum_{i \in \mathcal{C}_n} y_{in} \frac{1}{P_n(i|\mathcal{C}_n)} \frac{\partial P_n(i|\mathcal{C}_n)}{\partial V_{jn}}. \tag{7.62}$$

If V_{jn} is linear-in-parameters, the quantity $\partial V_{jn}/\partial \beta_k$ in (7.61) is the variable x_{jnk} with coefficient β_k in V_{jn}. Similarly, for $m = 1, \dots, M$, we have

$$\frac{\partial \mathcal{L}}{\partial \mu_m}(\beta_1, \dots, \beta_K, \mu_1, \dots, \mu_M) = \sum_n \sum_{i \in \mathcal{C}_n} y_{in} \frac{1}{P_n(i|\mathcal{C}_n)} \frac{\partial P_n(i|\mathcal{C}_n)}{\partial \mu_m}. \tag{7.63}$$

Consider the nested logit model (7.24), where the index n has been removed to simplify the notation:

$$P(i|\mathcal{C}) = \Pr(i|m) \Pr(m|\mathcal{C}) = \frac{e^{\mu_m V_i}}{\sum_{j \in \mathcal{C}_m} e^{\mu_m V_j}} \frac{\left(\sum_{\ell \in \mathcal{C}_m} e^{\mu_m V_\ell}\right)^{\frac{\mu}{\mu_m}}}{\sum_{p=1}^{M} \left(\sum_{\ell \in \mathcal{C}_p} e^{\mu_p V_\ell}\right)^{\frac{\mu}{\mu_p}}}. \tag{7.64}$$

The derivation with respect to any quantity y can be written as follows.

$$\frac{\partial P(i|\mathcal{C})}{\partial y} = \frac{\partial \Pr(i|m)}{\partial y} \Pr(m|\mathcal{C}) + \Pr(i|m) \frac{\partial \Pr(m|\mathcal{C})}{\partial y},$$

where

$$\frac{\partial \Pr(i|m)}{\partial y} = P(i|m)\left(\frac{\partial \mu_m V_i}{\partial y} - \sum_{j\in\mathcal{C}_m} P(j|m)\frac{\partial \mu_m V_j}{\partial y}\right)$$

and

$$\frac{\partial \Pr(m|\mathcal{C})}{\partial y} = \Pr(m|\mathcal{C})\Big(\frac{\partial \mu/\mu_m}{\partial y}\ln\sum_{j\in\mathcal{C}_m} e^{\mu_m V_j} + \frac{\mu}{\mu_m}\sum_{j\in\mathcal{C}_m}\Pr(j|m)\frac{\partial \mu_m V_j}{\partial y}$$

$$-\sum_{p=1}^{M}\frac{\mu}{\mu_p}\Pr(p|\mathcal{C})\sum_{\ell\in\mathcal{C}_p}\Pr(\ell|p)\frac{\partial \mu_p V_\ell}{\partial y}\Big).$$

Now, for (7.62), we need $\partial P(i|\mathcal{C})/\partial V_j$, $j \in \mathcal{C}$. We simplify the above equations when $y = V_j$. We distinguish three cases:

1. $j = i$,
2. $j \neq i$, $j \in \mathcal{C}_m$, that is j is in the same nest as i,
3. $j \in \mathcal{C}_p$, $p \neq m$, that is j is not in the same nest as i.

Case $j = i$ We have

$$\frac{\partial \Pr(i|m)}{\partial V_i} = \mu_m \Pr(i|m)\,(1 - \Pr(i|m)) \tag{7.65}$$

and

$$\frac{\partial \Pr(m|\mathcal{C})}{\partial V_i} = \mu P(i|\mathcal{C})(1 - \Pr(m|\mathcal{C})). \tag{7.66}$$

Therefore,

$$\frac{\partial P(i|\mathcal{C})}{\partial V_i} = \mu_m P(i|\mathcal{C})(1 - \Pr(i|m)) + \mu P(i|\mathcal{C})(\Pr(i|m) - P(i|\mathcal{C})). \tag{7.67}$$

Case $j \neq i$, $j \in \mathcal{C}_m$ We have

$$\frac{\partial \Pr(i|m)}{\partial V_j} = -\mu_m \Pr(i|m)\Pr(j|m) \tag{7.68}$$

and

$$\frac{\partial \Pr(m|\mathcal{C})}{\partial V_j} = \mu P(j|\mathcal{C})(1 - \Pr(m|\mathcal{C})). \tag{7.69}$$

Therefore,

$$\frac{\partial P(i|\mathcal{C})}{\partial V_j} = -\mu_m P(i|\mathcal{C})\Pr(j|m) + \mu P(i|\mathcal{C})(\Pr(i|m) - P(i|\mathcal{C})). \tag{7.70}$$

Case $j \in \mathcal{C}_p$, $p \neq m$ We have

$$\frac{\partial \Pr(i|m)}{\partial V_j} = 0, \tag{7.71}$$

and

$$\frac{\partial \Pr(m|\mathcal{C})}{\partial V_j} = -\mu \Pr(m|\mathcal{C}) P(j|\mathcal{C}). \tag{7.72}$$

Therefore,

$$\frac{\partial P(i|\mathcal{C})}{\partial V_j} = -\mu P(i|\mathcal{C}) P(j|\mathcal{C}). \tag{7.73}$$

Finally, for (7.63) we need $\partial P(i|\mathcal{C})/\partial \mu_m$ and $\partial P(i|\mathcal{C})/\partial \mu_p$, where m is the nest containing alternative i and p is any other nest. We have

$$\frac{\partial \Pr(i|m)}{\partial \mu_m} = \Pr(i|m) \left(V_i - \sum_{j \in \mathcal{C}_m} \Pr(j|m) V_j \right), \tag{7.74}$$

and

$$\frac{\partial \Pr(m|\mathcal{C})}{\partial \mu_m} = \frac{\mu}{\mu_m} \Pr(m|\mathcal{C}) \left((1 - \Pr(m|\mathcal{C})) \sum_{j \in \mathcal{C}_m} \Pr(j|m) V_j - \frac{1}{\mu_m} \ln \sum_{j \in \mathcal{C}_m} e^{\mu_m V_j} \right). \tag{7.75}$$

Therefore,

$$\begin{aligned} \frac{\partial \Pr(i|\mathcal{C})}{\partial \mu_m} &= P(i|\mathcal{C}) \left(V_i - \sum_{j \in \mathcal{C}_m} \Pr(j|m) V_j \right) \\ &+ \frac{\mu}{\mu_m} P(i|\mathcal{C}) \left((1 - \Pr(m|\mathcal{C})) \sum_{j \in \mathcal{C}_m} \Pr(j|m) V_j - \frac{1}{\mu_m} \ln \sum_{j \in \mathcal{C}_m} e^{\mu_m V_j} \right). \end{aligned} \tag{7.76}$$

If $p \neq m$, we have

$$\frac{\partial \Pr(i|m)}{\partial \mu_p} = 0, \tag{7.77}$$

and

$$\frac{\partial \Pr(m|\mathcal{C})}{\partial \mu_p} = -\frac{\mu}{\mu_p} \Pr(m|\mathcal{C}) \Pr(p|\mathcal{C}) \sum_{\ell \in \mathcal{C}_p} \Pr(\ell|p) V_\ell. \tag{7.78}$$

Therefore,

$$\frac{\partial P(i|\mathcal{C})}{\partial \mu_p} = -\frac{\mu}{\mu_p} P(i|\mathcal{C}) \Pr(p|\mathcal{C}) \sum_{\ell \in \mathcal{C}_p} \Pr(\ell|p) V_\ell. \tag{7.79}$$

7.B Elasticities of the nested logit model

As described in Section 10.5.1, the disaggregate *direct elasticity* with respect to one of the continuous variables x_{ink}'s is given by

$$E_{x_{ink}}^{P_n(i)} = \frac{\partial P_n(i)}{\partial x_{ink}} \frac{x_{ink}}{P_n(i)} = \frac{\partial P_n(i)}{\partial V_{in}} \frac{\partial V_{in}}{\partial x_{ink}} \frac{x_{ink}}{P_n(i)}, \tag{7.80}$$

where $\partial P_n(i)/\partial V_{in}$ is given by (7.67). In the case of a linear-in-parameter utility function, $\partial V_{in}/\partial x_{ink}$ is the coefficient β_{ik} of variable x_{ink}.

Similarly the *disaggregate cross elasticity* of the probability alternative i that is selected with respect to an attribute of alternative j is

$$E_{x_{jnk}}^{P_n(i)} = \frac{\partial P_n(i)}{\partial x_{jnk}} \frac{x_{jnk}}{P_n(i)} = \frac{\partial P_n(i)}{\partial V_{jn}} \frac{\partial V_{jn}}{\partial x_{jnk}} \frac{x_{jnk}}{P_n(i)}, \tag{7.81}$$

where $\partial P_n(i)/\partial V_{jn}$ is given by (7.70) if j is in the same nest as i, and by (7.73) if j is not in the same nest as i.

Chapter 8

Multivariate Extreme Value models

Contents

The nested logit model presented in Chapter 7 is designed to relax the assumption of independence of the error terms in a logit model. The model is simple to use but the type of correlation which can be captured is rather limited. Namely, the assumption that each alternative must belong to exactly one nest, and the resulting block diagonal structure of the correlation matrix, are often restrictive in real applications.

In the particular case of multi-dimensional choice sets, the independence assumption of the logit does not hold, and the partition of the choice sets into nests is not natural. More formally, a two-dimensional choice set $\mathcal{C}_n$ for

individual n is the cartesian product of two choice sets $\mathcal{D}_1$ and $\mathcal{D}_2$, where the set of combinations infeasible for n, denoted by $\mathcal{C}_n^*$, has been removed, that is

$$\mathcal{C}_n = (\mathcal{D}_1 \times \mathcal{D}_2) \setminus \mathcal{C}_n^*.$$

To illustrate this concept, consider the analysis of travel for shopping purposes, where both the destination and the transportation mode are relevant. In this case, $\mathcal{D}_1 = \{m_1, \ldots, m_{J_M}\}$ can represent the set of all possible transportation modes for shopping, and $\mathcal{D}_2 = \{d_1, \ldots, d_{J_D}\}$ the set of all possible destinations for shopping. In this case

$$\begin{aligned}\mathcal{D}_1 \times \mathcal{D}_2 =& \{(m_1, d_1), (m_1, d_2), \ldots, (m_1, d_{J_d}), (m_2, d_1), \ldots, (m_2, d_{J_d}), \ldots, \\ & (m_{J_M}, d_1), \ldots, (m_{J_M}, d_{J_d})\}\end{aligned}$$

contains all possible combinations of transportation modes and destinations. The set $\mathcal{C}_n^*$ contains all such combinations that are not in the choice set of individual n. For instance, a shopping mall may not be accessible by train, or traveler n may not be aware of that. A *multidimensional* choice set is similarly defined, when more than two choice sets may be involved in the cartesian product.

Alternatives along each dimension within a multidimensional choice set are sharing attributes, several of which are likely to be unobserved. If a nested logit model is defined using a partition of the choice set along one dimension, the alternatives in each nest are still sharing the attributes of the other dimensions. With the example above, the choice set could be partitioned across transportation modes, that is each mode is associated with a nest. Such a model would capture the correlation among the utility of all alternatives sharing the same mode. But the utility of the alternatives sharing the same destination would be independent random variables, despite the fact that they are likely to share unobserved attributes. A nested logit model with a partition based on destination would have the same limitations, as the correlation among alternatives sharing the same mode cannot be captured.

8.1 Illustration

We illustrated the nested logit for the airline itinerary choice example in Section 7.4. We postulated the nesting structure described in Figure 7.1. This structure reflects the assumption that the alternative "One stop, same airline" and "One stop, multiple airlines" share common unobserved variables, and that their error terms may therefore be correlated. It also means that the error term of the "Non stop" alternative is independent from the two others.

Clearly, this assumption is arbitrary, and other assumptions could be considered. For instance, we may want to assume that the "Non stop" alternative shares with the "One stop, same airline" alternative unobserved attributes associated with the fact that a unique airline is used. This assumption would correspond to the nesting structure presented in Figure 8.1. The estimation of this model provides the exact same results as the logit model reported in Table 6.18. Therefore, this nesting structure is rejected.

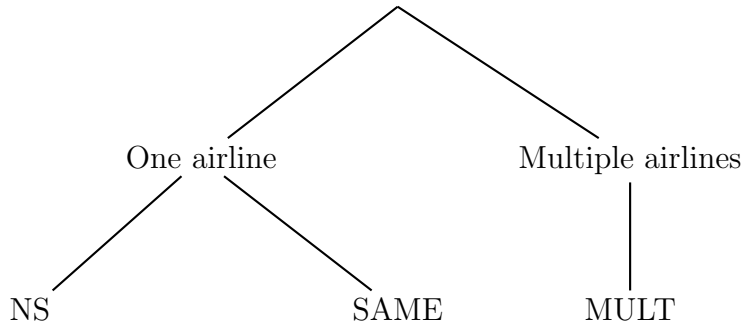


Figure 8.1: Another tree structure for the airline itinerary example

But none of these models allows us to test the assumption that the alternative "One stop, one airline" shares unobserved attributes with the "Non stop" alternative (as they both involve only one airline) **and** with the "One stop, multiple airlines" alternative (as they both involve one stop). This assumption is represented in Figure 8.2. The nested logit model requires that each alternative unambiguously belongs to one and only nest, excluding the possibility to capture such a structure. In this case, a more flexible model, called the *cross nested logit* model, is needed.

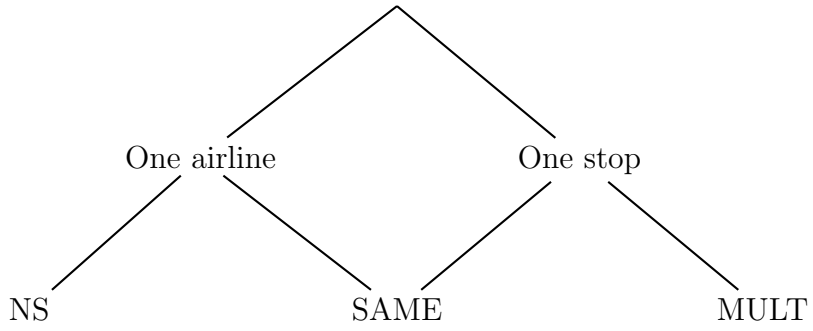


Figure 8.2: Cross nested structure for the airline itinerary example

The cross nested logit model is formally introduced later in the chapter. We illustrate it here to introduce the concept. The specification of the utility function is the same as the nested logit model estimated in Section 7.4. The

cross nested structure, depicted in Figure 8.2 consists of two nests labeled "One airline" and "One stop". The alternative "Non stop" belongs only to the "One airline" nest, the alternative "One stop, multiple airlines" belongs only to the "One stop" nest, and the alternative "One stop, same airline" belongs to both nests, as discussed above. The cross nested logit associates a parameter with each combination of alternative and nest. These parameters have a value between 0 and 1 that can be somehow interpreted as the "degree of membership" of the alternative to the nest. For a given alternative, these parameters should sum up to one. If an alternative belongs only to one nest, its degree of membership is 1 for this nest, and 0 for any other nest. In the example above, this is the case for the "Non stop" alternative in the "One airline" nest, and the "One stop, multiple airlines" in the "One stop" nest. Consequently, a nested logit model is a cross nested logit model where each membership parameter is either 0 or 1.

In the structure represented in Figure 8.2, alternative SAME belongs to the two nests. For the sake of illustration, we assume that it belongs equally to each nest, and sets its degree of membership to 0.5 for each of the two nests. This assumption will be relaxed in Section 8.9. The estimation results of this model are presented in Table 8.1.

The above example illustrates the need to go beyond the nested logit model, and to investigate models with a more flexible error structure. The cross nested model informally introduced above is one such model. Together with the logit and the nested logit models, it belongs to a family of models called *multivariate extreme value* (MEV) models. Not only they allow for a more complex structure of the correlation among the alternatives, but they also share with the logit and nested logit models the interesting property of having a closed form formula for the choice probability.

The general formulation of the random utility model in the multivariate context is introduced in Section 8.2. The multivariate extreme value model is derived in Section 8.3, and several of its properties are discussed in Section 8.4. Specific instances are then discussed: the logit model (Section 8.5), the nested logit model (Section 8.6), the cross-nested logit model (Section 8.7) and the network MEV model (Section 8.8). The airline itinerary example is revisited in Section 8.9. Section 8.10 contains a summary of the chapter.

Parameter number	Description	Coeff. estimate	Asympt. std. error	t-stat	p-value
1	One stop–same airline dummy	-0.674	0.185	-3.64	0.00
2	One stop–multiple airlines	-1.10	0.175	-6.29	0.00
3	Round trip fare ($100)	-1.55	0.170	-9.10	0.00
4	Elapsed time (0–2 hours)	-0.783	0.210	-3.72	0.00
5	Elapsed time (2–8 hours)	-0.177	0.0627	-2.82	0.00
6	Elapsed time (> 8 hours)	-0.832	0.274	-3.03	0.00
7	Leg room (inches), if male (non stop)	0.0904	0.0305	2.97	0.00
8	Leg room (inches), if female (non stop)	0.174	0.0302	5.77	0.00
9	Leg room (inches), if male (one stop)	0.0998	0.0227	4.40	0.00
10	Leg room (inches), if female (one stop)	0.0640	0.0200	3.20	0.00
11	Being early (hours)	-0.128	0.0175	-7.30	0.00
12	Being late (hours)	-0.0747	0.0154	-4.86	0.00
13	More than two air trips per year (one stop–same airline)	-0.241	0.120	-2.01	0.04
14	More than two air trips per year (one stop–multiple airlines)	-0.0964	0.132	-0.73	0.47
15	Round trip fare / income ($100/$1000)	-17.9	7.68	-2.34	0.02
16	$\mu_{\text{One airline}}$	1.11	0.122	0.86[1]	0.39
17	$\mu_{\text{One stop}}$	2.38	0.392	3.51[1]	0.00

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\hat{\beta}) &= -1615.470 \\ -2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 2358.799 \\ \rho^2 &= 0.422 \\ \bar{\rho}^2 &= 0.416 \end{aligned}$$

[1] t-test against 1

Table 8.1: Airline itinerary: cross nested logit

8.2 Multidimensional random utility

We are now considering the utility functions of the alternatives together, in a multidimensional way, using a vector notation:

$$\begin{pmatrix} U_{1n} \\ \vdots \\ U_{Jn} \end{pmatrix} = \begin{pmatrix} V_{1n} \\ \vdots \\ V_{Jn} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1n} \\ \vdots \\ \varepsilon_{Jn} \end{pmatrix} \tag{8.1}$$

or, in compact form,

$$U_n = V_n + \varepsilon_n \tag{8.2}$$

where $V_n \in \mathbb{R}^{J_n}$ is a vector of real numbers of dimension J_n and ε_n is a vector of random variables of the same dimension. The derivation of a choice model is based on assumptions about the distribution of each error term ε_n. The *mean* of the random vector is a vector with J_n entries, each of them containing the mean of the corresponding random variable. The *variance-covariance* matrix Σ_n of the random vector ε_n is a $J_n \times J_n$ matrix such that

- each diagonal entry $(\Sigma_n)_{ii}$ is the variance of the random variable $(\varepsilon_n)_i$, and
- each off-diagonal entry $(\Sigma_n)_{ij}$, $i \neq j$ is the covariance between the random variables $(\varepsilon_n)_i$ and $(\varepsilon_n)_j$.

Note that a variance-covariance matrix is symmetric (as the covariance between $(\varepsilon_n)_i$ and $(\varepsilon_n)_j$ is the same as the covariance between $(\varepsilon_n)_j$ and $(\varepsilon_n)_i$) and positive semidefinite[1].

For example, the logit model in this context is derived from the assumption that each entry of the ε_n vector is a random variable with an (identical) extreme value distribution, that is

$$(\varepsilon_n)_j \sim EV(0, 1),\ j = 1, \ldots, J_n,$$

and that these random variables are independent. The mean of ε_n is a vector of length J_n such that each entry is 0. The independence assumption implies that the off diagonal entries of the variance-covariance matrix are 0. The distributional assumption implies that the diagonal entries (i.e. the variance of each $(\varepsilon_n)_i$, are $\pi^2/6\mu^2$ (see Section B.2.6 about the Extreme Value

[1]The concept of *positive semidefinite* is the extension to matrices of the concept of being non negative for scalars

distribution). Therefore, the variance-covariance matrix of the error term ε_n of the logit model is

$$\Sigma_n^{\text{logit}} = \begin{pmatrix} \frac{\pi^2}{6\mu^2} & 0 & \dots & 0 \\ 0 & \frac{\pi^2}{6\mu^2} & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & \frac{\pi^2}{6\mu^2} \end{pmatrix},$$

and is the same for each individual n. Because of the independence assumption, the CDF and the pdf of the joint distribution of ε_n are simply the product of the CDF and pdf (respectively) of each entry, that is

$$F_{\varepsilon_n}(\xi_1, \dots, \xi_{J_n}) = \prod_{j=1}^{J_n} F_{(\varepsilon_n)_j}(\xi_j) = \prod_{j=1}^{J_n} \exp\left(-e^{-\xi_j}\right), \tag{8.3}$$

and

$$f_{\varepsilon_n}(\xi_1, \dots, \xi_{J_n}) = \prod_{j=1}^{J_n} f_{(\varepsilon_n)_j}(\xi_j) = \prod_{j=1}^{J_n} e^{-\xi_j} \exp\left(-e^{-\xi_j}\right). \tag{8.4}$$

Clearly, a multidimensional approach is not particularly of interest in the independent case. The logit example above illustrates the concept of a multidimensional distribution, but does not involve any correlation.

In the previous chapter, we have introduced the nested logit model, based on the partition of the choice set into M nests, where each nest m contains J_m alternatives. The error terms of the alternatives within a nest are assumed to be correlated, while the error terms of alternatives belonging to different nests are assumed to be independent. Consequently, the variance-covariance matrix of the error term is block diagonal, where each block corresponds to a different nest.

$$\Sigma_n^{\text{nested}} = \begin{pmatrix} D_1 & 0 & \dots & 0 \\ 0 & D_2 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & D_M \end{pmatrix},$$

where D_m, $m = 1, \ldots, M$ is a submatrix of size $J_m \times J_m$, defined as

$$D_m = \begin{pmatrix} \frac{\pi^2}{6\mu^2} & \frac{\pi^2}{6\mu^2} - \frac{\pi^2}{6\mu_m^2} & \cdots & \frac{\pi^2}{6\mu^2} - \frac{\pi^2}{6\mu_m^2} \\ \frac{\pi^2}{6\mu^2} - \frac{\pi^2}{6\mu_m^2} & \frac{\pi^2}{6\mu^2} & \cdots & \frac{\pi^2}{6\mu^2} - \frac{\pi^2}{6\mu_m^2} \\ & & \ddots & \\ \frac{\pi^2}{6\mu^2} - \frac{\pi^2}{6\mu_m^2} & \frac{\pi^2}{6\mu^2} - \frac{\pi^2}{6\mu_m^2} & \cdots & \frac{\pi^2}{6\mu^2} \end{pmatrix},$$

and the zeros represent submatrices of appropriate size containing only zeros. Although the nested logit model indeed relaxes the independence assumption associated with the logit model, the structure of the variance-covariance matrix is not flexible. It has to be block diagonal, and all off-diagonal entries in a block have to be the same.

The probit model offers more flexibility. It is derived from the assumption that the random vector ε_n follows a multivariate normal distribution, with mean 0 and a variance-covariance matrix Σ_n^{probit}, that is

$$\varepsilon_n \sim N(0, \Sigma_n^{\text{probit}}), \tag{8.5}$$

where 0 represents here a vector with J_n entries, all zeros, and Σ_n^{probit} is a $J_n \times J_n$ variance-covariance matrix. As the matrix Σ_n^{probit} is positive semidefinite, it is possible to define it in terms of its Cholesky factors[2], that is

$$\Sigma_n^{\text{probit}} = L_n L_n' \tag{8.6}$$

where L_n is a $J_n \times J_n$ lower triangular matrix with non negative elements on the diagonal.

For example, a probit model for airline choice example described above can be derived based on the assumption that the vector ε_n follows a multivariate normal distribution $N(0, \Sigma_n^{\text{probit}})$, where

$$\Sigma_n^{\text{probit}} = \begin{pmatrix} \sigma_{\text{NS}}^2 & \sigma_{\text{NS,SAME}} & 0 \\ \sigma_{\text{NS,SAME}} & \sigma_{\text{SAME}}^2 & \sigma_{\text{SAME,MULT}} \\ 0 & \sigma_{\text{SAME,MULT}} & \sigma_{\text{MULT}}^2 \end{pmatrix}. \tag{8.7}$$

[2]Intuitively, the Cholesky factors can be considered as the squared root of the matrix.

The Cholesky factor of Σ_n^{probit} has the form:

$$L_n = \begin{pmatrix} \ell_1 & 0 & 0 \\ \ell_{21} & \ell_2 & 0 \\ 0 & \ell_{32} & \ell_3 \end{pmatrix}, \tag{8.8}$$

where

$$\begin{aligned} \sigma^2_{\text{NS}} &= \ell_1^2, \\ \sigma^2_{\text{SAME}} &= \ell_{21}^2 + \ell_2^2, \\ \sigma^2_{\text{MULT}} &= \ell_{32}^2 + \ell_3^2, \\ \sigma_{\text{NS,SAME}} &= \ell_1 \ell_{21}, \\ \sigma_{\text{SAME,MULT}} &= \ell_2 \ell_{32}. \end{aligned}$$

The ℓ parameters are estimated from data (subject to proper normalization).

The derivation of the random utility model from a given multivariate distribution is detailed in Appendix 3.A. If $f_{\varepsilon_n}(\varepsilon)$ is the probability density function of the multivariate random variable ε_n and $F_{\varepsilon_n}(\varepsilon)$ its cumulative distribution function, the choice probability writes

$$P_n(i|\mathcal{C}_n) = \int_{\varepsilon_i=-\infty}^{+\infty} \int_{\varepsilon_1=-\infty}^{V_{in}-V_{1n}+\varepsilon_i} \cdots \int_{\varepsilon_{J_n}=-\infty}^{V_{in}-V_{J_n n}+\varepsilon_i} f_{\varepsilon_{1n},\varepsilon_{2n},\ldots,\varepsilon_{J_n}}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{J_n}) d\varepsilon, \tag{8.9}$$

or, if the CDF is available, the model is given by (3.62), that is

$$P_n(i|\mathcal{C}_n) = \int_{\varepsilon=-\infty}^{+\infty} \frac{\partial F_{\varepsilon_{1n},\varepsilon_{2n},\ldots,\varepsilon_{J_n}}}{\partial \varepsilon_i}(\ldots, V_{in}-V_{(i-1)n}+\varepsilon, \varepsilon, V_{in}-V_{(i+1)n}+\varepsilon, \ldots) d\varepsilon, \tag{8.10}$$

Although (8.10) involves only a unidimensional integral, it may be quite cumbersome to compute. In the probit case, with a multivariate normal distribution, the CDF is not available in a closed form, and (8.9) must be used. This approach is in general not operational due to the multifold integral. Although various estimation techniques have been proposed in the literature (Bolduc, 1999, Natarajan et al., 2000), they are limited to choice models with a few alternatives (typically less than 10).

We now introduce the multivariate extreme value model, a family of models derived from an assumption about the CDF, so that (3.62) can be used to obtain a closed form choice model.

8.3 The Multivariate Extreme Value model

As the logit model is derived from the assumption that the error terms follow an extreme value distribution, we now consider the multivariate version of this distribution.

A vector $\varepsilon_n = (\varepsilon_{1n}, \ldots, \varepsilon_{Jn})$ follows a *multivariate extreme value*[3] distribution if it is characterized by the following cumulative distribution function:

$$F_{\varepsilon_n}(\xi_1, \ldots, \xi_J) = e^{-G(e^{-\xi_1}, \ldots, e^{-\xi_J})}, \tag{8.11}$$

where $G : \mathbb{R}_+^{J_n} \to \mathbb{R}_+$ is a positive function accepting positive arguments, denoted by $y_1, \ldots, y_J$ in the following. Note from (8.11) that, in our context, $y_i = \exp(-\xi_i)$, guaranteeing the positivity of the arguments. To be a valid CDF, the function F_{ε_n} must verify some properties, and so does G:

1. F_{ε_n} goes to zero when any argument goes to $-\infty$, that is

   $$F_{\varepsilon_n}(\xi_1, \ldots, -\infty, \ldots, \xi_J) = 0.$$

   When ξ_i goes to $-\infty$, then $y_i = \exp(-\xi_i)$ goes to $+\infty$. Consequently, the corresponding condition on G is

   $$G(y_1, \ldots, +\infty, \ldots, y_J) = +\infty, \tag{8.12}$$

   that is the function must go to infinity whenever one of its arguments does. It is called the *limit* property.

2. F_{ε_n} goes to one when all of its arguments go to $+\infty$, that is

   $$F_{\varepsilon_n}(+\infty, \ldots, +\infty) = 1.$$

   When ξ_i goes to $+\infty$, then $y_i = \exp(-\xi_i)$ goes to 0. Consequently, the corresponding condition on G is

   $$G(0, \ldots, 0) = 0. \tag{8.13}$$

3. Any partial derivative of F_{ε_n} defines a density function of a marginal distribution. To be a valid density function, it has to be non negative. More precisely, for any set of $\widehat{J}_n \leq J_n$ distinct indices $i_1, \ldots, i_{\widehat{J}_n}$,

   $$\frac{\partial^{\widehat{J}_n} F_{\varepsilon_n}}{\partial \varepsilon_{i_1 n} \cdots \partial \varepsilon_{i_{\widehat{J}_n} n}}(\varepsilon_{1n}, \ldots, \varepsilon_{J_n n}) \geq 0.$$

   In particular, if $\widehat{J}_n = J_n$, we obtain the density function of the entire distribution of ε_n, that is

   $$f_{\varepsilon_n}(\varepsilon_{1n}, \ldots, \varepsilon_{J_n n}) = \frac{\partial^{J_n} F_{\varepsilon_n}}{\partial \varepsilon_{1n} \cdots \partial \varepsilon_{J_n n}}(\varepsilon_{1n}, \ldots, \varepsilon_{J_n n}) \geq 0.$$

[3]There are several families of multivariate extreme value distributions. We refer the interested reader to Pickands (1981), Joe (1997) or Kotz and Nadarajah (2001), among others.

Considering (8.11), the above condition says that any level of differentiation must correspond to a non-negative result. It appears that the right-hand side of (8.11) changes sign each time it is differentiated, except the first time. Indeed, $\partial y_{in}/\partial \varepsilon_{in} = -\exp(-\varepsilon_{in}) = -y_{in}$. To compensate that and always obtain a non negative sign, the function G must also change sign each time it is differentiated. This condition is called the *strong alternating sign property*, and states that the cross partial derivatives of G have alternative signs. That is, at the first degree,

$$G_i = \partial G/\partial y_i \geq 0,$$

for $i = 1, \ldots, J_n$. At the second degree,

$$G_{ij} = \partial G_i/\partial y_j = \partial^2 G/\partial y_i \partial y_j \leq 0,$$

for $i \neq j$. For higher degrees, and for any set of $\widehat{J}_n$ distinct indices $i_1, \ldots, i_{\widehat{J}_n}$,

$$(-1)^{\widehat{J}_n - 1} G_{i_1,\ldots,i_{\widehat{J}_n}} \geq 0. \tag{8.14}$$

If these properties are verified, (8.11) is a valid CDF. We also need an additional condition on G: *homogeneity*. A function G is homogeneous of degree μ, or μ-homogeneous, if

$$G(\alpha y) = \alpha^\mu G(y), \ \forall \alpha > 0 \text{ and } y \in \mathbb{R}^J_+. \tag{8.15}$$

The homogeneity condition implies two important properties of the model. We show below that, if G is homogeneous,

- the marginals of (8.11) are univariate extreme value distributions, so that the valid CDF indeed corresponds to a multivariate extreme value distribution, and,
- the corresponding choice model has a closed form.

The ith marginal distribution of (8.11) is given by

$$F_{\varepsilon_n}(+\infty, \ldots, +\infty, \varepsilon_{in}, +\infty, \ldots, +\infty) = e^{-G(0,\ldots,0,e^{-\varepsilon_{in}},0,\ldots,0)}. \tag{8.16}$$

If G is μ-homogeneous, we have

$$G(0, \ldots, 0, e^{-\varepsilon_{in}}, 0, \ldots, 0) = e^{-\mu\varepsilon_{in}} G(0, \ldots, 0, 1, 0, \ldots, 0),$$

or equivalently,

$$G(0, \ldots, 0, e^{-\varepsilon_{in}}, 0, \ldots, 0) = e^{-\mu\varepsilon_{in} + \log G(0,\ldots,0,1,0,\ldots,0)},$$

The quantity $\log G(0, \ldots, 0, 1, 0, \ldots, 0)$ is a constant. Call it $\mu\eta$, so that the CDF of the i^{th}marginal distribution of ε_n is

$$F_{\varepsilon_n}(+\infty, \ldots, +\infty, \varepsilon_{in}, +\infty, \ldots, +\infty) = \exp\left(-e^{-\mu(\varepsilon_{in}-\eta)}\right), \tag{8.17}$$

which is the CDF of a univariate extreme value distribution with location parameter η and scale parameter μ.

We have now established that F is the CDF of a multivariate extreme value distribution if G verifies a handful of properties:

M1: the strong alternating sign property (8.14),

M2: the μ-homogeneity property (8.15), and,

M3: the limit property (8.12).

Note that the condition (8.13) is not included in the above list, as it is a direct consequence of the homogeneity property.

The corresponding choice model is obtained by incorporating (8.11) into (3.62) (see the derivation in Appendix 8.A).

$$P_n(i) = \frac{e^{V_{in}+\log G_i\left(e^{V_{1n}}, \ldots, e^{V_{Jn}n}\right)}}{\sum_j e^{V_{jn}+\log G_j\left(e^{V_{1n}}, \ldots, e^{V_{Jn}n}\right)}}. \tag{8.18}$$

This is the *multivariate extreme value* (MEV) model. The function G is called a *choice probability generating function* (CPGF).

Formulation (8.18) is interesting because it has a similar structure as the logit model. Indeed, it can be interpreted as a logit model, where each systematic utility V_i is shifted by $\log G_i(\cdot)$. As a consequence, relaxing the independence assumption associated with the logit model can be accommodated by an appropriate correction of the utility functions, while keeping the functional form of the logit. However, it has to be remembered that the utility of an alternative i depends on the variables of all alternatives in the MEV context. Moreover, if the G function has a closed form, so has the choice model.

Before we analyze concrete instances of this family in more details, we comment on some properties.

8.4 Properties of the MEV model

We present here various comments, properties and features of the MEV model. This section can be skipped without loss of continuity.

Daniel McFadden was born in 1937 in North-Carolina. He joined UC Berkeley in 1963, moved to the Massachusetts Institute of Technology in 1977 and returned to the University of California in 1991, where he is E. Morris Cox Professor of Economics, and director of the econometrics laboratory. The pioneering work of Daniel McFadden on discrete choice models has been acknowledged by "The Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel 2000". He owns a small farm and vineyard in the Napa Valley. "*Farm work clears the mind, and the vineyard is a great place to prove theorems*".

Figure 8.3: Daniel L. McFadden

- The multivariate extreme value model was first proposed by McFadden (1978), under the name "Generalized Extreme Value" model. In order to avoid any confusion with the Generalized Extreme Value distribution presented in Section B.2.7, and to emphasize that we are dealing with multivariate distributions, we refer to this model as *multivariate extreme value* (MEV).

- In the context of random utility, a random vector $U_n = (U_{1n}, \ldots, U_{J_n n}) = (V_{1n} + \varepsilon_{1n}, \ldots, V_{J_n} + \varepsilon_{J_n})$ with a MEV distribution is such that its CDF is

$$F_{U_n}(\xi_1, \ldots, \xi_{J_n}) = \Pr(U_n \leq \xi_n) = e^{-G(e^{V_{1n}-\xi_1}, \ldots, e^{V_{J_n n}-\xi_{J_n}})}. \tag{8.19}$$

- The marginal distributions of $(U_n)_j$ for $j = 1, \ldots, J_n$ are extreme value distributed, with
  - means:
$$V_{jn} + \frac{\log G(0, \ldots, 1, \ldots, 0) + \gamma}{\mu}, \tag{8.20}$$
    where γ is Euler's constant (4.78),
  - variances: $\pi^2/6\mu^2$, for each j, and
  - moment generating functions
$$e^{tV_{jn}} G(0, \ldots, 1, \ldots, 0)^{\frac{t}{\mu}} \Gamma\left(1 - \frac{t}{\mu}\right), \tag{8.21}$$

where $\Gamma(\cdot)$ is the Gamma function

$$\Gamma(t) = \int_0^{+\infty} z^{t-1} e^{-z} dz.$$

- The variance covariance matrix of a MEV model is derived from its CDF (8.11). The covariance between the error terms of two alternatives i and j is given by

$$\begin{aligned} \text{Cov}(\varepsilon_{in}, \varepsilon_{jn}) &= \text{E}[\varepsilon_{in}\varepsilon_{jn}] - \text{E}[\varepsilon_{in}]\,\text{E}[\varepsilon_{jn}] \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \xi_i \xi_j \frac{\partial^2 F_{\varepsilon_n}(\xi_i, \xi_j)}{\partial \xi_i \partial \xi_j} d\xi_i d\xi_j - \gamma^2, \end{aligned} \tag{8.22}$$

  where $\text{E}[\varepsilon_{in}] = \gamma$,

$$F_{\varepsilon_n}(\xi_i, \xi_j) = F_{\varepsilon_n}(\ldots, +\infty, \xi_i, +\infty, \ldots, +\infty, \xi_j, +\infty, \ldots) \tag{8.23}$$

  is the bivariate marginal cumulative distribution, and

$$\frac{\partial^2 F_{\varepsilon_{in},\varepsilon_{jn}}(\xi_i, \xi_j)}{\partial \xi_i \partial \xi_j} = F_{\varepsilon_{in},\varepsilon_{jn}}(\xi_i, \xi_j) e^{-\xi_i} e^{-\xi_j} (G_i^{ij} G_j^{ij} - G_{ij}^{ij}) \tag{8.24}$$

  where

$$G_i^{ij} = \frac{\partial G(\ldots, 0, e^{-\xi_i}, 0, \ldots, 0, e^{-\xi_j}, 0, \ldots)}{\partial y_i} \tag{8.25}$$

  and

$$G_{ij}^{ij} = \frac{\partial^2 G(\ldots, 0, e^{-\xi_i}, 0, \ldots, 0, e^{-\xi_j}, 0, \ldots)}{\partial y_i \partial y_j}. \tag{8.26}$$

  In the general case, this double integral can only be computed numerically. It is recommended to apply first the change of variables $z_i = \exp(-\exp(-\xi_i))$ and $z_j = \exp(-\exp(-\xi_j))$.

- Contrarily to the probit model, the variance-covariance matrix does not characterize the distribution. Indeed, higher moments of the MEV distribution exist, and different MEV models can share the same variance-covariance (or the same correlation) matrix, as it is illustrated by the example in Section 8.9.

- McFadden's original result was derived from the assumption that G is a 1-homogeneous function. It is always possible through the normalization

$$G(y_1, \ldots, y_J) = G^*(y_1^{1/\mu}, \ldots, y_J^{1/\mu}) \tag{8.27}$$

to convert a μ-homogeneous function G^* into a 1-homogeneous function G. Indeed,

$$\begin{aligned} G(\alpha y_1, \ldots, \alpha y_J) &= G^*((\alpha y_1)^{1/\mu}, \ldots, (\alpha y_J)^{1/\mu}) \\ &= G^*(\alpha^{1/\mu} y_1^{1/\mu}, \ldots, \alpha^{1/\mu} y_J^{1/\mu}) \\ &= (\alpha^{1/\mu})^\mu G^*(y_1^{1/\mu}, \ldots, y_J^{1/\mu}) \\ &= \alpha G(y_1, \ldots, y_J), \end{aligned}$$

where the first and last equations use (8.27), and the third is a consequence of the μ-homogeneity of G^*. In the choice model (8.18), the arguments y of the G function (or its derivatives) are e^V. Therefore, the normalization $y^{1/\mu}$ is $e^{V/\mu}$, which amounts to rescale the V's. Like for the logit model, the μ parameter is not identified from data and can be normalized to one.

- The logarithm of the CPGF is the expected maximum utility of the choice set for the model, that is

$$E[\max_{j \in \mathcal{C}_n} U_{jn}] = \frac{1}{\mu}(\log G(e^{V_{1n}}, \ldots, e^{V_{Jnn}}) + \gamma), \tag{8.28}$$

where γ is Euler's constant. As utilities are defined up to a constant, it is common to ignore the γ. Also, the parameter μ is usually normalized to one. Under this interpretation, the choice model can also be obtained from (3.82), that is

$$P_n(i) = \frac{\partial\, E[\max_{j \in \mathcal{C}_n} U_{jn}]}{\partial V_{in}}, \ \forall i \in \mathcal{C}_n,$$

which is

$$P_n(i) = \frac{\partial \log G(e^{V_{1n}}, \ldots, e^{V_{Jnn}})}{\partial V_{in}} = \frac{e^{V_{in}} G_i(e^{V_{1n}}, \ldots, e^{V_{Jnn}})}{G(e^{V_{1n}}, \ldots, e^{V_{Jnn}})}, \tag{8.29}$$

justifying the name "choice probability generating function". Comparing (8.29) with (8.18), it is seen that G must be such that

$$G(e^{V_{1n}}, \ldots, e^{V_{Jnn}}) = \sum_j e^{V_{jn} + \log G_j\left(e^{V_{1n}}, \ldots, e^{V_{Jnn}}\right)} \tag{8.30}$$

or, equivalently,

$$G(e^{V_{1n}}, \ldots, e^{V_{Jnn}}) = \sum_j e^{V_{jn}} G_j\left(e^{V_{1n}}, \ldots, e^{V_{Jnn}}\right). \tag{8.31}$$

This latter condition actually characterizes homogeneous functions, and is known as *Euler's theorem* (B.112).

- It can be shown that some operations maintain the properties of CPGF functions. Therefore, MEV functions can be constructed from others, and they all correspond to valid choice models. The following results are adapted from the *inheritance theorem* proposed by Daly and Bierlaire (2006). A MEV function which is homogeneous of degree μ is called here a μ-MEV function.

  Consider a choice set $\mathcal{C}$ with J alternatives. Consider also M subsets of alternatives $\mathcal{C}_m$, $m = 1, \ldots, M$, and let J_m be the number of alternatives in subset m. Let $G^m : \mathbb{R}_+^{J_m} \longrightarrow \mathbb{R}$, $m = 1, \ldots, M$ be M μ_m-MEV functions on $\mathcal{C}_m$. Then, the function

$$G : \mathbb{R}_+^J \longrightarrow \mathbb{R} : y \rightsquigarrow G(y) = \sum_{m=1}^{M} (\alpha_m G^m([y]_m))^{\frac{\mu}{\mu_m}} \tag{8.32}$$

  is a μ-MEV function if $\alpha_m > 0$, $\mu > 0$ and $\mu_m \geq \mu$, $m = 1, \ldots, m$, where $[y]_m$ denotes a vector of dimension J_m with entries y_i, where the indices i correspond to the elements in $\mathcal{C}_m$. An application of this result is presented in Section 8.8.

  This result has some interesting corollaries.

  1. If $G(y)$ is a μ-MEV function, so is $\alpha G(y)$, with $\alpha > 0$. The inheritance theorem can be invoked with $M = 1$ and $\mu_m = \mu$.
  2. If $G(y)$ is a μ-MEV function and $\widehat{\mu} \geq 1$, then $G(y^{\widehat{\mu}})^{1/\widehat{\mu}}$ is also a μ-MEV function. Indeed, $G^*(y) = G(y^{\widehat{\mu}})$ is a $(\mu\widehat{\mu})$-MEV function. By the theorem,
$$G^*(y)^{\frac{\mu}{\mu\widehat{\mu}}} = G(y^{\widehat{\mu}})^{\frac{1}{\widehat{\mu}}}$$
     is a μ-MEV function, as $\mu\widehat{\mu} \geq \mu$.
  3. Any linear combination of μ-MEV functions is also a μ-MEV function if the multipliers are non negative and at least one is strictly positive.
  4. If $P_m(i)$ is the choice model derived from the μ_m-MEV function G^m, then the choice model $P(i)$ derived from the μ-MEV function G defined by (8.32) is
$$P(i) = \sum_{m=1}^{M} \frac{(\alpha_m G^m(e^V))^{\frac{\mu}{\mu_m}}}{\sum_{p=1}^{M} (\alpha_p G^p(e^V))^{\frac{\mu}{\mu_p}}} P_m(i). \tag{8.33}$$

8.5 The logit model as MEV

The logit model is a MEV model derived from the following choice probability generating function:

$$G(y) = \sum_{i=1}^{J} y_i^{\mu}. \tag{8.34}$$

Properties [M1]–[M3] are trivially verified:

M1 The strong alternating sign property is a consequence of the fact that $\mu > 0$, $y_i > 0$ and

$$G_i(y) = \frac{\partial G}{\partial y_i} = \mu y_i^{\mu-1}. \tag{8.35}$$

Derivatives of higher orders are all zero:

$$G_{ij}(y) = \frac{\partial^2 G}{\partial y_i \partial y_j} = 0, \text{ if } i \neq j.$$

M2 The function is μ-homogeneous as

$$G(\alpha y) = \sum_{i=1}^{J} (\alpha y_i)^{\mu} = \alpha^{\mu} \sum_{i=1}^{J} y_i^{\mu} = \alpha^{\mu} G(y).$$

M3 The limit property is also verified, as

$$\begin{aligned} G(y_1, \ldots, y_{j-1}, +\infty, y_{j+1}, \ldots, y_J) &= \lim_{y_j \to +\infty} \sum_{i=1}^{J} y_i^{\mu} \\ &= \sum_{i \neq j} y_i^{\mu} + \lim_{y_j \to +\infty} y_j^{\mu} = +\infty. \end{aligned}$$

Consequently, (8.34) defines a μ-MEV function. From (8.11), the CDF is

$$\begin{aligned} F_{\varepsilon}(\xi_1, \ldots, \xi_J) &= e^{-G(e^{-\xi_1}, \ldots, e^{-\xi_J})} \\ &= e^{-\sum_{i=1}^{J} e^{-\mu \xi_i}} \\ &= \prod_{i=1}^{J} e^{-e^{-\mu \xi_i}}, \end{aligned}$$

which is exactly (8.3). Substituting

$$e^{V_i + \log G_i(e^{V_1}, \ldots, e^{V_J})} = e^{V_i + \log \mu + (\mu - 1) \log e^{V_i}} = e^{\log \mu + \mu V_i}$$

into (8.18), we obtain the choice probability

$$P(i) = \frac{e^{\log \mu + \mu V_{in}}}{\sum_{j \in C} e^{\log \mu + \mu V_{jn}}} = \frac{e^{\mu V_{in}}}{\sum_{j \in C} e^{\mu V_{jn}}},$$

which is indeed the choice probability (5.14) of a logit model.

From (8.28), the expected maximum utility is

$$\frac{1}{\mu} \log G\left(e^{V_{1n}}, \ldots, e^{V_{Jn}}\right) = \frac{1}{\mu} \log \sum_{i=1}^{J} e^{\mu V_{in}},$$

which is (5.31) in Section 5.3.3.

We also illustrate the computation of the covariance from (8.22). We consider two distinct alternatives i and j. The double integral in (8.22) writes

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \xi_i \xi_j \frac{\partial^2 F_{\varepsilon_i, \varepsilon_j}(\xi_i, \xi_j)}{\partial \xi_i \partial \xi_j} d\xi_i d\xi_j =$$

$$\int_{-\infty}^{+\infty} \xi_i e^{-e^{-\xi_i}} e^{-\xi_i} d\xi_i \int_{-\infty}^{+\infty} \xi_j e^{-e^{-\xi_j}} e^{-\xi_j} d\xi_j,$$

where index n has been dropped for notational convenience. Applying the change of variable $t_i = e^{-\xi_i}$, we have

$$\int_{-\infty}^{+\infty} \xi_i e^{-e^{-\xi_i}} e^{-\xi_i} d\xi_i = -\int_0^{+\infty} \log t_i e^{-t_i} dt_i = \gamma,$$

so that

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \xi_i \xi_j \frac{\partial^2 F_{\varepsilon_i, \varepsilon_j}(\xi_i, \xi_j)}{\partial \xi_i \partial \xi_j} d\xi_i d\xi_j = \gamma^2. \tag{8.36}$$

Using (8.36) in (8.22), we obtain that the covariance between any pair of alternatives is 0, which is expected for the logit model.

8.6 The nested logit model as MEV

We consider a nested logit model, where the membership of alternatives to nests is characterized by the parameters

$$\alpha_{im} = \begin{cases} 1 & \text{if alternative } i \text{ belongs to nest } m \\ 0 & \text{otherwise.} \end{cases} \tag{8.37}$$

As each alternative belongs to exactly one nest, we have

$$\sum_{m} \alpha_{im} = 1. \tag{8.38}$$

The nested logit model with M nests is a MEV model, with CPGF

$$G(y) = \sum_{m=1}^{M} \left(\sum_{\ell \in \mathcal{C}} (\alpha_{\ell m} y_\ell)^{\mu_m} \right)^{\mu/\mu_m}. \tag{8.39}$$

If $0 < \mu \leq \mu_m$, for all m, (8.39) defines a μ-MEV function. Properties [M2] and [M3] are trivially verified. In order to check the strong alternating sign property [M1], we consider alternative i in nest m. We have

$$G_i(y) = \frac{\partial G}{\partial y_i}(y) = \mu y_i^{\mu_m - 1} \left(\sum_{\ell \in \mathcal{C}} (\alpha_{\ell m} y_\ell)^{\mu_m} \right)^{\frac{\mu}{\mu_m} - 1}. \tag{8.40}$$

It is non-negative when $y \in \mathbb{R}^J_+$. Consider now another alternative $j \neq i$. If j does not belong to the same nest as i, that is nest m, we have $\alpha_{jm} = 0$ and, consequently,

$$\frac{\partial^2 G}{\partial y_i y_j}(y) = 0. \tag{8.41}$$

If j belongs to m, we have

$$\frac{\partial^2 G}{\partial y_i y_j}(y) = \mu\mu_m \left(\frac{\mu}{\mu_m} - 1\right) y_i^{\mu_m - 1} y_j^{\mu_m - 1} \left(\sum_{\ell \in \mathcal{C}} (\alpha_{\ell m} y_\ell)^{\mu_m} \right)^{\frac{\mu}{\mu_m} - 2}. \tag{8.42}$$

We need to verify that (8.42) is non-positive. It is the case as $\mu > 0$ and $\mu \leq \mu_m$. Indeed, all terms in the product are non-negative, except for the term $(\mu/\mu_m - 1)$, which is non-positive. Actually, this condition is exactly (7.30) in Section 7.2. Each additional differentiation involves an additional factor of the form $(\mu/\mu_m - k)$, $k > 1$, which is always negative. Therefore, each differentiation leads to a change of sign, and [M1] is verified. Substituting (8.40) into (8.18), we obtain the nested-logit choice model (7.24).

We now derive the covariance between the error terms of two alternatives i and j in nest m. We first normalize all utilities by μ is order to transform G into a 1-homogenous function, without loss of generality, as suggested by (8.27). Then we consider the bivariate marginal cumulative distribution

$$\begin{aligned} F_{\varepsilon_n}(\xi_i, \xi_j) &= F_{\varepsilon_n}(\ldots, +\infty, \xi_i/\mu, +\infty, \ldots, +\infty, \xi_j/\mu, +\infty, \ldots) \\ &= \exp\left(-G(\ldots, 0, \exp(-\xi_i/\mu), 0, \ldots, 0, \exp(-\xi_j/\mu), 0, \ldots)\right) \\ &= \exp\left(- \left(\exp(-\frac{\mu_m}{\mu}\xi_i) + \exp(-\frac{\mu_m}{\mu}\xi_i) \right)^{\frac{\mu}{\mu_m}} \right). \end{aligned}$$

This is the CDF of a bivariave logistic model (see Kotz et al., 2000, p. 628) with parameter $m = \mu_m/\mu$ and correlation

$$\rho = 1 - m^{-2} = 1 - \frac{\mu^2}{\mu_m^2}. \tag{8.43}$$

The multiple level nested logit model is also a MEV model. Consider for instance a 3-level case, where the choice set is partitioned into p groups, each of them partitioned into M_p nests. The model is derived from the following CPGF:

$$G(y) = \sum_{p=1}^{P} \left(\sum_{m=1}^{M_p} \left(\sum_{i=1}^{J_{mp}} y_i^{\mu_{mp}} \right)^{\mu_p/\mu_{mp}} \right)^{\mu/\mu_p}, \tag{8.44}$$

where J_{mp} is the number of alternatives in the m^{th} nest within group p. It can be verified that the condition:

$$0 \leq \mu \leq \mu_m \leq \mu_{pm}, \text{ for all } m, p,$$

is sufficient for (8.44) to define a CPGF function.

8.7 The cross nested logit model

The cross nested logit model is a MEV model that generalizes the nested logit model by allowing an alternative to belong to more than one nest, in order to capture a wider variety of correlation structures.

We have informally introduced the model in the beginning of the chapter with the airline itinerary choice example. As it involves only three alternatives, the cross nested structure, represented by Figure 8.2, is relatively simple.

As another example, consider a transportation mode choice model with 5 alternatives: ride a bus, ride a train, drive a car, walk or bike. A nested logit structure may be used to capture the correlation between the *bus* and *train* alternatives, as they share unobserved attributes related to public transportation modes, while *car*, *walk* and *bike* share unobserved attributes of private transportation modes. Such a structure is depicted in Figure 8.4. Another nested logit specification may be based on the fact that *bus*, *train* and *car* are motorized modes, while *walk* and *bike* are non-motorized (Figure 8.5). But there is no nested logit structure that captures the fact that the car alternative shares some unobserved attributes with *bus* and *train*, and some other unobserved attributes with *walk* and *bike*. The desired structure, depicted in Figure 8.6, can be modeled by a cross nested logit model.

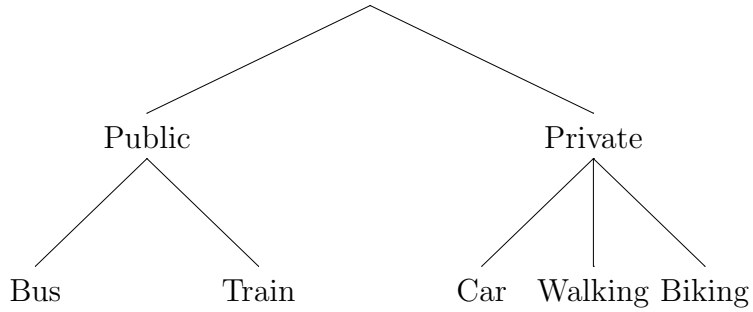


Figure 8.4: A nested logit structure for transportation mode choice

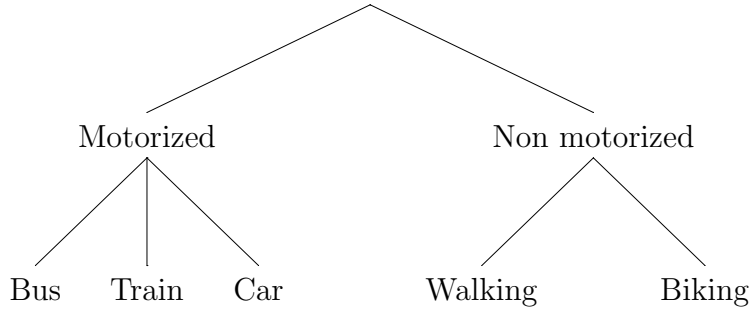


Figure 8.5: Another nested logit structure for transportation mode choice

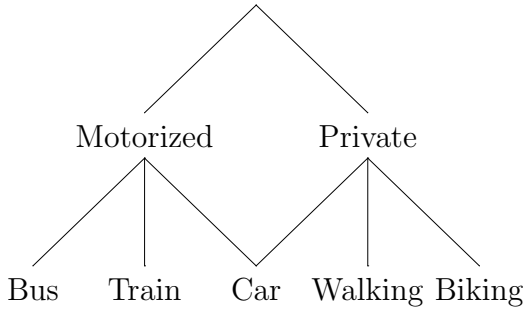


Figure 8.6: A cross nested structure for transportation mode choice

If we consider M nests, the cross nested logit model is a MEV model based on the following CPGF:

$$G(y) = \sum_{m=1}^{M} \left(\sum_{j=1}^{J} \alpha_{jm}^{\frac{\mu_m}{\mu}} y_j^{\mu_m} \right)^{\mu/\mu_m}, \tag{8.45}$$

where μ_m is a parameter associated with nest m (it plays a similar role as the μ_m parameter in the nested logit model), and α_{jm} are parameters capturing the level of membership of alternative j in nest m. We immediately note that the nested logit model is a special instance of the cross nested logit model where $\alpha_{im} = 1$ if alternative i belongs to nest m, and $\alpha_{im} = 0$ otherwise. For this reason, Wen and Koppelman (2001) prefer to call the model the *Generalized Nested Logit Model*, although this terminology does not prevail in the literature.

Bierlaire (2006b) has shown that the conditions

1. $\alpha_{im} \geq 0$, $\forall i, m$,
2. $\sum_m \alpha_{im} > 0$, $\forall i$, and
3. $0 < \mu \leq \mu_m$, $\forall m$,

are sufficient for (8.45) to be a CPGF. Note that condition 3 is the same as for the nested logit model (see Section 8.6).

The CNL model must be normalized before being estimated. As any MEV model, the normalization $\mu = 1$ is applicable. Moreover, if the α parameters are estimated, not all of them are identified and the following normalization is appropriate:

$$\sum_{m=1}^{M} \alpha_{im} = 1, \quad \forall i = 1, \ldots, J. \tag{8.46}$$

This normalization is consistent with the interpretation of α_{im} as the level of membership of alternative i in nest m. The derivative of (8.45) is

$$G_i(y) = \frac{\partial G(y)}{\partial y_i} = \mu \sum_{m=1}^{M} \alpha_{im}^{\frac{\mu_m}{\mu}} y_i^{\mu_m - 1} \left(\sum_{j=1}^{J} \alpha_{jm}^{\frac{\mu_m}{\mu}} y_j^{\mu_m} \right)^{\frac{\mu}{\mu_m} - 1}. \tag{8.47}$$

Substituting (8.47) in (8.18), we obtain the cross nested logit model:

$$P_n(i) = \sum_{m=1}^{M} \frac{\left(\sum_{j \in \mathcal{C}_n} \alpha_{jm}^{\mu_m/\mu} e^{\mu_m V_{jn}} \right)^{\frac{\mu}{\mu_m}}}{\sum_{p=1}^{M} \left(\sum_{j \in \mathcal{C}_n} \alpha_{jp}^{\mu_p/\mu} e^{\mu_p V_{jn}} \right)^{\frac{\mu}{\mu_p}}} \frac{\alpha_{im}^{\mu_m/\mu} e^{\mu_m V_{in}}}{\sum_{j \in \mathcal{C}_n} \alpha_{jm}^{\mu_m/\mu} e^{\mu_m V_{jn}}}, \tag{8.48}$$

which can nicely be interpreted as

$$P_n(i) = \sum_{m=1}^{M} P_n(m|\mathcal{C}_n) P_n(i|m), \tag{8.49}$$

where

$$P_n(i|m) = \frac{\alpha_{im}^{\mu_m/\mu} e^{\mu_m V_{in}}}{\sum_{j\in\mathcal{C}_n} \alpha_{jm}^{\mu_m/\mu} e^{\mu_m V_{jn}}}, \tag{8.50}$$

is the choice probability conditional to nest m, and

$$P_n(m|\mathcal{C}_n) = \frac{\left(\sum_{j\in\mathcal{C}_n} \alpha_{jm}^{\mu_m/\mu} e^{\mu_m V_{jn}}\right)^{\frac{\mu}{\mu_m}}}{\sum_{p=1}^{M} \left(\sum_{j\in\mathcal{C}_n} \alpha_{jp}^{\mu_p/\mu} e^{\mu_p V_{jn}}\right)^{\frac{\mu}{\mu_p}}}, \tag{8.51}$$

is the probability associated with nest m.

The choice model can also be written in a form where the utilities are shifted:

$$P_n(i) = \sum_{m=1}^{M} \frac{\left(\sum_{j\in\mathcal{C}_n} \exp(\mu_m(V_{jn} + \frac{1}{\mu}\log\alpha_{jm}))\right)^{\frac{\mu}{\mu_m}}}{\sum_{p=1}^{M} \left(\sum_{j\in\mathcal{C}_n} \exp(\mu_p(V_{jn} + \frac{1}{\mu}\log\alpha_{jp}))\right)^{\frac{\mu}{\mu_p}}} \frac{e^{\mu_m(V_{in}+\frac{1}{\mu}\log\alpha_{im})}}{\sum_{j\in\mathcal{C}_n} e^{\mu_m(V_{jn}+\frac{1}{\mu}\log\alpha_{jm})}}. \tag{8.52}$$

The correlation structure must be computed using (8.22), where

$$F_{\varepsilon_i,\varepsilon_j}(\xi_i,\xi_j) = \exp\left(-\sum_{m=1}^{M}\left((\alpha_{im}^{\frac{1}{\mu}} e^{-\xi_i})^{\mu_m} + (\alpha_{jm}^{\frac{1}{\mu}} e^{-\xi_j})^{\mu_m}\right)^{\frac{1}{\mu_m}}\right). \tag{8.53}$$

The cross nested logit model provides an intuitive way to capture complex correlation structures. Indeed, any source of correlation assumed by the analyst can be represented by a nest, and the alternatives involved are associated with the nest. As each alternative can potentially belong to more than one nest, a great deal of flexibility is provided. Actually, Fosgerau et al. (2013) have shown than any additive random utility model can be approximated by a cross nested logit model.

8.8 The network MEV model

The family of MEV models contains other members than the logit, the nested logit and the cross nested logit models. The network MEV model introduced

by Daly and Bierlaire (2006) is based on an extension of the representation illustrated in Figure 8.6 for the cross nested logit model. The model is derived from a network representation, as illustrated in Figure 8.7. A network is a set of nodes and directed arcs connecting these nodes. The nodes represent the alternatives, and groups of alternatives (or nests). An arc linking node a to node b means that the alternatives associated with node b also belongs to the set associated with node a. This interpretation is exactly the same as the tree representation of nested logit models.

The network must verify the following properties:

1. it does not contain any loop,
2. there is exactly one node (called the *root*) with no predecessor, and,
3. there are J nodes (called the *leaves*) with no successor, corresponding to the J alternatives in the choice set $\mathcal{C}$.

All the other nodes are *nests*. Several parameters are associated with the network structure.

- Each nest m is associated with a nest parameter μ_m.
- The parameter associated with the root is μ. It cannot be identified and is normalized to 1.
- Each arc linking node m to its successor node p is associated with a parameter α_{pm}, which captures the level of membership, in a similar way as the α parameters of the cross nested logit model.

The model recursively defines for each node m a subset $\mathcal{C}_m$ of the choice set $\mathcal{C}$ and a μ_m-MEV function G^m. The recursion starts at the bottom nodes, corresponding to the alternatives. The subset associated with these nodes contain only one alternative, and the associated 1-MEV function is $G^m : \mathbb{R} \to \mathbb{R} : G(y) = y$. Then, the recursion combines the choice sets. The choice set associated with any node is the union of the choice sets associated with its successors. The MEV function associated with each node is based on the inheritance theorem (8.32).

We describe it with an example. In Figure 8.7, a network is represented, with a set of nodes and arcs. The choice set is composed of 4 alternatives, corresponding to nodes 1, 2, 3 and 4. The nest represented by node 5 is associated with the subset $\mathcal{C}_5 = \{1, 2\}$, and the CPGF

$$G^5(y_1, y_2) = (\alpha_{15} y_1)^{\mu_5} + (\alpha_{25} y_2)^{\mu_5}. \tag{8.54}$$

It is a μ_5-MEV function. Similarly, the nest represented by node 6 is associated with the subset $\mathcal{C}_6 = \{2, 3\}$, and the μ_6-MEV function

$$G^6(y_2, y_3) = (\alpha_{26} y_2)^{\mu_6} + (\alpha_{36} y_3)^{\mu_6}, \tag{8.55}$$

and the nest represented by node 7 is associated with the subset $\mathcal{C}_7 = \{3, 4\}$, and the μ_7-MEV function

$$G^7(y_3, y_4) = (\alpha_{37} y_3)^{\mu_7} + (\alpha_{47} y_4)^{\mu_7}. \tag{8.56}$$

At the upper level, the nest represented by node 8 is associated with the subset $\mathcal{C}_8 = \{1, 2, 3\}$, and the μ_8-MEV function is

$$\begin{aligned} G^8(y_1, y_2, y_3) &= (\alpha_{58} G^5(y_1, y_2))^{\frac{\mu_8}{\mu_5}} + (\alpha_{68} G^6(y_2, y_3))^{\frac{\mu_8}{\mu_6}} \\ &= (\alpha_{58}((\alpha_{15} y_1)^{\mu_5} + (\alpha_{25} y_2)^{\mu_5}))^{\frac{\mu_8}{\mu_5}} \\ &\quad + (\alpha_{68}((\alpha_{26} y_2)^{\mu_6} + (\alpha_{36} y_3)^{\mu_6}))^{\frac{\mu_8}{\mu_6}}. \end{aligned} \tag{8.57}$$

Similarly, the nest represented by node 9 is associated with the subset $\mathcal{C}_9 = \{2, 3, 4\}$, and the CPGF is

$$\begin{aligned} G^9(y_2, y_3, y_4) &= (\alpha_{69} G^6(y_2, y_3))^{\frac{\mu_9}{\mu_6}} + (\alpha_{79} G^7(y_3, y_4))^{\frac{\mu_9}{\mu_7}} \\ &= (\alpha_{69}((\alpha_{26} y_2)^{\mu_6} + (\alpha_{36} y_3)^{\mu_6}))^{\frac{\mu_9}{\mu_6}} \\ &\quad + (\alpha_{79}((\alpha_{37} y_3)^{\mu_7} + (\alpha_{47} y_4)^{\mu_7}))^{\frac{\mu_9}{\mu_7}}. \end{aligned} \tag{8.58}$$

The node corresponding to nest 10 has only one successor. The associated subset is $\mathcal{C}_{10} = \{3, 4\}$ and the CPGF is

$$\begin{aligned} G^{10}(y_3, y_4) &= (\alpha_{7,10} G^7(y_3, y_4))^{\frac{\mu_{10}}{\mu_7}} \\ &= (\alpha_{7,10}((\alpha_{37} y_3)^{\mu_7} + (\alpha_{47} y_4)^{\mu_7}))^{\frac{\mu_{10}}{\mu_7}}. \end{aligned} \tag{8.59}$$

Finally, the root node is associated with the full choice set, and its CPGF is

$$\begin{aligned} G(y_1, y_2, y_3, y_4) = (\alpha_{8r} G^8(y_1, y_2, y_3))^{\frac{\mu}{\mu_8}} + (\alpha_{9r} G^9(y_2, y_3, y_4))^{\frac{\mu}{\mu_9}} + \\ (\alpha_{10r} G^{10}(y_3, y_4))^{\frac{\mu}{\mu_{10}}}, \end{aligned} \tag{8.60}$$

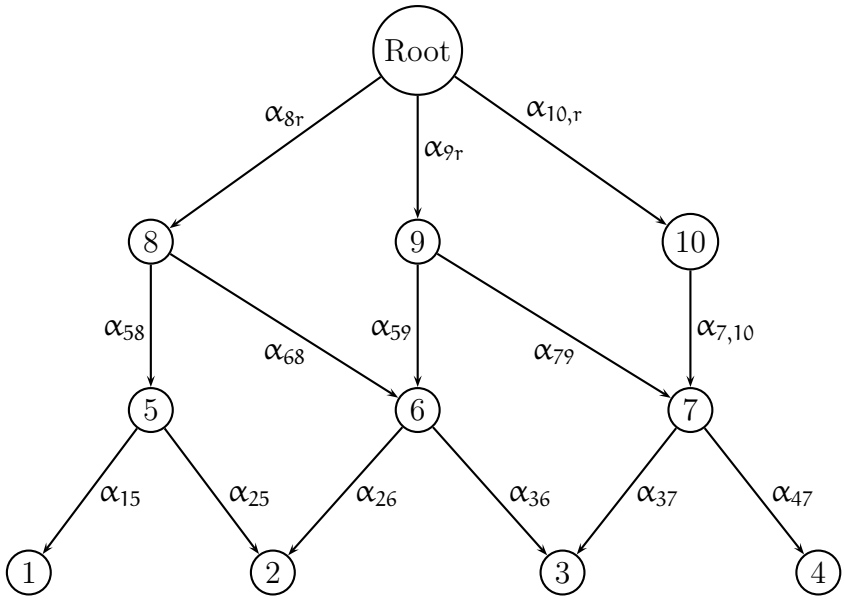


Figure 8.7: A network to derive a network MEV model

that is

$$\begin{aligned}
&G(y_1, y_2, y_3, y_4) = \\
&(\alpha_{8r}((\alpha_{58}((\alpha_{15}y_1)^{\mu_5} + (\alpha_{25}y_2)^{\mu_5}))^{\frac{\mu_8}{\mu_5}} + (\alpha_{68}((\alpha_{26}y_2)^{\mu_6} + (\alpha_{36}y_3)^{\mu_6}))^{\frac{\mu_8}{\mu_6}}))^{\frac{\mu}{\mu_8}} \\
&+ (\alpha_{9r}((\alpha_{69}((\alpha_{26}y_2)^{\mu_6} + (\alpha_{36}y_3)^{\mu_6}))^{\frac{\mu_9}{\mu_6}} + (\alpha_{79}((\alpha_{37}y_3)^{\mu_7} + (\alpha_{47}y_4)^{\mu_7}))^{\frac{\mu_9}{\mu_7}}))^{\frac{\mu}{\mu_9}} \\
&+ (\alpha_{10r}((\alpha_{7,10}((\alpha_{37}y_3)^{\mu_7} + (\alpha_{47}y_4)^{\mu_7}))^{\frac{\mu_{10}}{\mu_7}}))^{\frac{\mu}{\mu_{10}}}. \quad (8.61)
\end{aligned}$$

The inheritance theorem guarantees that each node of the tree, and in particular the root node, is associated with a valid MEV function. It is therefore associated with a valid choice model.

The normalization of the parameters may be complicated, and depends on the topology of the underlying network structure. Newman (2008) proposes a preliminary analysis of the normalization conditions, which may involve non-linear constraints and, therefore, complicate the estimation procedure. The most natural application is a multiple level cross-nested model as presented in the example above.

The advantage of the network MEV approach is that the model is entirely determined by the topology of the underlying network. The MEV function is obtained in a constructive and recursive way using the inheritance theorem, as illustrated above, and the choice model derived directly from this MEV

function. Note that, in most practical cases, it is seldom useful to develop a model involving more than two layers of nests.

8.9 Airline itinerary choice

We revisit here the airline itinerary choice example. And we consider again a cross-nested logit model based on the structure represented by Figure 8.2, where the alternative SAME belongs to both nests "One airline" and "One stop". In Section 8.1, the degree of membership has been arbitrarily set to 0.5 for each nest. We now relax this assumption and estimate the degree of membership from data. All parameters (known and unknown) associated with the cross-nested structure are reported in Table 8.2. A value of 0 means that the alternative does not belong to the corresponding nest. A value of 1 means that the alternative "fully" belongs to the nest. For the "One stop, same airline" alternative, the "degree of membership" to the nest "One stop" is represented by the parameter α, to be estimated from data. Consequently, the degree of membership to the nest "One airline" is $1-\alpha$. As the value of α is between 0 and 1, so is the value of $1-\alpha$. Note that for each alternative, the sum of the associated parameters is always equal to 1. Like the nested logit model, each nest of the cross nested logit model is associated with a nest parameter, which must be larger than 1. Therefore, we obtain a model with 18 parameters: 15 parameters in the utility function, two nest parameters, and one membership parameter.

The estimation of this model leads to an invalid model. Indeed, the estimated value for the parameter of the nest "One airline" is 0.785, which is lesser than 1. The model is therefore rejected, and this nest parameter is constrained to 1, leading to a model with 17 parameters. The result of the estimation is reported in Table 8.3.

Comparing the nested logit model (Table 7.5) with the cross nested logit model, first note the increase of the log likelihood function from -1613.858 to -1611.670. A likelihood ratio test rejects the nested logit model at the 95% level[4]. Indeed,

$$-2(-1613.858-(-1611.670))=4.32$$

which is above the 95% quantile of the χ^2 with one degree of freedom, which is 3.84. It shows that the improvement in fit is significant, and the nested logit model is rejected.

[4]The likelihood ratio test can be applied as the nested logit model is a restricted version the cross nested logit model.

The parameter associated with the nest "One stop" is 2.19, which is consistent with the condition that it must be larger than 1. It is also significantly different from 1, as the t-test is 3.72. The cross nested parameter α is equal to 0.798. The t statistic against 0 is 8.98. The t statistic against 1 is 2.27. Therefore, the cross nested parameter is significantly different from both 0 and 1. This shows that the additional flexibility obtained from relaxing the requirement about the unambiguous assignment of alternatives to nests allows to fit the data better.

The correlation matrix corresponding to this model is block diagonal:

$$\Sigma_{\text{CNL}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.695 \\ 0 & 0.695 & 1 \end{pmatrix}$$

It has the same structure as the correlation matrix of a nested logit model, because the nest parameter of the "One airline" nest is equal to 1, and alternatives within this nest are not correlated. Therefore, we have two different models with the exact same correlation structure. It illustrates the fact that, contrarily to probit models, the correlation matrix of a cross nested logit model does not fully characterize it.

To show that, we have estimated the 15 parameters of a nested logit model associated with the correlation matrix Σ_{CNL} (not only the structure but also the values). From (7.32), the value of its nest parameter is

$$\sqrt{\frac{1}{1-0.695}} = 1.810.$$

Constraining the parameter to that value, we re-estimate the other parameters. We obtain a final log likelihood equal to -1613.866. It does not fit the data as well as the cross nested logit model, providing an evidence that it is not the same model, although it has the exact same correlation matrix. It illustrates that there exists different cross nested logit models sharing the same correlation matrix.

8.10 Summary

Multivariate Extreme Value models represent a family of choice models accounting for the correlation among the error terms of the utility functions. They are characterized by a choice probability generating function (CPGF), denoted by G, if G verifies some technical properties:

Alternative	Nest	Cross nested param.	
Non stop	One airline	1	fixed
Non stop	One stop	0	fixed
One stop, same airline	One airline	$1-\alpha$	calculated
One stop, same airline	One stop	α	estimated
One stop, multiple airlines	One airline	0	fixed
One stop, multiple airlines	One stop	1	fixed

Table 8.2: Airline itinerary choice example: the cross nested parameters

M1 Strong Alternating Sign Property The alternating sign property of G states that the cross partial derivatives of G have alternative signs, that is for any set of $\widehat{j}$ distinct indices $i_1, \ldots, i_{\widehat{j}}$,

$$(-1)^{\widehat{j}-1} G_{i_1,\ldots,i_{\widehat{j}}}(y) \geq 0. \tag{8.62}$$

M2 μ-homogeneity property The μ-homogeneity property states that, for each $\alpha > 0$ and $y \in \mathbb{R}^J_+$,

$$G(\alpha y) = \alpha^\mu G(y). \tag{8.63}$$

M3 Limit property The limit property states that, for $i = 1, \ldots J$,

$$\lim_{y_i \to +\infty} G(y_1, \ldots, y_i, \ldots, y_J) = +\infty. \tag{8.64}$$

Given G, the MEV model is defined as

$$P_n(i) = \frac{e^{V_{in} + \log G_i\left(e^{V_{1n}}, \ldots, e^{V_{Jn}n}\right)}}{\sum_j e^{V_{jn} + \log G_j\left(e^{V_{1n}}, \ldots, e^{V_{Jn}n}\right)}},$$

where G_i denotes the partial derivative of G with respect to its i^{th}argument. The MEV models can be interpreted as a logit model where the utility of each alternative is adjusted to account for the correlation among the error terms. The adjustment term is a function of all utilities in the model. The logarithm of the CPGF is the expected maximum utility of the associated choice model, that is

$$E[\max_{j \in \mathcal{C}_n} U_{jn}] = \frac{1}{\mu} \log G(e^{V_{1n}}, \ldots, e^{V_{Jn}n}).$$

Parameter number	Description	Coeff. estimate	Asympt. std. error	t-stat	p-value
1	One stop, same airline dummy	-0.703	0.165	-4.27	0.00
2	One stop, multiple airlines	-0.975	0.172	-5.67	0.00
3	Travel time (hours) (0–2 hours)	-0.806	0.214	-3.76	0.00
4	Travel time (hours) (2–8 hours)	-0.182	0.0593	-3.07	0.00
5	Travel time (hours) ($\geq$ 8 hours)	-0.866	0.271	-3.20	0.00
6	Round trip fare ($100) / Income ($1000)	-18.8	7.53	-2.50	0.00
7	Round trip fare ($100)	-1.54	0.150	-10.26	0.00
8	More than two air trips per year (one stop, same airline)	-0.244	0.123	-1.99	0.05
9	More than two air trips per year (one stop, multiple airlines)	-0.109	0.131	-0.83	0.41
10	Leg room (inches), if female (non-stop)	0.179	0.0296	6.06	0.00
11	Leg room (inches), if male (non-stop)	0.0918	0.0309	2.97	0.00
12	Leg room (inches), if female (one-stop)	0.0607	0.0187	3.24	0.00
13	Leg room (inches), if male (one-stop)	0.0952	0.0211	4.52	0.00
14	Being early (hours)	-0.127	0.0157	-8.10	0.00
15	Being late (hours)	-0.0711	0.0141	-5.03	0.00
16	μ One stop	2.19	0.320	3.72[1]	0.00
17	α One stop / One stop, same airline	0.798	0.0889	8.98	0.00

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\hat{\beta}) &= -1611.670 \\ -2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 2366.400 \\ \rho^2 &= 0.423 \\ \bar{\rho}^2 &= 0.417 \end{aligned}$$

[1] t-test against 1

Table 8.3: Airline itinerary: cross nested logit

The CDF of the error terms is given by

$$F_{\varepsilon_n}(\varepsilon_{1n}, \ldots, \varepsilon_{Jn}) = e^{-G(e^{-\varepsilon_{1n}}, \ldots, e^{-\varepsilon_{Jn}})}.$$

The most important instances of the family are the logit, the nested logit and the cross nested logit models, the latter allowing to capture more general correlation structures:

$$P_n(i) = \sum_{m=1}^{M} \frac{\left(\sum_{j \in \mathcal{C}_n} \alpha_{jm}^{\mu_m/\mu} e^{\mu_m V_{jn}}\right)^{\frac{\mu}{\mu_m}}}{\sum_{p=1}^{M} \left(\sum_{j \in \mathcal{C}_n} \alpha_{jp}^{\mu_p/\mu} e^{\mu_p V_{jn}}\right)^{\frac{\mu}{\mu_p}}} \frac{\alpha_{im}^{\mu_m/\mu} e^{\mu_m V_{in}}}{\sum_{j \in \mathcal{C}_n} \alpha_{jm}^{\mu_m/\mu} e^{\mu_m V_{jn}}},$$

Generating functions can be combined to generate more complex valid generating function, inheriting the required properties. The network MEV model is derived directly from this inheritance.

Chapter appendix

The MEV model is first derived in Section 8.A. Finally, a short discussion about the link between MEV models and copulas in provided in Section 8.B.

8.A Derivation of the MEV model

As discussed in Section 8.3, the choice model is obtained by incorporating (8.11)

$$F_{\varepsilon_n}(\varepsilon_{1n}, \dots, \varepsilon_{Jn}) = e^{-G(e^{-\varepsilon_{1n}}, \dots, e^{-\varepsilon_{Jn}})},$$

into (3.62)

$$P_n(i) = \int_{\varepsilon=-\infty}^{+\infty} \frac{\partial F_{\varepsilon_{1n}, \varepsilon_{2n}, \dots, \varepsilon_{Jn}}}{\partial \varepsilon_i}(\dots, V_{in} - V_{(i-1)n} + \varepsilon, \varepsilon, V_{in} - V_{(i+1)n} + \varepsilon, \dots) d\varepsilon.$$

We have

$$\begin{aligned}
&\frac{\partial F_{\varepsilon_{1n}, \varepsilon_{2n}, \dots, \varepsilon_{Jn}}}{\partial \varepsilon_i}(\dots, V_{in} - V_{(i-1)n} + \varepsilon, \varepsilon, V_{in} - V_{(i+1)n} + \varepsilon, \dots) \\
&= e^{-\varepsilon} G_i(\dots, e^{-V_{in}+V_{(i-1)n}-\varepsilon}, e^{-\varepsilon}, e^{-V_{in}+V_{(i+1)n}-\varepsilon}, \dots) \\
&\qquad \exp\left(-G(\dots, e^{-V_{in}+V_{(i-1)n}-\varepsilon}, e^{-\varepsilon}, e^{-V_{in}+V_{(i+1)n}-\varepsilon}, \dots)\right) \\
&= e^{-\varepsilon} e^{-(\mu-1)\varepsilon} e^{-(\mu-1)V_{in}} G_i(\dots, e^{V_{(i-1)n}}, e^{V_{in}}, e^{V_{(i+1)n}}, \dots) \\
&\qquad \exp\left(-e^{-\mu\varepsilon} e^{-\mu V_{in}} G(\dots, e^{V_{(i-1)n}}, e^{V_{in}}, e^{V_{(i+1)n}}, \dots)\right),
\end{aligned}$$

because G is μ-homogeneous, which implies that G_i is $(\mu-1)$-homogeneous. We now denote

$$e^V = \left(\dots, e^{V_{(i-1)n}}, e^{V_{in}}, e^{V_{(i+1)n}}, \dots\right),$$

and simplify the terms to obtain

$$\begin{aligned}
&\frac{\partial F_{\varepsilon_{1n}, \varepsilon_{2n}, \dots, \varepsilon_{Jn}}}{\partial \varepsilon_i}(\dots, V_{in} - V_{(i-1)n} + \varepsilon, \varepsilon, V_{in} - V_{(i+1)n} + \varepsilon, \dots) \\
&= e^{-\mu\varepsilon} e^{-\mu V_{in}} e^{V_{in}} G_i(e^V) \exp\left(-e^{-\mu\varepsilon} e^{-\mu V_{in}} G(e^V)\right).
\end{aligned}$$

Therefore,

$$P_n(i) = e^{-\mu V_{in}} e^{V_{in}} G_i(e^V) \int_{\varepsilon=-\infty}^{+\infty} e^{-\mu\varepsilon} \exp\left(-e^{-\mu\varepsilon} e^{-\mu V_{in}} G(e^V)\right) d\varepsilon.$$

Defining $t = -\exp(-\mu\varepsilon)$, so that $dt = \mu\exp(-\mu\varepsilon)d\varepsilon$, we write

$$P_n(i) = e^{-\mu V_{in}} e^{V_{in}} G_i(e^V) \frac{1}{\mu} \int_{t=-\infty}^{0} \exp\left(te^{-\mu V_{in}} G(e^V)\right) dt,$$

which simplifies to

$$P_n(i) = \frac{e^{V_{in}} G_i(e^V)}{\mu G(e^V)}.$$

We finally invoke Euler's theorem (B.112) that characterizes homogeneous functions to obtain (8.18):

$$P_n(i) = \frac{e^{V_{in} + \log G_i\left(e^V\right)}}{\sum_j e^{V_{jn} + \log G_j(e^V)}}.$$

8.B Copulas

The concept of MEV functions is closely related to the concept of copulas in statistics. A copula is the CDF of a multivariate distribution such that every marginal distribution is uniform in the interval $[0, 1]$. We refer the reader to Nelsen (2006) for an introduction to copulas. A result by Sklar (1959) states that any multivariate distribution is entirely characterized by its marginals, and a copula. Intuitively, the copula captures the dependence among the various dimensions. More precisely, consider the CDF of a multivariate random vector ε

$$F(\varepsilon_1, \ldots, \varepsilon_J)$$

and denote $F_j(\varepsilon_j)$ the CDF of its univariate marginal distribution associated with dimension j. The copula of F is defined as

$$C : [0, 1]^J \to \mathbb{R} : (u_1, \ldots, u_J) \to C(u_1, \ldots, u_J) = F\left(F_1^{-1}(u_1), \ldots, F_J^{-1}(u_J)\right),$$

where $F_j^{-1} : [0, 1] \to \mathbb{R}$ denotes the inverse function of the marginals. Conversely, given a copula C, multivariate distributions can be constructed from the marginal distributions $F_j(\varepsilon_j)$:

$$F(\varepsilon_1, \ldots, \varepsilon_J) = C\left(F_1(\varepsilon_1), \ldots, F_J(\varepsilon_J)\right).$$

The link between copulas and MEV function is given by the following result (Joe, 1997): a multivariate random variable with CDF $F(\varepsilon_1, \ldots, \varepsilon_J)$ has a MEV distribution if and only if its copula satisfies the following condition:

$$C(u_1, \ldots, u_J)^\alpha = C(u_1^\alpha, \ldots, u_J^\alpha),$$

for $u \in [0,1]^J$ and $\alpha > 0$. Actually, $\log C$ plays the role of the choice probability generating function defined above.

These results, although quite technical, can be exploited to generate new MEV models from the theory of copulas. We refer the interested reader to Nikoloulopoulos and Karlis (2008), Bhat and Sener (2009) or Fosgerau et al. (2013) for more details.

Chapter 10

Prediction

"*Prediction is very difficult, especially about the future*", Niels Bohr.

Contents

Up to this point we have focused almost exclusively on the problem of predicting individual behavior. The choice models derived in previous chapters predict the probabilities with which any particular individual will take various actions. However, predictions for a specific individual are generally of little use in helping to make investment or planning decisions. Instead, most real-world decisions are based (at least in part) on the forecast of some aggregate demand, such as the market shares of competing products or services. Some linkage between the disaggregate level models described previously and the aggregate level forecasts of interest to planners and decision makers is obviously needed. In addition to forecast of market shares, several indicators for policy analysis can be derived from the choice models.

This chapter focuses on various usages of choice models after their parameters have been estimated. Sections 10.1 and 10.2 describe how disaggregate choice models can be aggregated to derive indicators at the population level, with a special emphasis on market shares. Sections 10.3 and 10.4 deal with practical issues related to the application of the choice model for future scenarios: the calibration of the constants to match aggregate market shares, and the inclusion of a non existing alternative in the model. Various useful indicators for policy analysis are then discussed in Section 10.5. Section 10.6 deals with the important issue of reliability of the various indicators, explaning how to perform a sensitivity analysis and to derive confidence intervals. Before summarizing the chapter in Section 10.8, Section 10.7 illustrates the various concepts introduced in the chapter on a real case study.

10.1 Aggregate forecasting

The first issue to resolve in considering the problem of making an aggregate forecast is to define the relevant aggregate population. In some cases this may be quite simple. For example, we may be interested in launching a product in a given market (Switzerland, say). The relevant population is therefore the population of Switzerland. In some cases, we may target specific submarkets, such as some age categories for instance.

In the context of travel demand, we may be interested in total transit ridership within an urban area, and the relevant population will therefore be all residents of the city. For other purposes there may be many relevant subpopulations. We might, for example, want separate aggregate forecasts for different income groups, or for different origin-destination pairs. For instance, consider the case of Biogeme Airways, who plans to open a new service between Chicago and San Diego. The population of interest is then

all individuals traveling from Chicago to San Diego during a given time period (typically a year), or only those of them traveling in economy class.

Assume that the populations for which aggregate forecasts are needed have been defined. Consider any one of these populations and denote it as T. The next relevant issue is measuring the number of decision makers in T. Generally, this is done by drawing on existing data sources such as census counts, utility records, or telephone directories. In some cases, it may be necessary to estimate the population's size by conducting a supplementary survey or by combining information from a variety of sources. In the example of Biogeme Airways, the information can be collected from the Bureau of Transportation Statistics, from historical data of the airline and its competitors, or by conducting surveys.

Assuming that the size of T is known; we now turn to the core of the problem of aggregating across individuals — predicting the share of the population choosing each alternative. Let us denote the number of decision makers as N_T. We write the probability that an individual n in T chooses an alternative i as $P_n(i|x_n, \mathcal{C}_n)$, where x_n is defined as all the independent or explanatory variables affecting the choice that appear in some way in the model, regardless of which utility function they appear in, and $\mathcal{C}_n$ is the choice set considered or available to individual n. For the Biogeme Airways example, the company may plan to propose two itineraries, one non stop and one with one stop, so that the choice set would be composed of two alternatives for each individual:

$$\mathcal{C}_n = \{\text{Non stop}, \text{One stop–same airline}\}. \tag{10.1}$$

Considering the nested logit model described in Table 7.5, x_n would have all the following entries:

- characteristics of the traveler: gender, desired arrival time, number of trips per year, and income;
- attributes of each itinerary: fare, elapsed time, leg room, and arrival time.

If we knew the value of x_n *for every member of the population* T, then making an aggregate prediction would be at least conceptually straightforward. The total expected number of individuals in T choosing any alternative i, denoted by $N_T(i)$, would simply be

$$N_T(i) = \sum_{n=1}^{N_T} P_n(i|x_n, \mathcal{C}_n). \tag{10.2}$$

The above definitions are summarized in a tabular form in Table 10.1. The entries are the disaggregate probabilities and the aggregate demands are the column sums.

Population	Alternatives				Total
	1	2	$\cdots$	J	
1	$P(1\|x_1)$	$P(2\|x_1)$	$\cdots$	$P(J\|x_1)$	1
2	$P(1\|x_2)$	$P(2\|x_2)$	$\cdots$	$P(J\|x_2)$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N_T	$P(1\|x_{N_T})$	$P(2\|x_{N_T})$	$\cdots$	$P(J\|x_{N_T})$	1
Total	$N_T(1)$	$N_T(2)$	$\cdots$	$N_T(J)$	N_T

Table 10.1: Disaggregate Probabilities and Aggregate Demands

It should be noted that $N_T(i)$ is the *expected value* of the aggregate number of individuals in the population choosing i. It is both a consistent and unbiased estimate of the actual number of people choosing i. Since the choices of individuals are probabilistic, the actual aggregate number choosing i is a random variable. In most real-world forecasting situations, T is large enough so that the distinction between the actual share of the population using i and its expected value is negligible. In what follows we will assume that the population is large and ignore this distinction. Since for any alternative i each individual's choice can be viewed as a Bernoulli event with probabilities $P_n(i|x_n)$ and $1 - P(i|x_n)$, respectively, the actual share of the population using it is the average of N_T Bernoulli variables. If the choice probabilities are known with certainty, we can show that the variance of the aggregate usage of i must be less than or equal to $N_T/4$.

A slightly more convenient form of equation (10.2) is based on a prediction of the *share* of the population choosing i. If we let $W(i)$ denote the fraction of population T choosing alternative i, then

$$W(i) = \frac{1}{N_T} \sum_{n=1}^{N_T} P(i|x_n) = E\left[P(i|x_n)\right]. \tag{10.3}$$

The use of equation (10.2) or (10.3) requires that we know each individual's complete vector of choice-relevant variables, which is not possible in general. If the distribution of the attributes x_n in the population T is continuous and represented by the density function $p(x)$, then equation (10.3) can be expressed as

$$W(i) = \int_x P_n(i|x, \mathcal{C}) p(x) dx = E\left[P_n(i|x, \mathcal{C})\right], \tag{10.4}$$

where it is assumed that each individual are faced with the same choice set. If it is not the case, this quantity can be computed for each group sharing the same choice set.

However, $p(x)$ is generally unknown, and even if it were known, the computational burden of evaluating this integral may be prohibitive when the number of x's is large.

Stated briefly, *the problem of aggregating across individuals is to develop methods for reducing the required data and computation needed to predict aggregate usage of various alternatives.*

Finally, the reader should recognize that in actual applications the analyst never knows $P_n(i|x_n, \mathcal{C}_n)$. Only an estimate of it is available because the underlying parameters are unknown.

10.2 Aggregation Methods

The methods for aggregating across individuals approximate equation (10.2) or (10.3) in some way, thereby reducing the needed data and computation at the expense of the accuracy of the forecast. From the analyst's point of view the goal is to find a method that provides the best combination of accuracy and cost for a specific forecasting situation. Each procedure reduces the problem of aggregating forecasts across individuals by making some simplifying assumptions about the choice model, the population, or both.

We define four general types of aggregation procedures:

1. Average individual: the population is divided into G nearly homogeneous subgroups with sizes $N_{T_1}, N_{T_2}, ..., N_{T_G}$. An "average individual" is constructed for each subgroup and the choice probability for that average individual is calculated within each subgroup. $N_T(i)$ is estimated as the weighted sum of the G average individual forecasts, where the weights are the values of N_{T_g}, $g = 1, ..., G$. This is described in details in Section 10.2.1.

2. Enumeration: the data on individual decision-makers is used to calculate predicted values of individual probabilities. Such data may represent the entire population, a synthetically generated population (Section 10.2.2), or a sample from the population (Section 10.2.3). Equation (10.3) is then applied with these individual probabilities, and the

resulting value of $W(i)$ is used as an estimate of the population's value:

$$W(i) \cong \frac{1}{N} \sum_{n=1}^{N} P(i|x_n)$$

where N is the number of individuals in the synthetic population or the sample.

3. Microsimulation: Use the choice probabilities in a simulation to draw choice realizations and then apply (10.3) with $(0, 1)$ realizations replacing the individual probabilities (Section 10.2.4).

In the following sections we develop the theory of using each of these methods in greater detail and comment on some of their properties.

10.2.1 Average individual

To apply the average individual-aggregation procedure, we create a "representative individual," using his or her characteristics to represent the entire population. More formally, define $\bar{x}$ as the mean of $p(x)$. Then, in using this procedure, we approximate $W(i)$ as $P(i|\bar{x})$.

A simple graphical example serves to illustrate the error that the approximation produces. Suppose the population consists of exactly two individuals with characteristics x_1 and x_2, respectively, and suppose further that x is a scalar. Let $P(i|x)$ be as illustrated in figure 10.1. Here $P(i|\bar{x})$ is marked as the forecast, and $W(i)$ denotes the actual value, where

$$W(i) = \frac{P(i|x_1) + P(i|x_2)}{2}. \tag{10.5}$$

The difference, denoted by Δ, is what we will term the aggregation error. Consider, for example, a simple binary logit model $P(i|x_n) = 1/(1 + e^{-x_n})$, and the following numerical values:

$$\begin{aligned} x_1 &= 2.0, \\ x_2 &= 0.5, \\ \bar{x} &= 1.25. \end{aligned}$$

The correct share of alternative 1 is found to be $P(1|x_1) = 0.88$, $P(1|x_2) = 0.62$, so that $W(1) = 0.75$. The average individual forecast is $P(1|\bar{x}) = 0.78$; hence the aggregation error is 0.03. Note that because both x_1 and x_2 are

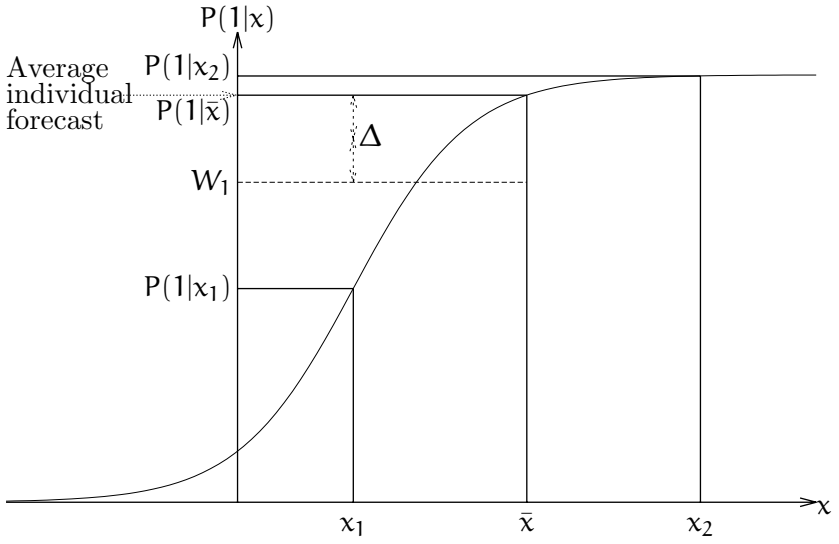


Figure 10.1: Error due to average individual aggregation

positive, the aggregation is performed over a concave region of $P(1|x_n)$, and therefore we have in this example $P(1|\bar{x}) > W(1)$.

The average individual procedure has the following properties:

1. If the choice model is linear over the range of x in the population, Δ will be zero. This is because for any *linear* function h,

$$E[h(x)] = h(E[x]).$$

   Generally, $|\Delta|$ is small when $h(x)$ is nearly linear (in a loosely defined sense).

2. The value of Δ can be positive or negative depending both on the form of $P(i|x)$ and $p(x)$. If $P(i|x)$ is concave over the range of x in the population, $E[P(i|x)] \leq P(i|\bar{x})$, and the average individual choice probability may overestimate $W(i)$. Conversely, if $P(i|x)$ is convex over the range of x, then $P(i|x)$ may underestimate $W(i)$.

3. Koppelman (1975) analyzes the value of Δ under some simplified conditions. Let

$$\bar{V}_i = \int_{x_i} V(x_i)p(x_i)dx_i, \tag{10.6}$$

where x_i is the subvector of x affecting the systematic utility of alternative i and $p(x_i)$ is its distribution. For example, if V is linear in its parameters, then $\bar{V}_i = \beta^T \bar{x}_i$. Koppelman shows that under certain conditions there will be only one value of V for which $\Delta = 0$. Also the skewness of the distribution of the V's will affect Δ.

σ^2_V (variance	$\lvert\bar{V}\rvert$ (mean of $V_{in} - V_{jn}$)						
of $V_{in} - V_{jn}$)	0.0	0.5	1.0	1.5	2.0	2.5	3.0
1	0	0.05	0.08	0.07	0.06	0.03	0.02
2	0	0.08	0.12	0.12	0.10	0.06	0.04
3	0	0.09	0.15	0.16	0.14	0.10	0.07
∞	0	0.19	0.34	0.43	0.48	0.49	0.499

Table 10.2: $|\Delta|$ for average individual aggregation as function of mean and variance of utilities

σ^2_V (variance	$\lvert\bar{V}\rvert$ (mean of $V_{in} - V_{jn}$)						
of $V_{in} - V_{jn}$)	0.0	0.5	1.0	1.5	2.0	2.5	3.0
1	0	7.8	10.5	8.1	6.5	3.1	2.0
2	0	13.1	16.7	14.8	11.4	6.5	4.2
3	0	15.0	21.7	20.8	16.7	11.2	7.5
∞	0	38.0	68.0	86.0	96.0	98.0	100.0

Table 10.3: $|\Delta|$ as a percentage of $W(i)$ for average individual aggregation procedure

4. In simple numerical examples the value of $|\Delta|$ can be shown to increase as the variance of the distribution $p(x)$ grows. The errors for even reasonable situations can be quite substantial. Consider, for example, a simple binary choice situation in which the probit model is used:

$$P_n(i) = \Phi(V_{in} - V_{jn}), \tag{10.7}$$

and $V_{in} - V_{jn}$ is normally distributed across the population with mean $\bar{V}$, variance σ^2_V. In this case $|\Delta|$ is given by table 10.2 as a function of $|\bar{V}|$ and σ^2_V. Table 10.3 expresses $|\Delta|$ as a percentage of the true value of $W(i)$. Note that the true value of $W(i)$ exceeds 0.5 for all the cases evaluated. The percentage errors when — $|\bar{V}|$ is used can be much greater because $W(i)$ will be small.

Clearly, the average individual aggregation applied to the full population is in general most inaccurate when the distribution of x has high variance. To

address that issue, we partition the population into relatively homogeneous subgroups (so that the variance of x within each group is low) and simply apply average individual aggregation to each subgroup.

Stated more formally, the method proceeds in the following steps:

1. Partition the population T into G mutually exclusive, collectively exhaustive subgroups, each corresponding to the portion of the population with a range of the attribute vector equal to set $\{X_g\}$, $g = 1, ..., G$.

2. Estimate the number of decision makers in each group, denoted by N_{T_g}, $g = 1, ..., G$.

3. For each group choose some representative value $\bar{x}_g$.

4. Approximate $W(i)$ as

$$W(i) \cong \sum_{g=1}^{G} \frac{N_{T_g}}{N_T} P(i|\tilde{x}_g). \tag{10.8}$$

Note that Equation (10.8) is equivalent to the approximation of the integral in equation (10.4) by the trapezoidal rule (see, for instance, Atkinson, 1989).

The key to classification lies in the partitioning of the entire space of possible attributes into subsets. Usually as the number of groups rises, the accuracy of the approximation in equation (10.8) will increase, but so will the data requirements and computational difficulties associated with making a forecast. Establishing a trade-off between the number of groups and the accuracy of the forecast represents an area of judgment for the analyst. Some useful insights into how different classifications affect the errors due to aggregation can be given.

1. It is important to stress that the attributes in x_n are combined in the model to form the systematic utility. The goal of a good classification system is therefore not to establish groups in which the within-group in x is small. *Rather, we want groups with small within-group variation in the V's.* Reid (1978) suggests classifying directly on the values of the systematic utilities rather than the underlying variables in x. This is a reasonable approach but requires knowledge of the fraction of the population in each utility class, which can only be derived by integrating $p(x)$ over various subsets. Inasmuch as the only practical way of doing this is often via classification of x, there may be little gained from utility-based classification.

2. It is generally infeasible to classify by every dimension in x. For example, if there are 6 different variables that constitute x, even if we divided each variable into only two categories, there would be $2^6 = 64$ different groups!

3. As a consequence of points 1 and 2, it generally makes sense to select a small number of independent variables on which to classify the population. In a loose sense the variables chosen should be "important" to the choice process in that they have a large effect on the systematic utility of at least one alternative and have a wide distribution across the population.

4. Each of the G classes contributes to the forecast $W(\mathsf{i})$ in proportion to its share of the population, N_{T_g}/N_T. All else equal, it therefore makes sense to avoid classes that are disproportionately small because the computational effort associated with forecasting for any single class is independent of the size of the class.

5. Classes defined for portions of the population for which $P(\mathsf{i}|\mathsf{x})$ is nearly linear do not need to be finely subdivided. This is because average individual aggregation works quite well for subpopulations with nearly linear choice probabilities.

6. Some classes will often have open sets $\{X_g\}$, making it difficult to define $\tilde{x}_g$. For example, if income were used as a basis for classification, one class might be all households with income over \$30,000 per year. It may be useful to obtain detailed information on the conditional distribution of income, given it exceeds \$30,000, and solve for the conditional expected value of income. One might also define the upper income range to include a small fraction of the population, thereby reducing the influence of the chosen $\tilde{x}_g$ on the total forecast $W(\mathsf{i})$. (This is only an apparent contradiction to guideline 4. The important aspect of point 4 was the phrase "all else equal." In this case the estimated value of $\tilde{x}_g$ for an open set $\{X_g\}$ may have much greater error than for the other groups.)

7. The different classes can vary in their available choices as well as their values of $\tilde{x}_g$.

8. Some variables that are components of x are often classified automatically as part of a demand analysis. For example, we generally divide a study area into geographical zones and predict electricity demand consumption between zones as a function of inter zonal temperature,

costs, and so forth. The zones serve as a "natural" classification for the cost variables that enter into the choice model.

9. To make forecasts using equation (10.8), it will be necessary to predict the values of class sizes $N_{T_g}, g = 1, ..., G$. This may be more difficult for some classification schemes than for others. For example, in a travel choice demand study, auto ownership is often predicted as part of a typical urban travel demand analysis, and it often is simple to use or adapt those forecasts for the purposes of using classification.

10.2.2 Synthetic population

The calculation of the population shares using (10.3) requires the knowledge of the complete vector of variables for each individual in the population. While it is not possible in general to have access to this information for the real population, it is possible to generate a synthetic population, where each synthetic individual is associated with the complete set of relevant variables. The synthetic population must share with the real population the same aggregate properties. Typically, the distribution of key variables, as well as their relevant interactions should be the same in the real and the synthetic population.

It is sometimes easier to describe the synthetic population as a *contingency table*. A contingency table is characterized by a list of discrete variables $x_1, \ldots, x_K$, where the variable x_i can take L_i different values. For example, we can consider $K = 2$ variables: x_1 is `age` and x_2 is `gender`. The variable `age` takes $L_1 = 4$ values, that is "`0 to 16`", "`17 to 25`", "`26 to 55`" and "`56 and above`". The variable `gender` takes $L_2 = 2$ values: "`male`" and "`female`". The contingency table $\pi(x_1, \ldots, x_K)$ contains as many entries as the number of possible combinaisons of values for all the variables, that is

$$\prod_{i=1}^{K} L_i. \tag{10.9}$$

In our example, the number of entries is $L_1 L_2 = 4 \cdot 2 = 8$. Each entry contains the number of persons in the population characterized by the associated combination of variables.

For instance, the entry π("`26 to 55`","`female`") represents the number of females with age between 26 and 55. Table 10.4 represents the contingency table of an hypothetical population of size 100,000. In this example, π("`26 to 55`","`female`")=11799, which is the number of such individuals in the population.

Age	Gender	
	Male	Female
`0 to 16`	10,238	4,774
`17 to 25`	21,906	7,661
`26 to 55`	8,435	11,799
`56 and above`	9,266	25,921

Table 10.4: Example of a contingency table π for the population

$\widehat{\pi} =$

Age	Gender	
	Male	Female
`0 to 16`	30	10
`17 to 25`	20	1
`26 to 55`	1	10
`56 and above`	5	10

Table 10.5: Example of a contingency table for the sample

In the simple example described in Section 1.2, a contingency table with two discrete variables was considered, one with 3 values, and one with 2 values, for a total of 6 entries in the table (see Table 1.1).

As discussed above, the contingency table of the full population is rarely available. Therefore, it has to be estimated from available data. Two different sources of data are usually available: a sample of individuals from the population, and aggregate properties of the population. The sample of individuals is such that the value of each relevant variable is known for each individual. If the choice model has been estimated on revealed preference data, the estimation sample can be used.

The sample (of size N_S) provides a first estimate of the contingency table, denoted by $\widehat{\pi}$. For example, assume that we have interviewed a sample of size 87 from the population, and obtained the contingency table $\widehat{\pi}$ presented in Table 10.5.

Aggregate properties of a population usually provide the marginals of the contingency table for the population. In our example,

- the total number of people of age 0 to 16: 15,012;
- the total number of people of age 17 to 25: 29,567;
- the total number of people of age 26 to 55: 20,234;

	Gender			
Age	Male	Female	Total	Target
`0 to 16`	30	10	40	15012
`17 to 25`	20	1	21	29567
`26 to 55`	1	10	11	20234
`56 and above`	5	10	15	35187
Total	56	31		
Target	49845	50155		

Table 10.6: IPF example: Aggregate values for the sample and the population

- the total number of people of age 56 and above: 35,187;
- the total number of males: 49,845,
- the total number of females: 50,155.

Clearly, in order for the problem to be well defined, the marginals in each dimension should sum up to the same number, that is the population size.

Iterative Proportional Fitting

Generating a synthetic population consists in modifying the sample contingency table $\widehat{\pi}$ in such a way that it keeps its structure as much as possible, while featuring the aggregate properties of the real population. The idea of the method called *Iterative Proportional Fitting* is to consider each marginal one at a time, and to adjust the contingency table of the sample to match the aggregate property of the population. It is easier to illustrate the algorithm on the example. In Figure 10.6, we report the aggregate value for both the sample and the population together with the contingency table.

The contingency table is first adjusted so that the marginals for age in each category match the marginals of the population. To do so,

- all entries of the first row are multipled by 15012/40 = 375,
- all entries of the second row are multipled by 29567 / 21 = 1408,
- all entries of the third row are multipled by 20234 / 11 = 1839,
- all entries of the fourth row are multipled by 35187 / 15 = 2346.

The result of these operations are reported in Table 10.7.

	Gender			
Age	Male	Female	Total	Target
`0 to 16`	11259.0	3753.0	15012	15012
`17 to 25`	28159.0	1408.0	29567	29567
`26 to 55`	1839.5	18394.5	20234	20234
`56 and above`	11729.0	23458.0	35187	35187
Total	52987	47013		
Target	49845	50155		

Table 10.7: IPF example: row totals are matched

	Gender			
Age	Male	Female	Total	Target
`0 to 16`	10591.3	4003.8	14595	15012
`17 to 25`	26489.0	1502.0	27991	29567
`26 to 55`	1730.4	19623.7	21354	20234
`56 and above`	11033.4	25025.5	36058.9	35187
Total	49845	50155		
Target	49845	50155		

Table 10.8: IPF example: column totals are matched

The number of persons per age category in the contingency table now matches the number in the population. Then, the contingency table is adjusted so that the marginals for each gender match the marginals of the population. To do so,

- all entries of the first column are multiplied by $49845/52987 = 0.9407$,
- all entries of the second column are multiplied by $50155/47013 = 1.0668$.

The result of these operations are reported in Table 10.8. The number of persons per gender in the contingency table now matches the number in the population. But the number per age category does not match anymore. Therefore, the process is repeated iteratively until all marginals match. After having applied the process several times, the contingengy table 10.9 is obtained, which corresponds to the true population in this simple example.

The algorithm can exploit more than the marginals. For instance, we may know the number of males (32144) and females (12435) who are less than 25 in the population. This information may be used in the exact same way as the marginals. In Table 10.8, the number of males less than 25 is 30660, and the number of females less than 25 is 11212. Therefore, the two cells of the table

	Gender			
Age	Male	Female	Total	Target
0 to 16	10565.6	4446.3	15012	15012
17 to 25	27811.2	1755.5	29567	29567
26 to 55	1485.1	18748.7	20234	20234
56 and above	9982.2	25204.5	35187	35187
Total	49845	50155		
Target	49845	50155		

Table 10.9: IPF example: after convergence

	Gender			
Age	Male	Female	Total	Target
0 to 16	9181.3	9042.6	18224	15012
17 to 25	22962.7	3392.4	26355	29567
26 to 55	1730.4	19623.7	21354.1	20234
56 and above	11033.4	25025.5	36058.9	35187
Total	44908	57084		
Target	49845	50155		

Table 10.10: IPF example: number of males and females under 25 are matched

corresponding to males less than 25 are multiplied by 32144/30660, and the two cells of the table corresponding to females less than 25 are multiplied by 12435/11212, to obtain the results in Table 10.10. After convergence, Table 10.11 is obtained. Clearly, the marginals, as well as any additional aggregate property exploited by the algorithm must be consistent.

It is straightforward to generalize this algorithm to contingency tables with more than two dimensions. However, in higher dimensions, the number of cells grows exponentially with the number of variables, and may easily exceed the population size. For instance, 10 variables associated with 5 values each would correspond to a contingency table with about 10 million cells. As a consequence, many cells are associated with a low number of people in the population, and several of them with zero. This generates several problems.

First, the nature of the algorithm generates contingency tables with real numbers, not integers. While rounding up or down large numbers may not affect too much the overall nature of the population, rounding up or down many small numbers may have a significant impact on the table. Various procedures have been proposed in the literature to deal with this problem

	Gender			
Age	Male	Female	Total	Target
0 to 16	6954.6	8057.4	15012	15012
17 to 25	25189.4	4377.6	29567	29567
26 to 55	2635.2	17598.6	20234	20234
56 and above	15065.1	20121.6	35187	35187
Total	49845	50155		
Target	49845	50155		

Table 10.11: IPF example exploiting the number of males and females under 25: after convergence

(see Lovelace and Ballas, 2013 for a review).

Second, the presence of an empty cell in the initial sample may mean two things: either there is no person corresponding to this combination of variables in the population (it is usually called a "structural zero" cell), or there are only few such persons, and none of them has been sampled (it is then called a "sampling zero" cell). The multiplicative nature of the algorithm maintains any cell initialized at zero to that value. While it is the requested outcome for structural zero cells, it is clearly not desirable for the sampling zero cells.

Third, the algorithm may fail to converge if a cell is empty in the initial contingency table. The practical solution is to replace the initial zero by a small number (such as 0.000001, for instance).

Because of these limitations, it is more appropriate to use the Gibbs sampling algorithm described below.

Gibbs Sampling

Gibbs sampling is a simulation algorithm that is designed to draw from a multivariate probability function using the conditional distribution of each variable involved. We generate a synthetic population of size N_T characterized by discrete random variables $X_1, \ldots, X_K$, with joint probability mass function (pmf)

$$p(X_1, \ldots, X_K). \tag{10.10}$$

As discussed above, the pmf is unknown, but its marginals are available. A sequence of synthetic individuals is generated by simulation. Suppose that individual number n is associated with variables $(X_1 = x_1^n, \ldots, X_K = x_K^n)$. The next individual $n+1$ of the sequence is generated as follows:

- select randomly an index k between 1 and K, with equal probability

for each index;

- draw the value x_k^{n+1} of the variable X_k from the conditional distribution

$$\Pr(X_k | X_1 = x_1^n, \ldots, X_{k-1} = x_{k-1}^n, X_{k+1} = x_{k+1}^n, \ldots, X_K = x_K^n); \quad (10.11)$$

- replace $X_k = x_k^n$ by $X_k = x_k^{n+1}$ to obtain the next individual of the sequence.

The synthetic population is then created as follows.

Initialization The first individual is associated with any arbitrary valid values of the variables:

$$X^1 = (X_1 = x_1^1, \ldots, X_K = x_K^1).$$

This individual is not included in the synthetic population.

Warm up Generate a sequence of M_w individuals using the procedure described above (typically, $M_w = 1000K$), and do not include them in the population.

Populate Generate the next individual from the sequence and include it in the population.

Skip Generate a sequence of M_s individuals using the procedure described above (typically, $M_s = 100K$), and do not include them in the population.

Iterate Repeat the steps "Populate" and "Skip" until the generated population contains N_T individuals.

In technical terms, the sequence of individuals is called a Markov chain, and the Gibbs sampling algorithm belongs to the class of Markov Chain Monte Carlo methods. The step "warm up" is designed to bring the sequence to a stationary state, independent of the first individual. A correct implementation should check that the Markov chain has indeed reached stationarity (Ross, 2012). The step "skip" is designed to avoid that the individuals included in the population are artificially too similar.

Note that this method does not suffer from the issue of real numbers discussed in the context of IPF. Individuals are added one by one to the population, so that the number of individuals is an integer by construction. Also, the method does not require an initial sample. Therefore, the issue of structural versus sampling zeros is not relevant here. Still, if a sample of real

individuals is available, it can easily be included in the synthetic population. The simulation method can produce different synthetic populations, all consistent with the requested distribution. Finally, the method can be extended to generate individuals characterized by both discrete and continuous variables. We refer the reader to Farooq et al. (2013) for a detailed discussion of the method.

We illustrate the procedure using the simple example described above. We consider first the case where only marginals are available. In this case, (10.11) simplifies, as any conditional probability is equal to the marginal probability:

$$\begin{array}{lcllcl}
\Pr(\text{Male}|\text{age=0–16}) & = & \Pr(\text{Male}|\text{age=17–25}) & = & & \\
\Pr(\text{Male}|\text{age=26–55}) & = & \Pr(\text{Male}|\text{age=56+}) & = & 0.49845, \\
\Pr(\text{Female}|\text{age=0–16}) & = & \Pr(\text{Female}|\text{age=17–25}) & = & & \\
\Pr(\text{Female}|\text{age=26–55}) & = & \Pr(\text{Female}|\text{age=56+}) & = & 0.50155, \\
\Pr(\text{0–16}|\text{Gender=Male}) & = & \Pr(\text{0–16}|\text{Gender=Female}) & = & 0.15012, \\
\Pr(\text{17–25}|\text{Gender=Male}) & = & \Pr(\text{17–25}|\text{Gender=Female}) & = & 0.29567, \\
\Pr(\text{26–55}|\text{Gender=Male}) & = & \Pr(\text{26–55}|\text{Gender=Female}) & = & 0.20234, \\
\Pr(\text{56+}|\text{Gender=Male}) & = & \Pr(\text{56+}|\text{Gender=Female}) & = & 0.35187.
\end{array}$$

Starting the process with a male of age from 0 to 16, an example of (selected individuals from) a possible sequence is reported in Table 10.12. A $*$ indicator has been associated with individuals included in the synthetic population. It is observed from the first 10 individuals in the sequence that only one variable is modified at a time. After a warmup period generating 2000 individuals not included in the population, the first individual considered is #2001. Then, 200 individuals of the sequence are skipped until the next one is included in the population, and so forth. In this example, individual number k in the population corresponds to individual number $2001 + 200(k-1)$ in the sequence.

An example of a population generated by the algorithm is reported in Table 10.13. Note that the sums over the rows and over the columns do not match exactly with the sums of the real population. Indeed, the algorithm does not enforce that.

Clearly, the structure of the algorithm allows to exploit more than the marginals. For instance, if we know the number of males (32144) and females (12435) who are less than 25 in the population, we can derive more precise conditional probabilities. For gender, we have

$$\begin{array}{lcllcl}
\Pr(\text{Male}|\text{age=0–16}) & = & \Pr(\text{Male}|\text{age=17–25}) & = & 0.721057 \\
\Pr(\text{Male}|\text{age=26–55}) & = & \Pr(\text{Male}|\text{age=56+}) & = & 0.31939, \\
\Pr(\text{Female}|\text{age=0–16}) & = & \Pr(\text{Female}|\text{age=17–25}) & = & 0.278943 \\
\Pr(\text{Female}|\text{age=26–55}) & = & \Pr(\text{Female}|\text{age=56+}) & = & 0.68061.
\end{array}$$

0	[Male,0–16]			...	
1	[Female,0–16]		2401	[Female,17–25]	*
2	[Female,56+]			...	
3	[Female,25-55]		2601	[Male,17–25]	*
4	[Male,25-55]			...	
5	[Female,25-55]		2801	[Male,0–16]	*
6	[Male,25-55]			...	
7	[Male,25-55]		3001	[Female,25-55]	*
8	[Male,25-55]			...	
9	[Female,25-55]		3201	[Female,56+]	*
10	[Male,25-55]			...	
	...		3401	[Male,17–25]	*
2001	[Female,56+]	*		...	
2002	[Male,56+]		3601	[Female,25-55]	*
2003	[Female,56+]			...	
	...		3801	[Female,25-55]	*
2201	[Male,0–16]	*		...	
2202	[Male,0–16]		20001801	[Female,56+]	*

Table 10.12: Selected individuals from a Gibbs sampling sequence (* corresponds to those included in the synthetic population)

	Gender			
Age	Male	Female	Total	Target
`0 to 16`	7533	7599	15132	15012
`17 to 25`	14628	14711	29339	29567
`26 to 55`	10147	10254	20401	20234
`56 and above`	17403	17725	35126	35187
Total	49711	50289		
Target	49845	50155		

Table 10.13: Gibbs sampling example with marginals only

The probability Pr(Male|age=0–16) is obtained by the number of males under 25 divided by the total number of people under 25, that is $32144/44579 = 0.721057$. The probability Pr(Male|age=17–25) is computed in the same way. In order to derive the probability to be a male if the age is over 26, we use the following formula:

$$\Pr(\text{Male}) = \Pr(\text{Male}|\text{age=0–25})\Pr(\text{age=0–25}) + \Pr(\text{Male}|\text{age=26+})\Pr(\text{age=26+}),$$

which means

$$0.49845 = 0.721057 \times 0.44579 + \Pr(\text{Male}|\text{age=26+}) \times 0.55421.$$

Solving the above equation, we obtain

$$\Pr(\text{Male}|\text{age=26+}) = 0.31939.$$

For age categories, we have

$$\begin{array}{lcl}
\Pr(\text{age=0–16}|\text{Male}) & = & 0.21716 \\
\Pr(\text{age=17–25}|\text{Male}) & = & 0.42771 \\
\Pr(\text{age=26–55}|\text{Male}) & = & 0.12965 \\
\Pr(\text{age=56+}|\text{Male}) & = & 0.22547 \\
\Pr(\text{age=0–16}|\text{Female}) & = & 0.08349 \\
\Pr(\text{age=17–25}|\text{Female}) & = & 0.16444 \\
\Pr(\text{age=26–55}|\text{Female}) & = & 0.27457 \\
\Pr(\text{age=56+}|\text{Female}) & = & 0.47749
\end{array}$$

These probabilities are obtained from Bayes theorem. For instance,

$$\Pr(\text{age=0–16}|\text{Male}) = \Pr(\text{Male}|\text{age=0–16})\ \Pr(\text{age=0–16})\ /\ \Pr(\text{Male}).$$

Applying the Gibbs sampling algorithm with these conditional probabilities, we obtain a synthetic popluation characterized by the contingency table 10.14.

10.2.3 Sample enumeration

Sample enumeration uses a random sample of the population as "representative" of the entire population. In some circumstances, the same sample used for estimation can be used for aggregation and prediction, if the data consists of revealed preferences. Stated preferences data are not appropriate for prediction.

	Gender			
Age	Male	Female	Total	Target
0 to 16	11057	4069	15126	15012
17 to 25	21228	8335	29563	29567
26 to 55	6415	13762	20177	20234
56 and above	11209	23925	35134	35187
Total	49909	49932		
Target	50091	50155		
Total 0–25	32285	12404		
Target 0–25	32144	12435		

Table 10.14: Gibbs sampling example exploiting the number of males and females under 25

The predicted share of the sample choosing alternative i is used as an estimate for $W(i)$:

$$\widehat{W}(i) = \frac{1}{N} \sum_{n=1}^{N} P(i|x_n) \tag{10.12}$$

where N is the number of individuals in the sample.

Sample enumeration can also be used when the sample is drawn non-randomly from the population. In that case a weighted average will replace the simple average of (10.12). For example, suppose the sample is (endogenously or exogenously) stratified so that different groups within the population are sampled at different rates. In this case it is straightforward first to classify the population into the groups used for stratification of the sample, apply the sample enumeration method to each group, and then compute an estimate of $W(i)$ as the weighted sum of the within-class forecasts. Mathematically this corresponds to equation (10.13):

$$\widehat{W}(i) = \sum_{g=1}^{G} \left(\frac{N_{T_g}}{N_T}\right) \frac{1}{N_{sg}} \sum_{n=1}^{N_{sg}} P(i|x_n) \tag{10.13}$$

where N_{sg} is the size of the sample for the gth group, and N_{T_g} and N_T are as before.

The following aspects of sample enumeration are worth noting:

1. The predicted aggregate shares are estimates and are consequently subject to sampling error. When the choice probabilities or the sample are small, the sampling error may be a large fraction of $W(i)$.

2. The estimator of $W(i)$ is consistent as long as the parameter estimates used are consistent. (This can be shown straightforwardly from the Slutsky theorem. The variance of the sample enumeration forecast will vary inversely with N_s, the sample size.)

3. It is easy to produce forecasts for different socioeconomic groups. All one need do is keep separate tabulations of the choice probabilities for each of the socioeconomic groups of interest. If, however, the groups are only a small fraction of the original N_s, individuals in the sample, the variance of the subpopulation forecasts may be quite high.

4. It is well suited for forecasting the effects of policies that differentially impact various population groups. To forecast the changes in aggregate shares under some policy, one simply changes the values of the appropriate independent variables for each affected individual in the sample. For example, suppose we wanted to test the effects of rebating the transit fares of low-income travelers using different sliding scales, each based on gross household income. For each formula we would appropriately change the transit fare for each traveler and perform a sample enumeration forecast. This type of procedure is straightforward to program on a computer because it essentially involves execution of the same steps for each observation in the sample.

10.2.4 Microsimulation

In some instances the complete choice set is very large because either the number of original alternatives is inherently great (as often occurs in destination choice models) or, when forecasts are made for successive time intervals, the choice in any period can depend on some or all prior choices (as might be the case for residential location decisions). In the case of time-dependent choices, if there are J choices and τ time intervals, there are J^τ possible choice "paths" over time.

In both these situations we can often adopt some form of Monte Carlo simulation to reduce the computational burden of producing a forecast. Two common applications are as follows:

1. Random sampling from the choice set. This approach applies only when we use logit. Because of the IIA property, if we first randomly sample a subset of each individual's possible choices and then use sample enumeration to estimate $W(i)$, the resulting estimator is consistent.

(As we will discuss further in section 12.3, this procedure and numerous variants can also be used in estimating the parameters of a logit model.)

2. Simulation of outcomes. When an individual faces a sequence of τ decisions, we are often not particularly interested in forecasting the fraction of the population selecting each of the possible J^τ choice paths. Rather, we are more concerned with estimating the fraction of the population making each choice at each time, resulting in J^τ forecasts. One way of exploiting this result is actually to simulate the choice process for each individual in the sample.

This latter procedure can be illustrated by means of a binary choice example. Consider an individual n at time 1. Let $P_n(i)$ and $P_n(j)$ denote the predicted probability that n chooses alternatives i and j, respectively. We then draw a random variable u_1, uniformly distributed between 0 and 1 and "assign" the individual to alternative i at time 1 if $P_n^1(i) < u_1$, or to alternative j otherwise. At time period 2 we forecast $P_n^2(i)$ and $P_n^2(j)$ *conditional on the assigned choice for time period* 1. We again draw a uniformly distributed random variable and assign the individual to i or j, as appropriate. This procedure continues through all τ time periods. In doing this for all N_s individuals, we keep a count of $N_s^t(i)$, the number in the sample "assigned" to i at time t. We then use

$$\widehat{W}^t(i) = \frac{N_s^t(i)}{N_s} \tag{10.14}$$

as an estimate of the share of the population choosing i at time t.

Note that using Monte Carlo in a forecasting procedure retains the property of consistently estimating $W(i)$, it generally increases the variance of the forecast.

10.3 Calibration of the constants

As discussed above, a disaggregate choice model is mainly used to compute aggregate quantities in order to perform policy analysis. When aggregate data is available, it is recommended to recalibrate some parameters of the model at the aggregate level. Typically, the alternative specific constants are calibrated against observed market shares. Let's denote $c_1, \ldots, c_{J-1}$ the alternative specific constants of the choice model $P(i|x_n; c_1, \ldots, c_{J-1})$. For each $i \in \mathcal{C}$, the aggregate market shares predicted by the model are given by

(10.3), that is

$$W(i; c_1, \ldots, c_{J-1}) = \frac{1}{N_T} \sum_{n=1}^{N_T} P(i|x_n; c_1, \ldots, c_{J-1}). \tag{10.15}$$

If $\overline{W}_i$ denotes the observed market share of alternative i, it is rarely the case that the predicted shares match the observed shares when the constants have been estimated from disaggregate data. Indeed, various sources of errors add up, including sampling errors and aggregation errors. Before using the model to forecast market shares, the constants must be calibrated by solving the following system of equations

$$\begin{aligned} W(1; c_1, \ldots, c_{J-1}) &= \overline{W}_1 \\ W(2; c_1, \ldots, c_{J-1}) &= \overline{W}_2 \\ &\vdots \\ W(J-1; c_1, \ldots, c_{J-1}) &= \overline{W}_{J-1}. \end{aligned} \tag{10.16}$$

It is a system of $J-1$ equations with $J-1$ unknowns (the constants). It can be solved with numerical methods such as Newton's algorithm (Dennis and Schnabel, 1996). Train (2003) suggests a heuristic method to adjust the constants empirically:

$$c_i^+ = c_i + \ln(\overline{W}_i) - \ln(W(i; c_1, \ldots, c_{J-1})). \tag{10.17}$$

If the predicted market shares $W(i; c_1^+, \ldots, c_{J-1}^+)$ still do not match the observed ones, start the process again until they do.

10.4 Including a new alternative

It is commonly desirable to forecast the demand in a context where a new alternative is introduced. How will the market react to the introduction of a new product? What will be the ridership of a new transportation service? In such a case, it is recommended to complement the revealed preferences data by collecting stated preferences data to assess the responses of the decision makers to the new alternative. We refer the reader to Section 2.2 for a discussion about the collection of stated preference data, and to Chapter 15 for a discussion about how to combine both types of data to estimate choice models.

If a joint model has been estimated using both revealed and stated preference data, the utility functions of the forecasting model are the utility functions from the revealed preference model for the existing alternative and

the utility function from the stated preference models for the new alternatives. However, note that the calibration of the alternative specific constant of a new alternative cannot be done objectively. Therefore, it may be considered as part of the forecasting scenario. As discussed in Section 10.3, making assumptions about the constant c^R_{new} is equivalent to making assumptions about the market shares for the base case scenario.

Several researchers (such has Daly and Rohr, 1998, Cherchi and Ortúzar, 2006 and Glerum et al., forthcoming) have proposed to exploit the stated preferences data to make these assumptions. The general idea is to obtain an estimate of the relative market shares of the new alternative compared to an existing alternative from the model U^S. It first requires that the individuals in the SP dataset are representative of the population, so that weighting the SP observations is likely to be necessary. As discussed in Section 2.2.4, the values of the variables in the SP questionnaire are generated by an experimental design, with the objective to increase the precision of the parameter estimates. They do not necessarily correspond to configurations that will be implemented in practice. The responses to unrealistic choice situations do not reveal anything about the market shares. They should be discarded from the sample for this analysis. Judgment must be used to decide what is realistic, and what is not.

Once the value of the constant has been assumed, it is useful to interpret it in terms of trade-offs. Indeed, the difference $c^R_{new} - c^R_i$ between the constant of the new alternative and an existing alternative i can be converted into interpretable units, such as currency or time, by dividing it by the corresponding coefficient (in a linear-in-parameters specification). For instance, the value

$$\frac{c^R_{new} - c^R_i}{\beta_{cost}} \tag{10.18}$$

is a price equivalent of the average preference toward the new alternative. In transportation demand analysis, it is also convenient to consider

$$\frac{c^R_{new} - c^R_i}{\beta_{time}} \tag{10.19}$$

which translates into minutes that preference. The realism of these quantities is easier to assess than the value of the constant itself.

When no stated preference data is available, the utility function of the new alternative must be borrowed from the existing alternative which is the most similar to the new one. For instance, assume that a transportation mode choice model for long distance travel has been estimated, with the following existing alternatives: car, train and plane. In order to analyze the

impact of a new high-speed train in this market, the corresponding alternative must be included in the model. In this example, it makes sense to borrow the specification of the utility function (including the estimates of the parameters) from the train alternatives. The values of the explanatory variables will be determined by the forecasting scenarios.

The main issue with this approach is the likely correlation between the error term of the new alternative and the error term of the borrowed existing alternative. If the estimated model is logit, a logit specification including the new alternative may not be valid. In this case, the two alternatives should be included in a nest, and a nested logit specification should be used for forecasting. As there is no way to estimate the nest parameter for such a specification, the future performance indicators should be derived for different assumptions about the nest parameters, from 1 (corresponding to the logit model with independent error terms) to $+\infty$ (corresponding to perfectly correlated error terms). In practice, the indicators are derived for the two extreme values (1 and $+\infty$) and an interval is reported.

Once the specification of the forecasting model is finalized, aggregation methods described in Section 10.2 can be applied. Remember that the values of the variables used for forecasting cannot be borrowed from the stated preference data set. Indeed, these values have been engineered in order to obtain efficient estimate of the coefficients. They do not reflect any realistic scenario of the future. Such scenarios have to be created, and the values of the variables have to be assumed accordingly. In the presence of new alternatives, the sample enumeration approach is less appropriate, as it is impossible to sample data from a non-existing reality.

10.5 Indicators for policy analysis

In addition to the market shares, various indicators can be dervied from choice models that are useful for policy analysis. We present here the most used in practice. Elasticities and willingness to pay provide information about the role of one or two specific variables on the outcome of the model. The incremental logit, the consumer surplus and the revenue calculator provide more aggregate information about future scenarios.

10.5.1 Point Elasticities

One useful property of demand models is the concept of an elasticity. Elasticity is the ratio of the percent change in one variable to the percent change in another variable. In our case, we are interested in an indicator reporting

the effect on the choice probability, or on the aggregate shares, of the change of one of the policy variables in the model. Depending on the type of change in the policy variable that is considered, we refer to *point* or *arc* elasticity. The *point* elasticity captures the impact of an infinitesimal modification of the (continuous) policy variable. It is defined as the incremental change of the logarithm of the dependent variable with respect to an incremental change of the logarithm of the independent variable. The *arc* elasticity captures the impact of a modification of one variable between two values, and is discussed in the next section. The *direct* elasticity of demand measures the responsiveness of the quantity demanded of an alternative to a change in an attribute of the same alternative. The *cross* elasticity of demand measures the responsiveness of the quantity demanded of an alternative to a change in an attribute of another alternative.

In defining elasticities for discrete choice models, we must distinguish between disaggregate and aggregate elasticities.

A disaggregate elasticity represents the responsiveness of an individual's choice probability to a change in the value of some continuous attribute. The simplest case is the elasticity $E_{x_{ink}}^{P_n(i)}$ of the probability of an individual n choosing alternative i with respect to a change in attribute k, that is

$$E_{x_{ink}}^{P_n(i)} = \frac{\partial P_n(i)}{\partial x_{ink}} \frac{x_{ink}}{P_n(i)} = \frac{\partial \ln P_n(i)}{\partial \ln x_{ink}}. \tag{10.20}$$

For example, consider the linear-in-parameters logit model:

$$P_n(i) = \frac{e^{V_{in}}}{\sum_j e^{V_{jn}}} = m \frac{e^{\sum_k \beta_k x_{ink}}}{\sum_j e^{\sum_k \beta_k x_{jnk}}}. \tag{10.21}$$

The derivatives with respect to the utility functions are

$$\frac{\partial P_n(i)}{\partial V_{in}} = P_n(i)(1 - P_n(i)), \tag{10.22}$$

and

$$\frac{\partial P_n(i)}{\partial V_{jn}} = -P_n(i)P_n(j). \tag{10.23}$$

Note that the logit model has the property that the cross derivatives are symmetric, that is

$$\frac{\partial P_n(i)}{\partial V_{jn}} = \frac{\partial P_n(j)}{\partial V_{in}}. \tag{10.24}$$

The derivatives with respect to the variables are

$$\frac{\partial P_n(i)}{\partial x_{ink}} = \frac{\partial P_n(i)}{\partial V_{in}} \frac{\partial V_{in}}{\partial x_{ink}} = \beta_k P_n(i)(1 - P_n(i)), \tag{10.25}$$

and

$$\frac{\partial P_n(i)}{\partial x_{jnk}} = \frac{\partial P_n(i)}{\partial V_{jn}} \frac{\partial V_{jn}}{\partial x_{jnk}} = -\beta_k P_n(i) P_n(j). \tag{10.26}$$

Therefore, the disaggregate *direct elasticity* of logit with linear-in-parameters utility functions is given by

$$E_{x_{ink}}^{P_n(i)} = (1 - P_n(i)) x_{ink} \beta_k. \tag{10.27}$$

Similarly the *disaggregate cross elasticity* of the probability that alternative i is selected with respect to an attribute of alternative j is

$$E_{x_{jnk}}^{P_n(i)} = \frac{\partial P_n(i)}{\partial x_{jnk}} \frac{x_{jnk}}{P_n(i)}, \text{ for } j \neq i. \tag{10.28}$$

For logit with linear-in-parameters utility functions, we obtain

$$E_{x_{jnk}}^{P_n(i)} = -P_n(j) x_{jnk} \beta_k, \text{ for } j \neq i. \tag{10.29}$$

One property of logit is that it has uniform cross elasticities — that is, the cross elasticities of all alternatives with respect to a change in an attribute affecting only the utility of alternative j are equal for all alternatives $i \neq j$. It is a direct consequence of (10.24). This aspect of the logit model is another manifestation of the IIA property, discussed in Section 5.3.1.

There are many cases where the formulas defined in equations (10.27) and (10.29) are not appropriate. One relatively simple instance is where the independent variable is a function of some attribute of interest, z_{ink}, such as $x_{ink} = h^k(z_{ink})$. In this case

$$E_{z_{ink}}^{P_n(i)} = (1 - P_n(i)) \beta_k \frac{\partial h^k}{\partial z_{ink}} z_{ink}. \tag{10.30}$$

If, for example, $x_{ink} = \ln(\text{travel time}_{in})$, then

$$E_{\text{travel time}_{in}}^{P_n(i)} = (1 - P_n(i)) \beta_k. \tag{10.31}$$

Similarly the cross elasticity for an attribute that enters into the utility of an alternative in transformed form is

$$E_{z_{jnk}}^{P_n(i)} = -P_n(j) \beta_k \frac{\partial h^k}{\partial z_{jnk}} z_{jnk}, \text{ for } j \neq i.$$

A more complicated case occurs when one of the attributes in the model is an alternative-specific decision-maker's characteristics that enters into the systematic utility of more than one alternative. In this case the effect of

an incremental change in the socioeconomic variable does not only shift one utility, it shifts all of the utilities in which that variable appears. Similarly, if some variable such as travel time appears in more than one independent variable (e.g., time and time squared), then an incremental change in travel time influences more than one independent variable.

One must be extremely careful when applying elasticity formulas to account fully for such complications. Although the functional forms of the elasticity for any particular situation is problem specific, it is always possible to apply the definitions (10.20) and (10.28) of a point elasticity directly, taking care to ensure that the derivatives of the choice probabilities with respect to the variable of interest appropriately account for how that variable enters all the utility functions.

Aggregate elasticities summarize the responsiveness of some group of decision makers rather than that of any individual. For example, suppose that we wish to know the effect of an incremental change in a variable on the *expected share* of the group choosing alternative i. To solve for this, we consider $W(i)$, the expected share of the group choosing alternative i, defined by (10.3). Suppose we now alter the value of some variable x_{jnk} for each individual by some increment so that

$$\frac{\partial x_{jnk}}{x_{jnk}} = \frac{\partial x_{jn'k}}{x_{jn'k}} = \frac{\partial x_{jk}}{x_{jk}}, \text{ for all } n, n' = 1, 2, \ldots, N_T, \tag{10.32}$$

where

$$x_{jk} = \frac{1}{N_T} \sum_{n=1}^{N_T} x_{jnk}.$$

(This corresponds to a uniform percentage change in x_{jnk} across all members of the group.) For this type of change the aggregate elasticity is as follows:

$$E_{x_{jk}}^{W(i)} = \frac{\partial W(i)}{\partial x_{jk}} \frac{x_{jk}}{W(i)}. \tag{10.33}$$

Using the definition (10.3) of $W(i)$, and multiplying every term by $P_n(i)/P_n(i)$, we obtain

$$E_{x_{jk}}^{W(i)} = \sum_{n=1}^{N_T} \frac{P_n(i)}{P_n(i)} \frac{\partial P_n(i)}{\partial x_{jk}} \frac{x_{jk}}{\sum_{n=1}^{N_T} P_n(i)}. \tag{10.34}$$

Now the aggregate elasticity can be expressed as follows:

$$E_{x_{jk}}^{W(i)} = \frac{\sum_{n=1}^{N_T} P_n(i) E_{x_{jnk}}^{P_n(i)}}{\sum_{n=1}^{N_T} P_n(i)}, \tag{10.35}$$

which is simply a weighted average of the individual level elasticities using the choice probabilities as weights.

Note that if the market shares is approximated by sample enumeration as described in Section 10.2.3, the aggregate elasticity can be derived from (10.13) to obtain

$$E_{x_{jk}}^{\widehat{W}(i)} = \frac{1}{\widehat{W}(i)} \sum_{g=1}^{G} \left(\frac{N_{T_g}}{N_T}\right) \frac{1}{N_{sg}} \sum_{n=1}^{N_{sg}} E_{x_{ink}}^{P_n(i)} P(i|x_n). \tag{10.36}$$

The disaggregate elasticities of equations (10.27) and (10.29) can be written as

$$E_{x_{jnk}}^{P_n(i)} = (\delta_{ij} - P_n(i)) x_{jnk} \beta_k,$$

where δ_{ij} is the Kronecker delta function, which equals 1 for $i = j$ and 0 for $i \neq j$. Substituting this expression in (10.35), we obtain

$$E_{x_{jk}}^{W(i)} = \frac{\beta_k}{N_T W(i)} \sum_{n=1}^{N_T} P_n(i)(\delta_{ij} - P_n(j)) x_{jnk}.$$

Note that for a cross elasticity we set $\delta_{ij} = 0$ and obtain an expression that depends on both i and j. This demonstrates once again that the uniform disaggregate cross elasticities that result from the IIA property need not hold at the aggregate level.

10.5.2 Arc elasticities

Using point elasticities is one way to predict changes due to modifications in the independent variables. In practice, policy variables are updated from one value to another, so that the use of arc elasticities is often more appropriate. Let x_{ink} be the current value of one variable, and $x_{ink}^+ = x_{ink} + \Delta x_{ink}$ the future value. Keeping all other variables at their current values, we denote $P_n(i)$ the current choice probability of alternative i, and $P_n^+(i) = P_n(i) + \Delta P_n(i)$ the choice probability involving x_{ink}^+. The disaggregate *arc elasticity* is defined as the ratio of the relative change of the choice probability and the relative change of the variable, that is

$$E_{\Delta x_{ink}}^{\Delta P_n(i)} = \frac{\Delta P_n(i)}{P_n(i)} \frac{x_{ink}}{\Delta x_{ink}}. \tag{10.37}$$

A formula where the relative change is computed based on the midpoint of the interval has also been proposed:

$$E_{\Delta x_{ink}}^{\Delta P_n(i)} = \frac{\Delta P_n(i)}{P_n(i) + \Delta P_n(i)/2} \frac{x_{ink} + \Delta x_{ink}/2}{\Delta x_{ink}} = \frac{\Delta P_n(i)}{P_n(i) + P_n^+(i)} \frac{x_{ink} + x_{ink}^+}{\Delta x_{ink}}. \tag{10.38}$$

Cross arc elasticities are defined similarly. The aggregate arc elasticity is the same concept, applied to the expected share $W(i)$, that is

$$E_{\Delta x_j}^{\Delta W(i)} = \frac{\Delta W(i)}{W(i)} \frac{x_{jk}}{\Delta x_{jk}}, \tag{10.39}$$

or

$$E_{\Delta x_{jk}}^{\Delta W(i)} = \frac{\Delta W(i)}{W(i) + \Delta W(i)/2} \frac{x_{jk} + \Delta x_{jk}/2}{\Delta x_{jk}}. \tag{10.40}$$

Note that the aggregate arc elasticity is not a weighted sum of the disaggregate arc elasticities.

10.5.3 Incremental logit

For the linear-in-parameters logit model there is a convenient form known as the *incremental logit* which can be used to predict changes in behavior on the basis of the *existing choice probabilities of the alternatives* and *changes in variables* that obviates the need to use the full set of independent variables to calculate the new choice probabilities.

Derivation of the incremental form of the logit model is relatively straightforward. The linear-in-parameters logit model predicts the probability that individual n chooses alternative i from the set of alternatives $\mathcal{C}_n$, as given by equation (5.21). The revised choice probability resulting from a change in utilities is given by

$$P_n'(i) = \frac{e^{V_{in}+\Delta V_{in}}}{\sum_{j\in\mathcal{C}_n} e^{V_{jn}+\Delta V_{jn}}}$$

where

ΔV_{in} = the change in utility for alternative i

$$= \sum_{k=1}^{K} \beta_k \Delta x_{ink},$$

Δx_{ink} = the change in the kth independent variable for alternative i and individual n.

Divide both the numerator and the denominator by $\sum_{j\in\mathcal{C}_n} e^{V_{jn}}$ to obtain the incremental logit model:

$$P_n'(i) = \frac{P_n(i)e^{\Delta V_{in}}}{\sum_{j\in\mathcal{C}_n} P_n(j)e^{\Delta V_{jn}}}.$$

Thus to predict changes with a linear-in-parameters logit choice model, we need to know the choice probabilities in the base case and the changes in utilities due only to the affected variables. It is not necessary to recalculate the full utilities.

10.5.4 Consumer surplus

The concept of consumer surplus has been derived from microeconomic consumer theory in the context of continuous goods in Section 3.5. In the case of discrete choice, the definition is the same: it is the difference between what a consumer is willing to pay for a good and what she actually pays for the good. It is equal to the area under the demand curve and above the market price. In discrete choice, the demand for an individual is characterized by the choice probability. Also, the role of price is taken by the utility of the good. To illustrate these differences, Figure 3.3 illustrating the consumer surplus has been updated for the specific case of a binary logit model in Figure 10.2. In Figure 3.3, the x-axis represents the quantity of the good. In the context of discrete choice, it corresponds to the number of times that the good is chosen by the consumer. The normalized version of the quantity is the choice probability, represented on the x-axis of Figure 10.2. The y-axis of Figure 3.3 is the price of the good. In the context of discrete choice, there are more variables explaining the behavior. Therefore, we prefer to use the indirect utility. Note that utility can be transformed into monetary units in various ways, such as by dividing this measure by a cost coefficient. In Figure 10.2, the y-axis represents the negative utility of alternative i, $-V_i$. The minus sign helps in obtaining the same interpretation as in Figure 3.3: going up the axis corresponds to a deterioration (increase of price in one case, decrease in utility in the other case). Simulating a future increase of the utility of item i (that is, a decrease of the quantity $-V_i$), while the utility of j is constant, the change in consumer surplus is represented by the hatched area. This area can be calculated by the following integral

$$\int_{V_i^1}^{V_i^2} P(i|V_i, V_j)\,dV_i = \int_{V_i^1}^{V_i^2} \frac{e^{\mu V_i}}{e^{\mu V_i} + e^{\mu V_j}}\,dV_i \tag{10.41}$$

which is

$$\frac{1}{\mu}\ln(e^{\mu V_i^2} + e^{\mu V_j}) - \frac{1}{\mu}\ln(e^{\mu V_i^1} + e^{\mu V_j}). \tag{10.42}$$

To generalize this result, we calculate the difference in an individual's consumer surplus between two situations corresponding to vectors of systematic

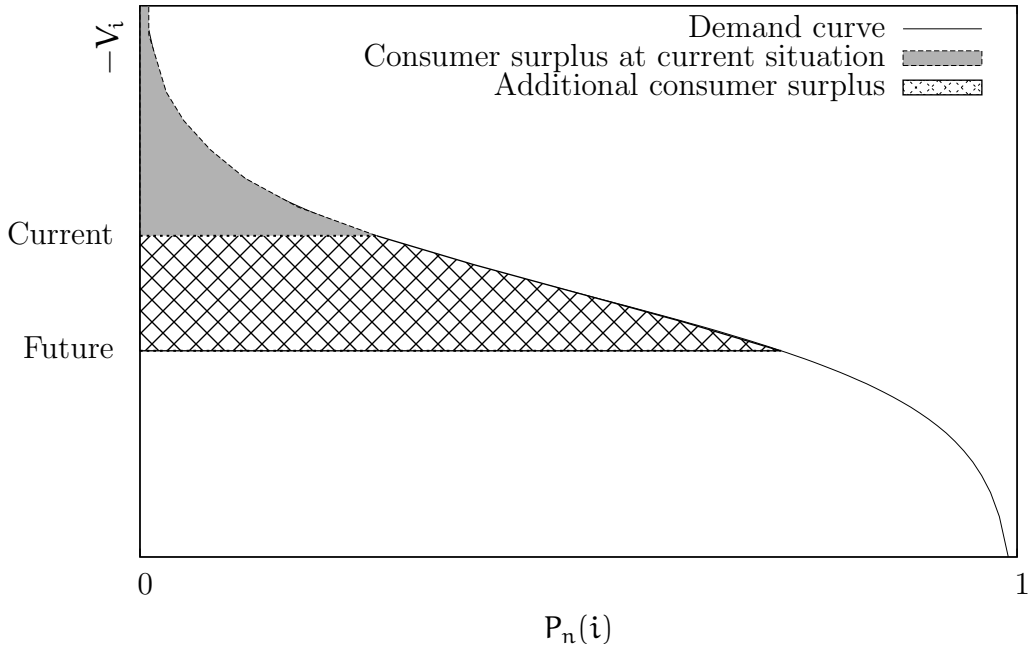


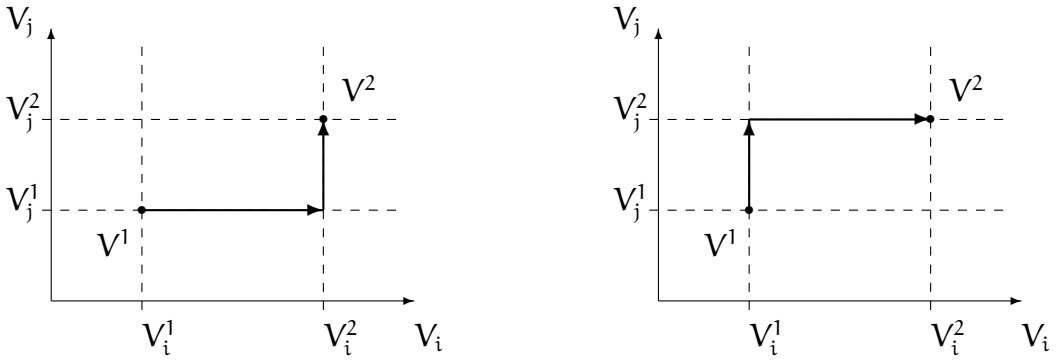
Figure 10.2: Illustration of the consumer surplus for binary logit

utilities V^1 and V^2 as follows:

$$\sum_{i \in \mathcal{C}} \int_{V^1}^{V^2} P(i|V) dV_i, \tag{10.43}$$

where the choice probability is denoted as conditional on the vector V of systematic utilities in order to make the dependency explicit. Note that the term $P(i|V)dV_i$ corresponds to the hatched area in Figure 10.2 where the change in the utility of i is infinitesimal. The difficulty here is that the utility of **all** alternatives are modified. Therefore, the integral in (10.43) is a line integral. And there are infinitely many ways to move from utility vector V^1 to utility vector V^2 in the J-dimensional space. This is illustrated for an example with two alternatives in Figure 10.3. In order to simplify the calculation of the integral, we consider paths that are updating each coordinate at a time. In Figure 10.3(a), the path moves first from $V^1 = (V_i^1, V_j^1)$ to (V_i^2, V_j^1), and then from (V_i^2, V_j^1) to $(V_i^2, V_j^2) = V^2$. The path in Figure 10.3(b) moves first from $V^1 = (V_i^1, V_j^1)$ to (V_i^1, V_j^2), and then from (V_i^1, V_j^2) to $(V_i^2, V_j^2) = V^2$.

For the example with two alternatives, where the path in Figure 10.3(a)



(a) First path (b) Second path

Figure 10.3: Moving from utility vector V_1 to utility vector V_2

is followed, (10.43) writes

$$\int_{V_i^1}^{V_i^2} P(i|V_i, V_j^1)dV_i + \int_{V_j^1}^{V_j^2} P(j|V_i^2, V_j)dV_j. \tag{10.44}$$

Assuming now a binary logit model, the first integral is

$$\int_{V_i^1}^{V_i^2} \frac{e^{\mu V_i}}{e^{\mu V_i} + e^{\mu V_j^1}} dV_i. \tag{10.45}$$

Let $t = e^{\mu V_i} + e^{\mu V_j^1}$ so that $dt = \mu e^{\mu V_i} dV_i$, we obtain

$$\frac{1}{\mu} \int_{e^{\mu V_i^1} + e^{\mu V_j^1}}^{e^{\mu V_i^2} + e^{\mu V_j^1}} \frac{dt}{t} = \frac{1}{\mu} \ln(e^{\mu V_i^2} + e^{\mu V_j^1}) - \frac{1}{\mu} \ln(e^{\mu V_i^1} + e^{\mu V_j^1}). \tag{10.46}$$

The second integral is

$$\int_{V_j^1}^{V_j^2} \frac{e^{\mu V_j}}{e^{\mu V_i^2} + e^{\mu V_j}} dV_j. \tag{10.47}$$

Let $t = e^{\mu V_i^2} + e^{\mu V_j}$ so that $dt = \mu e^{\mu V_j} dV_j$, we obtain

$$\frac{1}{\mu} \int_{e^{\mu V_i^2} + e^{\mu V_j^1}}^{e^{\mu V_i^2} + e^{\mu V_j^2}} \frac{dt}{t} = \frac{1}{\mu} \ln(e^{\mu V_i^2} + e^{\mu V_j^2}) - \ln(e^{\mu V_i^2} + e^{\mu V_j^1}). \tag{10.48}$$

Adding (10.46) and (10.48), we obtain the difference of logsum, that is

$$\frac{1}{\mu} \ln(e^{\mu V_i^2} + e^{\mu V_j^2}) - \frac{1}{\mu} \ln(e^{\mu V_i^1} + e^{\mu V_j^1}). \tag{10.49}$$

Calculating the integral following the path described in Figure 10.3(b) leads to the exact same result. We say that the calculation of the integral is *path independent*. Not all line integrals are path independent. If the choice model happens to have equal cross-derivatives, that is

$$\frac{\partial P(i|V,\mathcal{C})}{\partial V_j} = \frac{\partial P(j|V,\mathcal{C})}{\partial V_i}, \ \forall i,j \in \mathcal{C}, \tag{10.50}$$

the calculation of the integral in (10.43) is path independent.

As discussed in Section 10.5.1, the logit model has this property (see (10.24)). Therefore, for logit, we can select the path of integration. We calculate (10.43) using a path that first updates the utility of alternative 1, then of alternative 2, and so on until J. The k^{th} term integrates over V_k, with all utilities of alternatives 1 to $k-1$ set at the level V^2, while all utilities of alternatives $k+1$ to J are set at the level V^1. This term writes

$$\int_{V_k^1}^{V_k^2} \frac{e^{\mu V_k}}{\sum_{j=1}^{k-1} e^{\mu V_j^2} + e^{\mu V_k} + \sum_{j=k+1}^{J} e^{\mu V_j^1}} dV_k =$$
$$\frac{1}{\mu} \ln \left(\sum_{j=1}^{k-1} e^{\mu V_j^2} + e^{\mu V_k^2} + \sum_{j=k+1}^{J} e^{\mu V_j^1} \right) - \frac{1}{\mu} \ln \left(\sum_{j=1}^{k-1} e^{\mu V_j^2} + e^{\mu V_k^1} + \sum_{j=k+1}^{J} e^{\mu V_j^1} \right). \tag{10.51}$$

When summing up over k, most terms of the sum over alternatives cancel out, and the difference of consumer surplus for the logit model is

$$\frac{1}{\mu} \ln \sum_{j \in \mathcal{C}} e^{\mu V_j^2} - \frac{1}{\mu} \ln \sum_{j \in \mathcal{C}} e^{\mu V_j^1}. \tag{10.52}$$

which is the difference among expected maximum utilities in the two situations (see equation (5.31)). When the choice set changes from $\mathcal{C}^1$ to $\mathcal{C}^2$, the result of (10.43) is

$$\frac{1}{\mu} \ln \sum_{j \in \mathcal{C}^2} e^{\mu V_j^2} - \frac{1}{\mu} \ln \sum_{j \in \mathcal{C}^1} e^{\mu V_j^1}. \tag{10.53}$$

Note that the difference in consumer surplus can be calculated for choice models that are translationally invariant (see Section 9.1). For these models, property (10.50) is directly derived from (3.82), or (9.1), using the choice probability generating function (CPGF), and the integral (10.43) is also path independent. We obtain that the diffence in consumer surplus is

$$\ln G(e^{V^2}) - \ln G(e^{V^1}), \tag{10.54}$$

where G is the CPGF of the choice model. For instance, using the CPGF (8.39) of the nested logit model, we obtain the difference of consumer surplus for this model as

$$\frac{1}{\mu}\ln\sum_{m=1}^{M}\left(\sum_{\ell=1}^{J_m} e^{V_\ell^2 \mu_m}\right)^{\mu/\mu_m} - \frac{1}{\mu}\ln\sum_{m=1}^{M}\left(\sum_{\ell=1}^{J_m} e^{V_\ell^1 \mu_m}\right)^{\mu/\mu_m}. \tag{10.55}$$

We refer the reader to Neuburger (1971), Small and Rosen (1981), Hanemann (1984), McConnell (1995), Dagsvik and Karlström (2005) for more detailed discussions about consumer surplus.

10.5.5 Willingness to pay and willingness to accept

The *willingness to pay* (WTP) is a useful quantity that captures the trade off between the modification of a variable and money. Formally, it is defined as the net decrease in final income that makes the expected utility equal before and after the modification of the variable. Note that it is implicit to this definition that the modification of the variable **improves** the expected utility.

As discussed in Section 3.6, in the context of discrete choice, we must consider the income remaining, which is the income that is left to spend on the continuous goods after the discrete good i is purchased. We denote it $I_n - c_{in}$, where I_n is the income of individual n and c_{in} the cost of alternative i for individual n. When this variable appears linearly in the utility function of each alternative, that is

$$V_{in} = \beta_c(I_n - c_{in}) + \cdots = \beta_c I_n - \beta_c c_{in} + \cdots, \tag{10.56}$$

the income variable cancels out in the calculation of the choice probability, and can be removed from the model (as it was done for most examples presented in this book). We keep it here to derive the willingness to pay from its definition, which is based on income.

In transportation demand analysis, a typical willingness to pay indicator is the *value of travel time savings* (VTTS) or *value of time* (VOT), which is the price that travelers are willing to pay to save travel time. The VTTS is expressed in currency units per unit of time, like dollars per minute or euros per hour. In other fields such as environmental or health economics, the aim is to give a money value to non market resources, such as the quality of air, of the eradication of a disease. In these contexts, this is usually referred to as *contingent valuation*. The role of WTP is particularly important in cost-benefit analysis. Indeed, the decision to invest in a project that decreases the

travel time between two cities, improves the quality of air in the city center, or decreases the spread of cholera after a major catastrophic event, can be motivated by the comparison between the actual cost of the project, and its benefit, expressed in currency units using the willingness to pay indicator.

Let I_n be the income of individual n, c_{in} be the cost of alternative i for individual n and x_{ink} be the value of another variable of the model (such as the number of cubic meters of clean water in a lake.) Let $V_{in}(I_n - c_{in}, \ldots, x_{ink}, \ldots)$ be the value of the utility function of alternative i. We denote by

$$EM(I_n - c_{in}, x_{ink}) = E[\max_{j \in \mathcal{C}_n}(V_1, \ldots, V_{in}(I_n - c_{in}, \ldots, x_{ink}, \ldots), \ldots, V_{J_n})] \tag{10.57}$$

the expected maximum utility experienced by individual n for the current situation. Consider a scenario where the variable under interest increases to $x_{ink} + \Delta x$. We denote by ΔI the decrease in final income that would achieve the same expected maximum utility, that is the value of ΔI that solves the equation

$$EM(I_n - c_{in}, x_{ink}) = EM(I_n - \Delta I - c_{in}, x_{ink} + \Delta x). \tag{10.58}$$

Using the implicit function theorem, we can express ΔI as an implicit function of Δc, such that

$$WTP(I_n - c_{in}, x_{ink}) = \frac{\partial \Delta I}{\partial \Delta x} = -\frac{\partial EM(I_n - \Delta I - c_{in}, x_{ink} + \Delta x)/\partial \Delta x}{\partial EM(I_n - \Delta I - c_{in}, x_{ink} + \Delta x)/\partial \Delta I}. \tag{10.59}$$

Using the property (3.82) of expected maximum utility, and assuming that x_{ink} is a continuous variable[1], and that the utility function is differentiable[2] in x_{ink}, the willingness to pay for an infinitesimal modification of x_{ink} is

$$WTP(I_n - c_{in}, x_{ink}) = -\frac{\frac{\partial EM}{\partial V_{in}}\frac{\partial V_{in}}{\partial \Delta x}}{\frac{\partial EM}{\partial V_{in}}\frac{\partial V_{in}}{\partial \Delta I}} = -\frac{P_n(i)\frac{\partial V_{in}}{\partial \Delta x}}{P_n(i)\frac{\partial V_{in}}{\partial \Delta I}} = -\frac{\frac{\partial V_{in}}{\partial \Delta x}}{\frac{\partial V_{in}}{\partial \Delta I}}. \tag{10.60}$$

This definition of WTP assumes that the increase of x_{ink} improves the utility, that is $\partial V_{in}/\partial x_{ink} > 0$. This would be the case, for example, for a variable such as the amount of clean water in a lake in a swimming site choice context. As a decrease of income decreases the utility, that is $\partial V_{in}/\partial \Delta I < 0$, the

[1] If x_{ink} is a discrete variable, the willingness to pay is directly obtained by solving (10.58).

[2] This assumption will be relaxed later on.

definition (10.60) of the willingness to pay corresponds to a positive number, as expected.

If x_{ink} and $I_n - c_{in}$ appear linearly in the utility function, that is if

$$V_{in}(I_n - \Delta I - c_{in}, \ldots, x_{in} + \Delta x, \ldots) = \beta_c(I_n - \Delta I - c_{in}) + \beta_k(x_{ink} + \Delta x) + \cdots,$$

with $\beta_c, \beta_k > 0$, then the willingness to pay simplifies into

$$\mathrm{WTP}(I_n - c_{in}, x_{ink}) = \frac{\beta_k}{\beta_c}, \tag{10.61}$$

and is constant.

Now, consider a situation where the increase in the variable x_{ink} deteriorates the overall utility, that is $\partial V_{in}/\partial x_{ink} < 0$. This would be the case, for example, for a variable such as travel time in a mode choice context. In this case, we are interested in the net **increase** in income (or, equivalently, the decrease in cost) that makes the expected utility equal with the modified value of the variable. This quantity is called the *willingness to accept*. It can be derived in the exact same manner as the willingness to pay, by solving the equation

$$\mathrm{EM}(c_{in}, x_{ink}) = \mathrm{EM}(c_{in} - \Delta c, x_{ink} + \Delta x). \tag{10.62}$$

so that

$$\mathrm{WTA}(x_{ink}) = \frac{\dfrac{\partial V_{in}}{\partial x_{ink}}}{\dfrac{\partial V_{in}}{\partial c_{in}}}. \tag{10.63}$$

In the case of a linear-in-parameter specification, the WTA simplifies to

$$\mathrm{WTA}(x_{ink}) = \frac{\beta_k}{\beta_c}. \tag{10.64}$$

Now, the same derivations can be performed if the value of the variable **decreases** instead of increasing. For example, we can define the willingness to pay for a decrease in travel time (this is what we called the value-of-time in the beginning of this section), as well as the willingness to accept for a decrease of quantity of clean water in a lake. To do this, we observe that if x_{ink} decreases, $-x_{ink}$ increases and the above results apply. In summary,

- if x_{ink} is such that $\partial V_{in}/\partial x_{ink} > 0$ (such as the quantity of clean water in a lake),

$$\mathrm{WTP}(x_{ink}) = -\frac{\dfrac{\partial V_{in}}{\partial x_{ink}}}{\dfrac{\partial V_{in}}{\partial c_{in}}}, \tag{10.65}$$

and

$$\mathrm{WTA}(x_{ink}) = \frac{\dfrac{\partial V_{in}}{\partial - x_{ink}}}{\dfrac{\partial V_{in}}{\partial c_{in}}} = -\frac{\dfrac{\partial V_{in}}{\partial x_{ink}}}{\dfrac{\partial V_{in}}{\partial c_{in}}} = \mathrm{WTP}(x_{ink}), \tag{10.66}$$

and the willingness to pay and the willingness to accept are equal;

- if x_{ink} is such that $\partial V_{in}/\partial x_{ink} < 0$ (such as the travel time),

$$\mathrm{WTA}(x_{ink}) = \frac{\dfrac{\partial V_{in}}{\partial x_{ink}}}{\dfrac{\partial V_{in}}{\partial c_{in}}}, \tag{10.67}$$

and

$$\mathrm{WTP}(x_{ink}) = -\frac{\dfrac{\partial V_{in}}{\partial - x_{ink}}}{\dfrac{\partial V_{in}}{\partial c_{in}}} = \frac{\dfrac{\partial V_{in}}{\partial x_{ink}}}{\dfrac{\partial V_{in}}{\partial c_{in}}} = \mathrm{WTA}(x_{ink}), \tag{10.68}$$

and the willingness to pay and the willingness to accept are again equal.

The fact that the willingness to pay is equal to the willing to accept is a consequence of the assumption that the utility function is differentiable in x_{ink}. However, it has been shown empirically (Kahneman et al., 1991) that the willingness to accept may be higher than the willingness to pay, as people put more value in a good that they already have. Therefore, it may be relevant to relax the assumption that the utility function is differentiable in x_{ink}. It is actually sufficient that it is continuous and differentiable to the left or to the right for the derivation of WTP and WTA. Examples of such utility function are depicted in Figure 10.4, where the current value of the variable (the reference point) is represented by an horizontal line. In Figure 10.4(a), the utility function is increasing with the value of the variable, such as the example of the amount of clean water in the lake. In Figure 10.4(b), the utility function is decreasing with the value of the variable, such as the example of the travel time. The specification of the two utility functions is linear-in-parameters, where the slope is different for an increase and a decrease of the variable:

$$V_{in} = \beta_x^+ \max(x_i - x_n^{ref}, 0) + \beta_x^- \min(x_i - x_n^{ref}, 0) + \beta_c c_{in} \cdots, \tag{10.69}$$

where x_n^{ref} is the current value of the variable for individual n. These utility functions are not differentiable at the reference point, but they are everywhere else. In Figure 10.4(a), β_x^+ and β_x^- are positive. The willingness to pay for an increase of the variable is equal to $-\beta_x^+/\beta_c$, and the willingness to accept for a decrease of the variable is equal to $-\beta_x^-/\beta_c$. In Figure 10.4(b), β_x^+ and β_x^- are negative. The willingness to pay for a decrease of the variable is equal to β_x^-/β_c, and the willingness to accept for an increase of the variable is equal to β_x^+/β_c.

More formally, if $\partial_- V_{in}/\partial x_{ink}$ denotes the left derivative of the utility function, and $\partial_+ V_{in}/\partial x_{ink}$ its first derivative, we have:

- If x_{ink} is such that $\partial_+ V_{in}/\partial x_{ink} > 0$,

$$\text{WTP}(x_{ink}) = -\frac{\dfrac{\partial_+ V_{in}}{\partial x_{ink}}}{\dfrac{\partial_+ V_{in}}{\partial c_{in}}}. \tag{10.70}$$

- If x_{ink} is such that $\partial_- V_{in}/\partial x_{ink} > 0$,

$$\text{WTA}(x_{ink}) = -\frac{\dfrac{\partial_- V_{in}}{\partial x_{ink}}}{\dfrac{\partial_- V_{in}}{\partial c_{in}}}. \tag{10.71}$$

- If x_{ink} is such that $\partial_+ V_{in}/\partial x_{ink} < 0$,

$$\text{WTA}(x_{ink}) = \frac{\dfrac{\partial_+ V_{in}}{\partial x_{ink}}}{\dfrac{\partial_+ V_{in}}{\partial c_{in}}}. \tag{10.72}$$

- If x_{ink} is such that $\partial_- V_{in}/\partial x_{ink} < 0$,

$$\text{WTP}(x_{ink}) = \frac{\dfrac{\partial_- V_{in}}{\partial x_{ink}}}{\dfrac{\partial_- V_{in}}{\partial c_{in}}}. \tag{10.73}$$

We conclude this discussion by some additional remarks:

- If the utility function is non monotonic in x_{ink}, we may have two values for WTP or WTA. Indeed, if any modification of the variable (up or down) generates an increase (resp. a decrease) of the utility function, the associated quantities will be both WTP (resp. WTA).

- The derivation (10.60) assumes that both x_{ink} and c_{in} appears only in the utility function of alternative i. If it is not the case, (10.59) must be used with

$$\frac{EM(c_{in}, x_{ink})}{\partial x_{ink}} = \sum_{j \in \mathcal{C}_n} \frac{EM(c_{in}, x_{ink})}{\partial V_{jn}} \frac{\partial V_{jn}}{\partial x_{ink}}. \tag{10.74}$$

- When the utility function is linear in parameters, WTP and WTA are constant. If it is not the case, they depend on the current value of the variable.

10.5.6 Revenue calculator

The aggregation procedures described in Section 10.2, designed to calculate the market share of an alternative i (10.3) can easily be adapted to compute the revenues associated with alternative i. To do so, the cost c_{in} of alternative i for decision-maker n must be an explanatory variable of the model, that is

$$P_n(i|x_n) = P_n(i|c_{in}, \bar{x}_n),$$

where $\bar{x}_n$ is the vector of all explanatory variables, except c_{in}. In this case, the expecte revenue generated by individual n is

$$c_{in} P_n(i|c_{in}, \bar{x}_n).$$

Therefore, the total expected revenue R_i associated with alternative i is given by

$$R_i = \sum_{n=1}^{N_T} c_{in} P_n(i|c_{in}, \bar{x}_n). \tag{10.75}$$

In practice, c_{in} may happen to be constant across the population, and therefore denoted by c_i. Then, (10.75) simplifies to

$$R_i = c_i \sum_{n=1}^{N_T} P_n(i|c_i, \bar{x}_n) = c_i N_T W(i). \tag{10.76}$$

A relevant problem for the provider of alternative i is to optimize the cost c_i. The probability to use/purchase an alternative increases if the cost *decreases*. Also, the revenue per user increases if the cost *increases*. A trade-off must be found here by solving the following optimization problem:

$$\max_{c_i} c_i \sum_{n=1}^{N_T} P_n(i|c_i, \bar{x}_n). \tag{10.77}$$

Note that this formulation assumes that all other variables $\bar{x}_n$ are constant. In competitive markets, it is likely that competing alternatives will see their cost adjusted if the cost of alternative i is modified. In that case, solving (10.77) is not appropriate and a game theoretic approach (Fundenberg and Tirole, 1991, Bresnahan and Reiss, 1991) or an explicit simulation of the interactions of decision-makers (see Section 10.5.7) should be preferred.

We illustrate the cost optimization procedure on a simple binary logit model, with the following utility functions:

$$\begin{aligned} V_{1n} &= \beta_c c_1 - 0.5, \\ V_{2n} &= \beta_c c_2. \end{aligned}$$

We would like to analyze the revenues for alternative 1 as a function of its cost, for an homogenous market of 1000 individuals. Figure 10.5 illustrates the example, assuming that the cost of alternative 2 is 2 €, and the cost of alternative 1 ranges from 0 to 4 €. It appears that a low cost of the alternative is associated with high market shares, but low revenues. If the cost is high, both the market share and the revenue are low. In this example, the optimal value for the cost is c_1=1.436 €, associated with a market share of 65.2% and a total revenue of 936 €.

If the population is heterogenous, the revenue function may become more complicated and exhibit several local optima. For example, let's consider a population composed of two groups. Group 1 contains 400 individuals, while group 2 contains 600 individuals. They differ by their cost coefficient, which is $\beta_c = -2$ for group 1, and $\beta_c = -0.2$ for group 2. Individuals in group 2 are much less sensitive to cost that those from group 1. The revenue generated by alternative 1 as a function of its cost is plotted on Figure 10.6. The curve has two local maxima. One at a low price (c_1=1.829 €), generating revenue from both groups, and one at a higher price (c_2=6.264 €), generating revenue almost exclusively from group 2. The second option generates a slightly higher revenue (772 € instead of 760 €). In practice, it does not mean that the higher price should be necessarily preferred. A sensitivity analysis has to be performed to account for various imprecisions in the model and the data (see Section 10.6). Also, other business dimensions will have to be considered. For example, is there any risk to generate revenue only based on one group of the population? Is the solution to exclude group 1 from the market by setting a high price consistent with equity considerations?

10.5.7 Supply-demand interactions

10.6 Sensitivity analysis and confidence intervals

In Section 10.5, we have reviewed a list of policy indicators derived from the choice model

$$P(i|x_n;\theta). \tag{10.78}$$

Even if the model specification is assumed to be correct, the variables x_n and the parameters θ are subject to errors that propagate to the indicators through the calculation of the choice probability. The variables x_n are subject to measurement errors, and to approximations based on the definition of the scenarios. Therefore, they should be considered as random variables, with a distribution that depends on the measurement method, and the scenario definition. Modeling the impact of measurement errors is not a trivial task. It is specific to the application and to the type of data collection method (see for instance Biemer et al., 2004, Part V). The parameters θ are the maximum likehood estimates based on the estimation sample. The theory of maximum likelihood estimation tells us that they are random variables that are asymptotically normally distributed, and that their variance can be approximated.

As a consequence, the choice probability of each individual n is also a random variable, and so are the policy indicators. However, it is impossible to obtain an analytical derivation of the distribution of the choice probability, due to the non linearity of the model. Therefore, it is required to rely on Monte-Carlo simulation to perform a sensitivity analysis and to compute condifence intervals.

The method is quite simple. A sequence $(x_n^r, \theta^r)_{r=1}^R$ of R realizations of the random variables x_n and θ is generated (using the techniques described in Section B.3). For each r, we compute the choice probability $P^r(i|x_n^r;\theta^r)$, and derive the policy indicator. For instance, we can use (10.3) to derive the market share of alternative i, that is

$$W^r(i) = \frac{1}{N_T}\sum_{n=1}^{N_T} P^r(i|x_n^r;\theta^r). \tag{10.79}$$

Any statistic can now be derived from the generated sequence $(W^r(i))_{r=1}^R$, such as the mean

$$\overline{W}^R(i) = \frac{1}{R}\sum_{r=1}^{R} W^r(i), \tag{10.80}$$

and the standard deviation

$$\sigma^R_{W(i)} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(W^r(i) - \overline{W}^R(i))^2}. \tag{10.81}$$

As R goes to ∞, these values converge to the required statistics. It is useful to compute the 5% and the 95% quantiles. To do so, sort the values of $W^r(i)$ by increasing values, where $W^1(i)$ is the smallest and $W^R(i)$ the largest. Define $r_1(q) = \lceil Rq \rceil$[3], and $r_2(q) = \lfloor Rq + 1 \rfloor$[4]. The q-quantile is

$$Q(q) = \frac{W^{r_1(q)}(i) + W^{r_2(q)}(i)}{2}. \tag{10.82}$$

For instance, if there are $R = 1000$ values, the 5% quantile is the average of the values indexed $r_1(0.05) = 50$ and $r_2(0.05) = 51$. The 95% quantile is the mean of values indexed $r_1(0.95) = 950$ and $r_2(0.95) = 951$. If there are $R = 1001$ values, the 5% quantile is the value indexed $r_1(0.05) = r_2(0.05) = 51$, and the 95% quantile in the value indexed $r_1(0.95) = r_2(0.95) = 951$. The 90% confidence interval can now be defined as

$$[Q(0.05), Q(0.95)]\,. \tag{10.83}$$

10.7 Illustration

We illustrate the concepts discussed above using the transportation mode choice model described in Section 5.8. As it is a revealed preferences data set, we have used the estimation sample to perform an aggregation with sample enumeration (see Section 10.2.3). The observations where the income was not reported were excluded from the estimation sample. They were integrated in the forecasting sample. For those observation, income was infered from the level of education: the weighted average of income of all people with the same education level was imputed. We obtain a sample of 1818 individuals (as opposed to 1723 in the estimation sample). Each respondent has been associated with a weight. The weights have been computed using the iterative proportional fitting (IPF) algorithm (see Section 10.2.2) in order to reproduce the population shares for education, age and gender (as given by the 2000

[3] $\lceil x \rceil$ denotes the smallest integer that is not less than x. It is the value of x "rounded up" to the closest integer.

[4] $\lfloor x \rfloor$ denotes the largest integer that is not greater than x. It is the value of x "rounded down" to the closest integer.

Swiss Federal Census). They have been normalized so that they sum up to 1.

We first look at some indicators derived from the model. The distribution of the value of time for public transportation in the population is represented in Figure 10.7. It appears clearly that a significant share of the population has a value of time equal to zero (actually, 39%). Remember that the model has been specified in terms of marginal cost of public transportation. It is the price of the trip, excluding the cost of travelcards. In particular, this cost is zero for holders of a season ticket (see Section 5.8). The same distribution is represented in Figure 10.8, where those travelers have not been reported. The mean value of time across the whole population is 14.5 CHF/hour. The mean value of time calculated only on nonzero values is 21.2 CHF/hour. The value of time for the car is much lower: 1.14 CHF/hour for the whole population, and 1.71 CHF/hour for travelers with non zero value of time. This is explained by the exclusive use of the gasoline cost in the model. These costs correspond only to 15.1% of the total cost of a car (Touring Club Suisse, n.d.).

We then compute the market shares for the base case. It is the scenario corresponding to the current situation. We obtain 62.0% for car, 32.1% for public transportation and 5.90% for slow modes. We also compute the 90% condifence intervals, using 100 draws for the Monte-Carlo simulation described in Section 10.6. Note that we assume here that the explanatory variables are free of errors. The sensitivity analysis is performed only with respect to the coefficients of the choice model, that follow a multivariate normal distribution. We obtain $[56.05\% - 67.2\%]$ for car, $[27.0\% - 37.5\%]$ for public transportation and $[3.72\% - 9.11\%]$ for slow modes, as illustrated in Figure 10.9.

The distribution of disaggregate cost elasticity for car is depicted in Figure 10.10. Although some extreme values (-12) are reported, 94% of the population has an elasticity higher than -1, 73% higher than -0.1 and 28% higher than -0.01. The aggregate cost elasticity, calculated using (10.36) is equal to -0.102.

Now, we simulate the following situation. The public transportation company would like to analyze the impact of the fare on the market shares, as well as on its revenues. To do so, we are considering many scenarios, where the marginal cost that travelers are paying represents a percentage of today's marginal cost. This percentage is supposed to be the same for the whole population. Using the method described in Section 10.5.6, Figure 10.9 illustrates the analysis for price modification ranging from 0% (free service) to 1000% (that is ten times today's cost). We observe the non-concavity of the revenue curve. A first local optimum of the revenue function is observed

when the cost is increased by about 300%. After, the revenue decreases until about 400%, and start increasing again. The next local maximum lies beyond the 1000% increase of the marginal cost, and has not been reached by this analysis.

We have also computed the 90% condifence intervals, using 100 draws for the Monte-Carlo simulation described in Section 10.6. The results for the market shares are presented in Figure 10.12, and for the revenues in Figure 10.13. We observe two typical features of the confidence intervals: they are asymmetric, and their size increases as we deviate from the current situation.

10.8 Summary

This chapter has discussed the problem of using models of individual choice to obtain forecasts of the expected aggregate behavior of a population as well as indicators for policy analysis. Four qualitatively distinct approaches (some of which are special cases of others) were presented:

1. average individual,
2. synthetic population,
3. sample enumeration,
4. microsimulation.

The issues that should be considered in applying each of these procedures were summarized.

The chapter has also explained how to adjust the alternative specific constants of the model before applying it, and how to include a new alternative, with and without the availability of stated preferences data.

The following policy indicators have been derived:

- disaggregate direct point elasticity:

$$E_{x_{ink}}^{P_n(i)} = \frac{\partial P_n(i)}{\partial x_{ink}} \frac{x_{ink}}{P_n(i)},$$

- disaggregate cross point elasticity:

$$E_{x_{jnk}}^{P_n(i)} = \frac{\partial P_n(i)}{\partial x_{jnk}} \frac{x_{jnk}}{P_n(i)},$$

- aggregate direct point elasticiy:

$$E_{x_{ik}}^{W(i)} = \frac{\sum_{n=1}^{N_T} P_n(i) E_{x_{ink}}^{P_n(i)}}{\sum_{n=1}^{N_T} P_n(i)},$$

- aggregate cross point elasticiy:

$$E_{x_{jk}}^{W(i)} = \frac{\sum_{n=1}^{N_T} P_n(i) E_{x_{jnk}}^{P_n(i)}}{\sum_{n=1}^{N_T} P_n(i)},$$

- disaggregate direct arc elasticity

$$E_{\Delta x_{ink}}^{\Delta P_n(i)} = \frac{\Delta P_n(i)}{P_n(i)} \frac{x_{ink}}{\Delta x_{ink}},$$

- disaggregate cross arc elasticity

$$E_{\Delta x_j}^{\Delta W(i)} = \frac{\Delta W(i)}{W(i)} \frac{x_{jk}}{\Delta x_{jk}},$$

- the willingness to pay

$$\mathrm{WTP}(x_{ink}) = \pm \frac{\dfrac{\partial V_{in}}{\partial x_{ink}}}{\dfrac{\partial V_{in}}{\partial c_{in}}}.$$

- incremental logit

$$P_n'(i) = \frac{e^{V_{in}+\Delta V_{in}}}{\sum_{j\in\mathcal{C}_n} e^{V_{jn}+\Delta V_{jn}}} = \frac{P_n(i) e^{\Delta V_{in}}}{\sum_{j\in\mathcal{C}_n} P_n(j) e^{\Delta V_{jn}}},$$

- consumer surplus, which is equal to the difference of expected maximum utility between two scenarios. For logit, it is

$$\frac{1}{\mu} \ln \sum_{j\in\mathcal{C}^2} e^{\mu V_j^2} - \frac{1}{\mu} \ln \sum_{j\in\mathcal{C}^1} e^{\mu V_j^1},$$

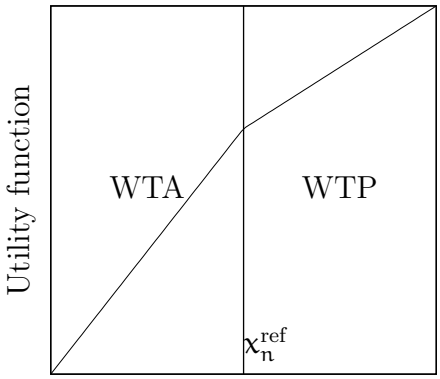
  and for nested logit, it is

$$\frac{1}{\mu} \ln \sum_{m=1}^{M} \left(\sum_{\ell=1}^{J_m} e^{V_\ell^2 \mu_m} \right)^{\mu/\mu_m} - \frac{1}{\mu} \ln \sum_{m=1}^{M} \left(\sum_{\ell=1}^{J_m} e^{V_\ell^1 \mu_m} \right)^{\mu/\mu_m}, \qquad (10.84)$$

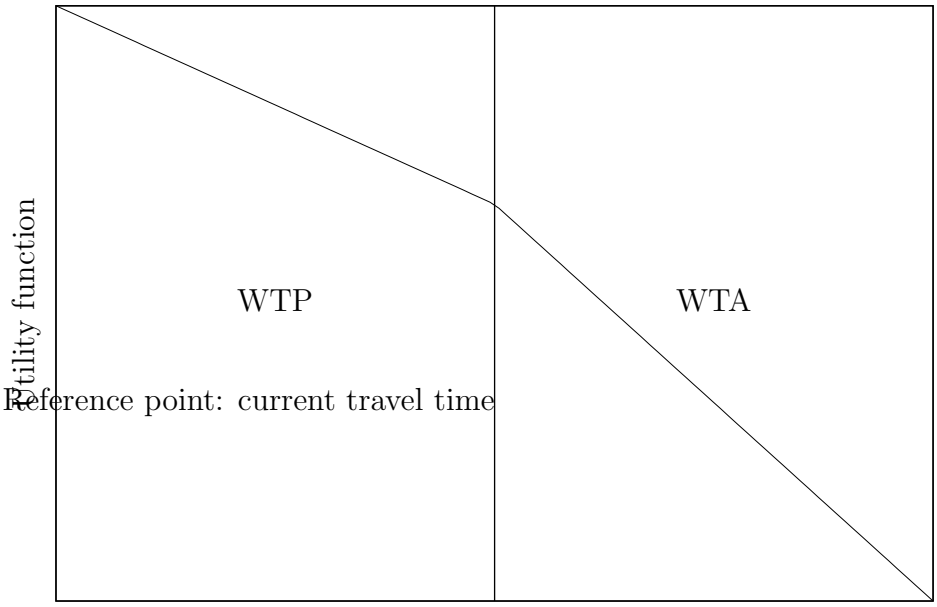
- a calculation of the total revenue generated by an alternative

$$R_i = \sum_{n=1}^{N_T} c_{in} P_n(i|c_{in}, \bar{x}_n),$$

The chapter has also shown how to use simulation to perform sensitivity analysis and generate confidence intervals.



(a) $\partial V_{in}/\partial x_{ink} > 0$



(b) $\partial V_{in}/\partial x_{ink} < 0$

Figure 10.4: Example of non differentiable utility functions

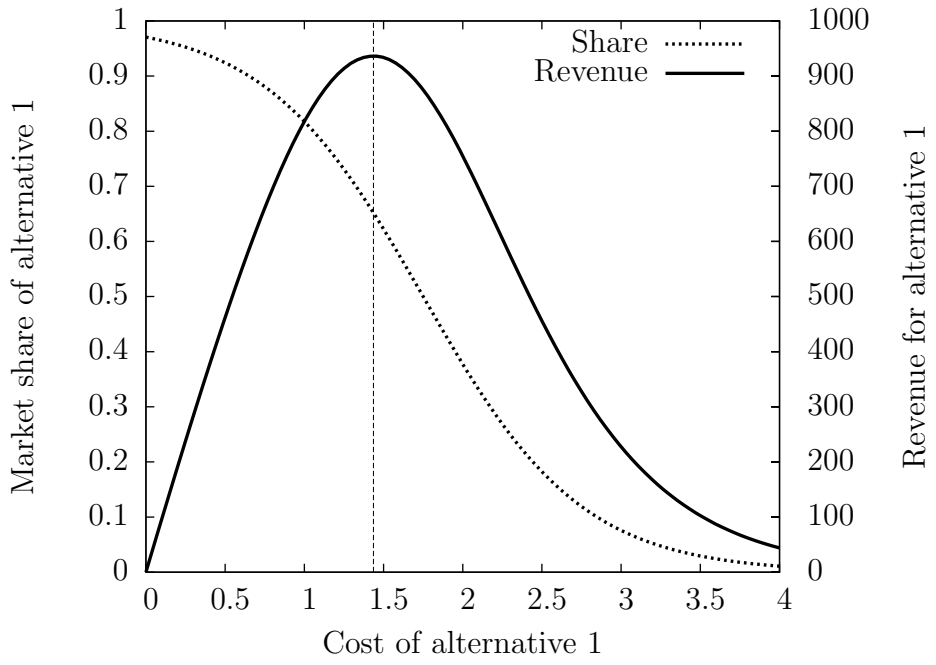


Figure 10.5: Market share and revenue as a function of cost - homogenous population

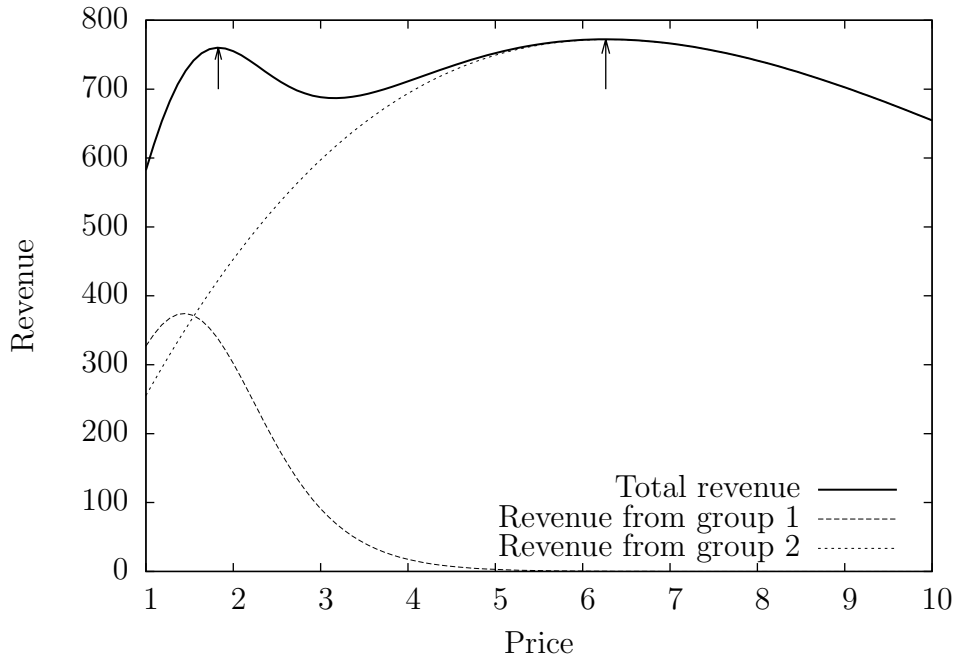


Figure 10.6: Revenue as a function of cost - heterogenous population

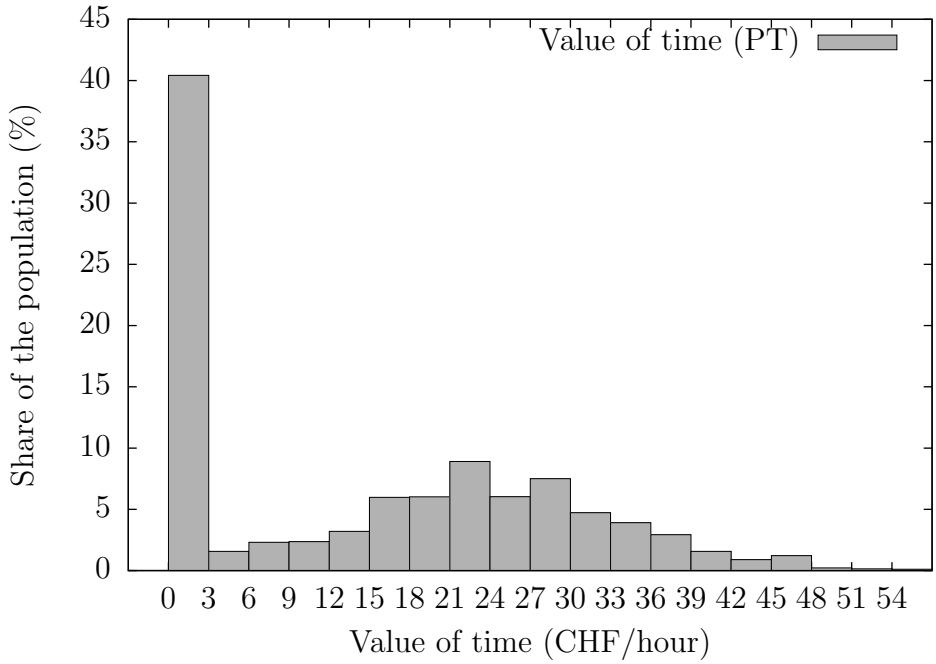


Figure 10.7: Distribution of value of time (PT) in the population

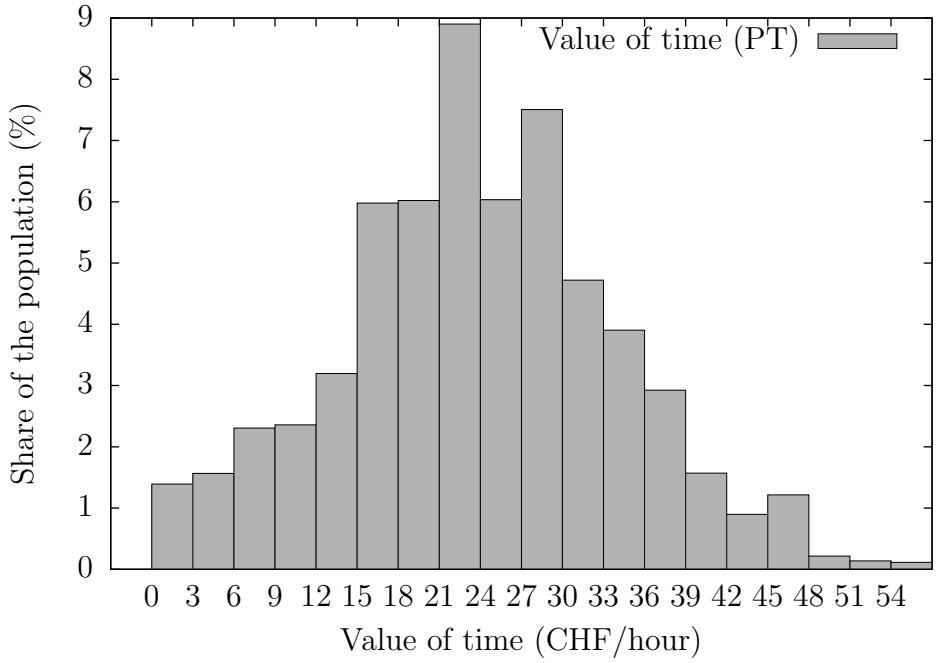


Figure 10.8: Distribution of value of time (PT) in the population (only non zero)

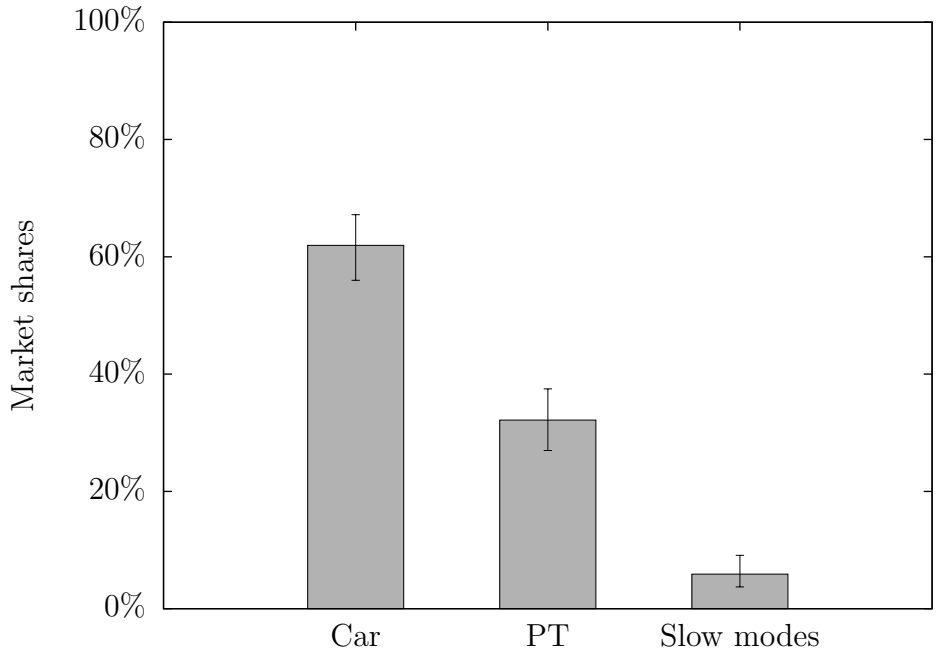


Figure 10.9: Swiss transport mode choice: market shares for the base case

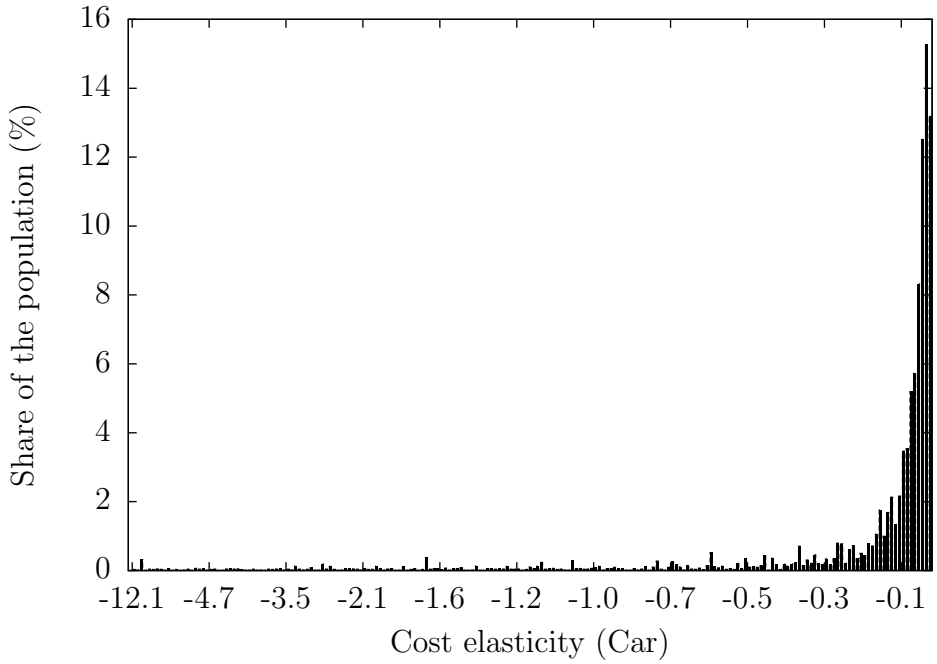


Figure 10.10: Distribution of the cost elasticity (car) in the population

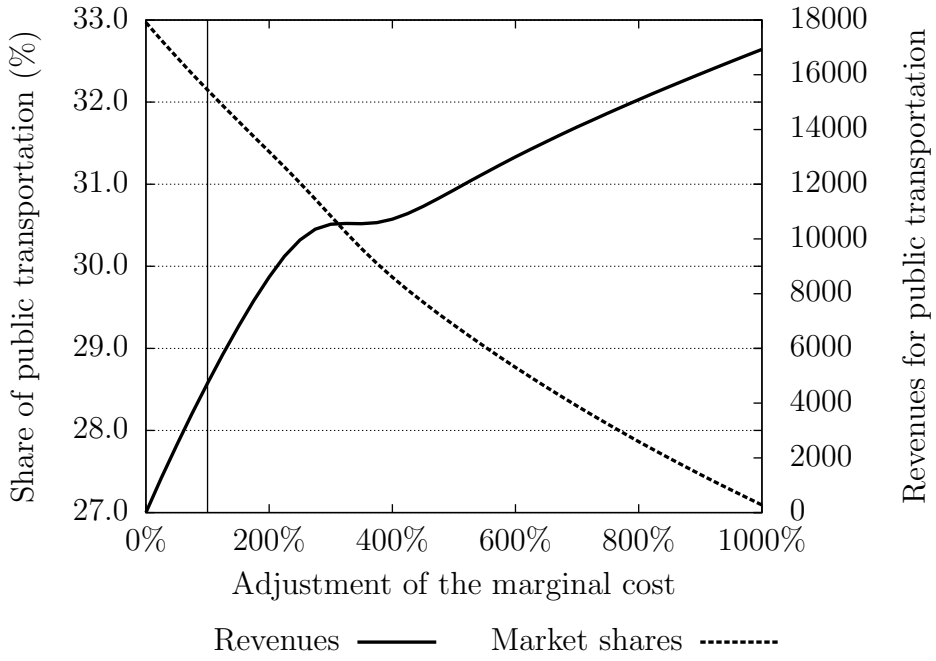


Figure 10.11: Impact of the marginal cost of public transportation on the market shares and the revenues

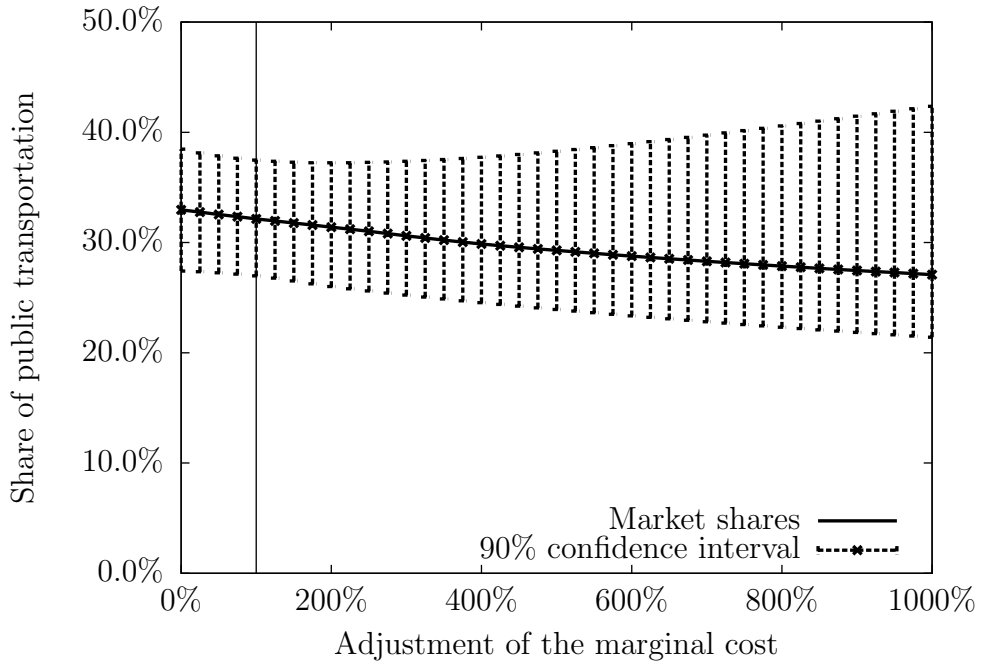


Figure 10.12: Impact of the marginal cost of public transportation on the market shares: confidence intervals

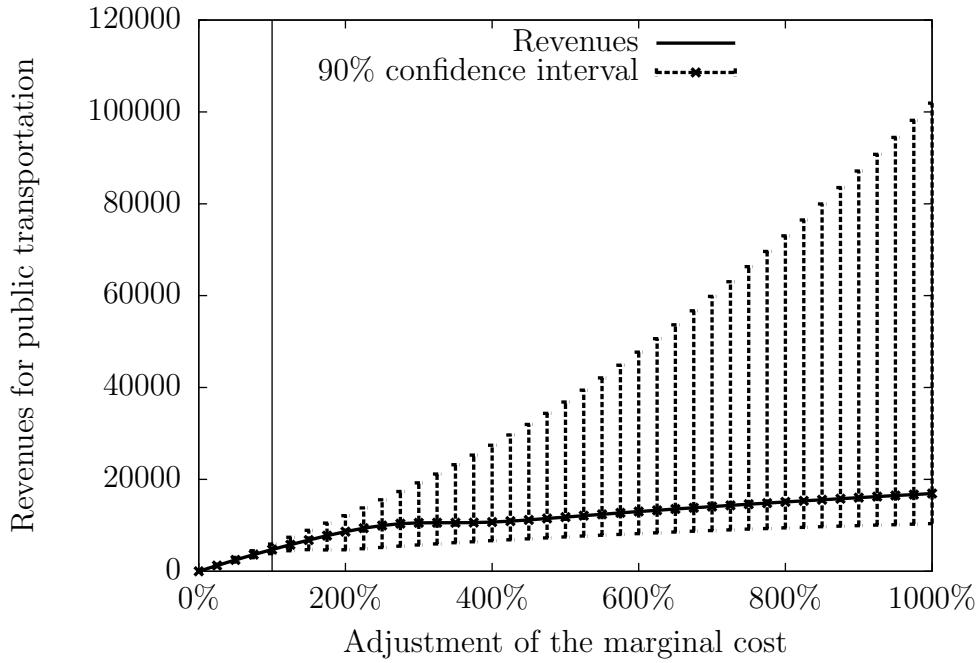


Figure 10.13: Impact of the marginal cost of public transportation on the revenues: confidence intervals

Part III

Advanced methods

Part IV

Additional material

Appendix A

Notations

> "*There are two things people kill each other for: parking and notations*"

It is assumed that all vectors are column vectors. If β is a vector, we usually denote its length by with K^β. Therefore,

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{K^\beta} \end{pmatrix} \text{ and } \beta^T = (\beta_1, \ldots, \beta_{K^\beta}),$$

where superscript T denotes the transpose operator. Also, $\beta^T x$ denotes the inner product between vectors β and x, that is

$$\beta^T x = \sum_{k=1}^{K_\beta} \beta_k x_k.$$

We now provide the list of notations and conventions used throughout the book. In order to avoid confusion due to an excess of formalism, these conventions may be relaxed in the text, when no ambiguity is possible.

$\mathbb{N}$	set of strictly positive natural numbers	
$\mathbb{R}$	set of real numbers	
$\mathbb{R}^+$	set of non negative real numbers	
$\mathbb{R}^K$	set of column vectors of size K	
$\mathbb{R}^{K\times L}$	set of matrix with K rows and L columns	
β_i	ith elements of vector i	
K^θ	length of vector θ	$\mathbb{N}$

K	by default, length of vector β, that is $K = K^\beta$	$\mathbb{N}$
n	index for individuals	$\mathbb{N}$
N	number of individuals in the sample	$\mathbb{N}$
N_T	number of individuals in the population	$\mathbb{N}$
$\mathcal{C}$	universal choice set	
$\mathcal{C}_n$	choice set of individual n	
$i, j \in \mathcal{C}_n$	indices of alternatives	
J	total number of alternatives in $\mathcal{C}$	$\mathbb{N}_0$
J_n	number of alternatives in $\mathcal{C}_n$	$\mathbb{N}$
A_n	availability matrix. It is a $J_n \times J$ matrix obtained from the $J \times J$ identity matrix removing the rows corresponding to alternatives that are unavailable to individual n	$\mathbb{R}^{J_n \times J}$
y_{nt}	vector of dependent variables for individual n at time t	
$\theta = (\theta_1, \ldots, \theta_{K^\theta})^T$	vector of all unknown parameters	$\mathbb{R}^{K^\theta}$
$\beta = (\beta_1, \ldots, \beta_K)^T$	vector of unknown coefficients in the systematic part of the utility	$\mathbb{R}^K$
$\gamma = (\gamma_1, \ldots, \gamma_{K^\gamma})^T$	vector of unknown parameters that are not coefficients	$\mathbb{R}^{K^\gamma}$
$\theta^V = (\beta^T \gamma^T)^T,$ $= (\theta_1, \ldots, \theta_{K^V})^T$	vector of all unknown parameters involved in the systematic part of the utility	$\mathbb{R}^{K^V}$
S_n	vector of characteristics of individual n	
z_{in}	vector of attributes describing alternative i as perceived by individual n	
$x_{in} = (x_{in1}, \ldots, x_{inK^x})^T$	vector of explanatory variables for alternative i and individual n, function of attributes and socio-economic characteristics, that is $x_{in} = h(z_{in}, S_n)$ (for notational simplification, K^x may be denoted by K or L in the text)	$\mathbb{R}^{K^x}$

$X_n^{\mathcal{C}} = (x_{1n}, \ldots, x_{Jn})^T$	matrix of explanatory variables for individual n and all alternatives	$\mathbb{R}^{J \times K^x}$
$X_n = A_n X_n^{\mathcal{C}}$	matrix of explanatory variables for individual n and available alternatives	$\mathbb{R}^{J_n \times K^x}$
$U_n^{\mathcal{C}}$	vector specific to individual n containing the utilities of all alternatives, where entries corresponding to unavailable alternatives are arbitrary	$\mathbb{R}^J$
$U_n = A_n U_n^{\mathcal{C}}$	vector containing the utilities of alternatives available to individual n	$\mathbb{R}^{J_n}$
$U_{in} = V_{in} + \varepsilon_{in}$	utility of alternative i for individual n	$\mathbb{R}$
$V_n^{\mathcal{C}}$	systematic part of $U_n^{\mathcal{C}}$	$\mathbb{R}^J$
$V_n = A_n V_n^{\mathcal{C}}$	systematic part of U_n	$\mathbb{R}^{J_n}$
V_{in}	systematic part of U_{in}	$\mathbb{R}$
$V(\theta^V, x_{in})$	systematic part of the utility as a function of parameters θ^V and explanatory variables x_{in}	
$\varepsilon_n^{\mathcal{C}}$	vector of error terms of $U_n^{\mathcal{C}}$	$\mathbb{R}^J$
$\varepsilon_n = A_n \varepsilon_n^{\mathcal{C}}$	vector of error terms of U_n	$\mathbb{R}^{J_n}$
ε_{in}	error term of U_{in}	$\mathbb{R}$
Σ_ξ	variance-covariance matrix of the random vector ξ	$\mathbb{R}^{K^\xi \times K^\xi}$
$\Pr(B)$	probability of an event B	
$P(i\|\mathcal{C}_n)$	probability of individual n choosing alternative i within $\mathcal{C}_n$	$\mathbb{R}$
$\Lambda(i\|\mathcal{C}_n)$	probability of individual n choosing alternative i within $\mathcal{C}_n$, as given by the multinomial logit model, that is $\Lambda_n(i\|\mathcal{C}_n) = e^{V_{in}} / \sum_{j \in \mathcal{C}_n} e^{V_{jn}}$	$\mathbb{R}$
ζ	standard normal distributed random variable	
ξ	normal distributed random variable	

ν	Gumbel distributed random variable	
$\mathcal{L}^*(\theta)$	likelihood function	$\mathbb{R}^{K^\theta} \to \mathbb{R}$
$\mathcal{L}(\theta)$	log-likelihood function, $\mathcal{L}(\theta) = \ln \mathcal{L}^*(\theta)$	$\mathbb{R}^{K^\theta} \to \mathbb{R}$,
$E[\xi]$	expectation operator	$\mathbb{R}^{K^\xi} \to \mathbb{R}^{K^\xi}$
$Var[\xi]$	variance-covariance matrix operator	$\mathbb{R}^{K^\xi} \to \mathbb{R}^{K^\xi \times K^\xi}$
$E^{P_n(i)}_{x_{ink}}$	point elasticity	$\mathbb{R}$
$E^{\Delta P_n(i)}_{\Delta x_{ink}}$	arc elasticity	$\mathbb{R}$
$f_\xi(\cdot)$	probability density function (pdf) of continuous random variable ξ	$\mathbb{R}^{K^\xi} \to \mathbb{R}^+$
$p_\xi(\cdot)$	probability mass function of discrete random variable ξ	$\mathbb{R}^{K^\xi} \to \mathbb{R}^+$
$F_\xi(\cdot)$	cumulative distribution function (CDF) of random variable ξ	$\mathbb{R}^{K^\xi} \to [0, 1]$
$\phi(\cdot)$	probability density function of the univariate standardized normal distribution	$\mathbb{R} \to \mathbb{R}^+$
$\Phi(\cdot)$	cumulative distribution function of the univariate standardized normal distribution	$\mathbb{R} \to [0, 1]$
$\mathbf{n}_\xi(E[\xi], Var(\xi))$	multivariate normal probability density function of $\xi \in R^m$ with mean $E[\xi]$ and variance-covariance matrix $Var(\xi)$	$\mathbb{R}^m \to \mathbb{R}$
R	number of draws in simulation context	$\mathbb{N}$
r	index of the draws in simulation context	$\mathbb{N}$
μ	scale parameter of the McFadden Multivariate Extreme Value (MEV) distribution	$\mathbb{R}^+$
σ	scale parameter of the univariate normal distribution	$\mathbb{R}$
Γ	Cholesky factor of Σ, that is $\Gamma\Gamma^T = \Sigma$	$\mathbb{R}^{m \times m}$
$F^{\mathcal{C}}_n$	factor loading matrix for individual n, with M factors	$\mathbb{R}^{J \times M}$

$F_n = A_n F_n^c$	factor loading matrix for available alternatives	$\mathbb{R}^{J_n \times M}$
T_n	number of observations specific to individual n in a panel data context	$\mathbb{N}$
t	index of observations specific to an individual in a panel data context	$\mathbb{N}$
Δ_i	it is a $J-1 \times J-1$ identity matrix where a column of -1 has been inserted in the ith position, in such a way that $\Delta_i U = (U_1 - U_i, \ldots, U_{i-1} - U_i, \ldots, U_{i+1} - U_i, \ldots, U_j - U_i)^T$	$\mathbb{R}^{J-1 \times J}$
G	number of market segments, or groups	$\mathbb{N}$
$G(\cdot)$	generating function for the McFadden-MEV model	$\mathbb{R}^J \to \mathbb{R}^+$
$U[a, b]$	uniform distribution between a and b	
$N(\beta, \sigma^2)$	normal distribution with mean β and standard deviation σ.	

Appendix D

Tables

K	90%	95%	99%	K	90%	95%	99%
1	2.706	3.841	6.635	21	29.615	32.671	38.932
2	4.605	5.991	9.210	22	30.813	33.924	40.289
3	6.251	7.815	11.345	23	32.007	35.172	41.638
4	7.779	9.488	13.277	24	33.196	36.415	42.980
5	9.236	11.070	15.086	25	34.382	37.652	44.314
6	10.645	12.592	16.812	26	35.563	38.885	45.642
7	12.017	14.067	18.475	27	36.741	40.113	46.963
8	13.362	15.507	20.090	28	37.916	41.337	48.278
9	14.684	16.919	21.666	29	39.087	42.557	49.588
10	15.987	18.307	23.209	30	40.256	43.773	50.892
11	17.275	19.675	24.725	31	41.422	44.985	52.191
12	18.549	21.026	26.217	32	42.585	46.194	53.486
13	19.812	22.362	27.688	33	43.745	47.400	54.776
14	21.064	23.685	29.141	34	44.903	48.602	56.061
15	22.307	24.996	30.578	35	46.059	49.802	57.342
16	23.542	26.296	32.000	36	47.212	50.998	58.619
17	24.769	27.587	33.409	37	48.363	52.192	59.893
18	25.989	28.869	34.805	38	49.513	53.384	61.162
19	27.204	30.144	36.191	39	50.660	54.572	62.428
20	28.412	31.410	37.566	40	51.805	55.758	63.691

Table D.1: 90%, 95% and 99% of the χ^2 distribution with K degrees of freedom

Bibliography

Ajzen, I. (1991). The theory of planned behavior, *Organizational Behavior and Human Decision Processes* **50**: 179–211.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Second international symposium on information theory*, Vol. 1, Springer Verlag, pp. 267–281.

Amemiya, T. (1975). Qualitative response models, *Ann. Econ. Social Measurement* **4**: 363–372.

Amemiya, T. (1980). Selection of regressors, *International Economic Review* **21**(2): 331–354.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators, *Journal of the Royal Statistical Society. Series B (Methodological)* **32**(2): 283–301.
URL: *http://www.jstor.org/stable/2984535*

Anderson, S. P., de Palma, A. and Thisse, J.-F. (1992). *Discrete Choice Theory of Product Differentiation*, MIT Press.

Ariely, D. (2010). *Predictably Irrational, Revised and Expanded Edition: the Hidden Forces That Shape Our Decisions*, Harper Perennial.

Armijo, L. (1966). Minimization of functions having continuous partial derivatives, *Pacific J. Math* **16**: 1–3.

Atherton, T. and Ben-Akiva, M. (1976). Transferability and updating of disaggregate travel demand models, *Transportation Research Record* **610**: 12–18.

Atkinson, K. E. (1989). *An Introduction to Numerical Analysis*, 2nd edn, John Wiley & Sons, Ltd., New York.

Axhausen, K., Koenig, A., Abay, G., Bates, J. J. and Bierlaire, M. (2004). Swiss value of travel time savings, paper presented at the European Transportation Conference.

Babbie, E. R. (2009). *The Practice of Social Research*, 12th edition edn, Wadsworth Publishing.

Becker, G. (1965). A theory of the allocation of time, *Econ. J.* **75**: 493–517.

Ben-Akiva, M. (1977). Choice models with simple choice set generating processes, *Working paper*, Dept. of Civil Engineering, MIT, Cambridge, MA.

Ben-Akiva, M. and Boccara, B. (1995). Discrete choice models with latent choice sets, *International Journal of Research in Marketing* **12**(1): 9 – 24. Consideration sets.
URL: *http://www.sciencedirect.com/science/article/B6V8R-40SFMGS-10/2/f781bc9df7013351f04552eeab3ff802*

Ben-Akiva, M., Bolduc, D. and Bradley, M. (1993). Estimation of travel model choice models with randomly distributed values of time, *Transportation Research Record* **1413**: 88–97.

Ben-Akiva, M. E. (1973). *Structure of Passenger Travel Demand Models*, PhD thesis, Massachusetts Institute of Technology.

Ben-Akiva, M. E. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, Ma.

Ben-Akiva, M. and Gershenfeld, S. (1998). Multi-featured products and services: analysing pricing and bundling strategies, *Journal of Forecasting* **17**(3-4): 175–196.

Ben-Akiva, M., Gunn, H. and Pol, H. (1983). Expansion of data from mixed random and choice based survey designs. Prepared for presentation at the conference on New Survey Methods in Transports, Sydney, Australia.

Ben-Akiva, M., Gunn, H. and Silman, L. (1984). Disaggregate trip distribution models, *Proceedings of the Japanese Society of Civil Engineers* .

Ben-Akiva, M. and Lerman, S. (1979). Disaggregate travel and mobility choice models and measures of accessibility, *in* D. Hensher and P. Stopher (eds), *Behavioural Travel Modelling*, Croom Helm, London.

Ben-Akiva, M. and Watanatada, T. (1981). Application of a continuous choice logit model, *in* C. Manski and D. McFadden (eds), *Structural Analysis of Discrete Data With Econometric Applications*, MIT Press, Cambridge, Mass.

Bentler, P. M. (1980). Multivariate analysis with latent variables, *Annual review of psychology* **31**: 419–456.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*, Springer.

Berkson, J. (1944). Application of the logistic function to bio-assay, *Journal of the American Statistical Association* **39**(227): pp. 357–365.
URL: *http://www.jstor.org/stable/2280041*

Berkson, J. (1953). A statistically precise and relatively simple method of estimating the bioassay with quantal response, based on the logistic function., *J. Amer. Stat. Assn.* **48**: 565–599.

Berndt, E. K., Hall, B. H., Hall, R. E. and Hausman, J. A. (1974). Estimation and inference in nonlinear structural models, *Annals of Economic and Social Measurement* **3/4**: 653–665.

Bertsekas, D. P. (1999). *Nonlinear Programming*, 2nd edn, Athena Scientific, Belmont.

Bhat, C. (1998). Accommodating flexible substitution patterns in multi-dimensional choice modeling: Formulation and application to travel mode and departure time choice, *Transportation Research Part B: Methodological* **32**(7): 455–466.

Bhat, C. and Sener, I. (2009). A copula-based closed-form binary logit choice model for accommodating spatial correlation across observational units, *Journal of Geographical Systems* **11**(3): 243–272–272.
URL: *http://dx.doi.org/10.1007/s10109-009-0077-9*

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A. and Sudman, S. (eds) (2004). *Measurement Errors in Surveys*, John Wiley & Sons, Ltd., Hoboken, NJ, USA.

Bierlaire, M. (2003). BIOGEME: a free package for the estimation of discrete choice models, *Proceedings of the Swiss Transport Research Conference*, Ascona, Switzerland.

Bierlaire, M. (2006a). *Introduction à l'optimisation différentiable*, Presses Polytechniques et Universitaires Romandes, Lausanne.

Bierlaire, M. (2006b). A theoretical analysis of the cross-nested logit model, *Annals of operations research* **144**(1): 287–300.

Bierlaire, M. (2015). *Optimization: principles and algorithms*, EPFL Press.

Bierlaire, M., Bolduc, D. and McFadden, D. (2008). The estimation of generalized extreme value models from choice-based samples, *Transportation Research Part B* **42**(4): 381–394.

Bierlaire, M., Hurtubia, R. and Flötteröd, G. (2010). An analysis of the implicit choice set generation using the constrained multinomial logit model, *Transportation Research Record* **2175**: 92–97.

Bierlaire, M., Lotan, T. and Toint, P. L. (1997). On the overspecification of multinomial and nested logit models due to alternative specific constants, *Transportation Science* **31**(4): 363–371.

Blackwell, R. D., Miniard, P. W. and Engel, J. F. (2005). *Consumer Behavior*, 10th edn, South-Western College Pub.

Bock, R. D. (1968). Estimating multinomial response relations, *Contributions to Statistics and Probability: Essays in Memory of S. N. Roy*, University of North Carolina Press, Chapel Hill.

Bock, R. and Jones, J. (1968). *The measurement and prediction of judgment and choice*, Holden-Day, Oxford, England.

Boersch-Supan, A. and Hajivassiliou, V. A. (1993). Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models, *Journal of Econometrics* **58**(3): 347 – 368.

Bolduc, D. (1999). A practical technique to estimate multinomial probit models in transportation, *Transportation Research Part B: Methodological* **33**(1): 63 – 79.
URL: *http://www.sciencedirect.com/science/article/B6V99-3V7JR3Y-4/2/7b1b8d3abcabc160c667f1ca7c30ab77*

Bolduc, D. and Ben-Akiva, M. (1991). A multinomial probit formulation for large choice sets, *Proceedings of the Sixth International Conference on Travel Behavior*, pp. 243–258.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society. Series B (Methodological)* **26**(2): 211–252.
URL: *http://www.jstor.org/stable/2984418*

Boyd, J. and Mellman, J. (1980). The effect of fuel economy standards on the u.s. automotive market: A hedonic demand analysis, *Transportation Research Part A: Policy and Practice* pp. 367–378.

Brenner, J. (2013). Pew internet: Mobile.
URL: *http://pewinternet.org/Commentary/2012/February/Pew-Internet-Mobile.aspx*

Bresnahan, T. F. and Reiss, P. C. (1991). Empirical models of discrete games, *Journal of Econometrics* **48**(1–2): 57–81.
URL: *http://www.sciencedirect.com/science/article/pii/0304407691900329*

Brey, R. and Walker, J. (2011). Latent temporal preferences: An application to airline travel, *Transportation Research Part A: Policy and Practice* **45**(9): 880–895.

Brog, W. and Meyburg, A. H. (1981). Consideration of nonresponse effects in large-scale mobility surveys, *Transportation Research Record* **807**: 39–46.

Brownstone, D., Bunch, D. and Train, K. (2000). Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles, *Transportation Research Part B* **34**: 315–338.

Brownstone, D. and Small, K. A. (1989). Efficient estimation of nested logit models, *Journal of Business and Economic Statistics* **7**(1): 67–74.

Cambridge Systematics Europe (1984). Estimation and application of disaggregate models of mode and destination choice, Draft report prepared for Direction des Etudes Générales, Régie Autonome des Transport Parisien, Paris, France.

Cardell, S. and Dunbar, F. (1980). Measuring the societal impacts of automobile downsizing, *Transportation Research Part A: Policy and Practice* pp. 423–434.

Carroll, J. D. and Green, P. E. (1995). Guest editorial: Psychometric methods in marketing research: Part i, conjoint analysis, *Journal of Marketing Research* **32**(4): 385–391.
URL: *http://www.jstor.org/stable/3152174*

Carson, R. and Louviere, J. (2011). A common nomenclature for stated preference elicitation approaches, *Environmental Resource Economics* **49**(4): 539–559.

Cascetta, E. and Papola, A. (2001). Random utility models with implicit availability/perception of choice alternatives for the simulation of travel demand, *Transportation Research Part C: Emerging Technologies* **9**(4): 249 – 263.
URL: *http://www.sciencedirect.com/science/article/B6VGJ-42WP4TY-2/2/2dfd858bbbf68374e4a4176fb5d67947*

Cherchi, E. and Ortúzar, J. (2006). On fitting mode specific constants in the presence of new options in RP/SP models, *Transportation Research Part A: Policy and Practice* **40**(1): 1–18.

Cho (2012). *Ngene 1.1.1 User Manual & Reference Guide*.

Chu, W. and Anderson, E. M. (1992). Capturing ordinal properties of categorical dependent variables: a review with application to modes of foreign entry, *International Journal of Research in Marketing* **9**(2): 149–160.

Clark, C. E. (1961). The greatest of a finite set of random variables, *Operations Research* **9**: 85–91.

Cochran, M. (1977). *Sampling technique*, 3rd edn, Wiley, New-York.

Converse, J. and Presser, S. (1986). *Survey Questions, Handcrafting the Standardized Questionnaire*, Vol. 07, Sage Publications, Inc., Thousand Oaks, Ca.

Cosslett, S. (1981a). Efficient estimation of discrete choice models, *in* C. Manski and D. McFadden (eds), *Structural analysis of discrete data with econometric applications*, MIT Press, Cambridge, Ma.

Cosslett, S. (1981b). Maximum likelihood estimation for choice-based samples, *Econometrica* **49**: 1289–1316.

Cosslett, S. (1983). Distribution-free maximum likelihood estimator of the binary choice model, *Econometrica* **51**: 765–782.

Cox, D. R. (1970). *Analysis of Binary Data*, Methuen, London.

Daganzo, C. F. (1980). Optimal sampling strategies for statistical models with discrete dependent variables, *Transportation Science* **14**.

Daganzo, C. F. and Kusnic, M. (1993). Two properties of the nested logit model, *Transportation Science* **27**(4): 395–400.

Dagsvik, J. K. and Karlström, A. (2005). Compensating variation and hicksian choice probabilities in random utility models that are nonlinear in income, *Review of Economic Studies* **72**(1): 57–76.

Daly, A. (1982). Estimating choice models containing attraction variables, *Transportation Research Part B* **16**(1): 5–15.

Daly, A. (1987). Estimating "tree" logit models, *Transportation Research Part B* **21**(4): 251–268.

Daly, A. and Bierlaire, M. (2006). A general and operational representation of generalised extreme value models, *Transportation Research Part B* **40**(4): 285–305.

Daly, A. J. and Van Zwam, H. H. P. (1981). Travel demand models for the zuidvleugel study, *Proc. PTRC Summer Annual Meeting*.

Daly, A. and Rohr, C. (1998). Forecasting demand for new travel alternatives, *in* T. Gärling, T. Laitila and K. Westin (eds), *Theoretical Foundation for Travel Choice Modelling*, Pergamon.

Davidson, R. and MacKinnon, J. (1981). Several tests for model specification in the presence of alternative hypotheses, *Econometrica: Journal of the Econometric Society* pp. 781–793.

De Groot, M. (1970). *Optimal statistical decisions*, McGraw-Hill, New-York.

de Palma, A. and Picard, N. (2005). Route choice decision under travel time uncertainty, *Transportation Research Part A* **39**(4): 295–324.

Debreu, G. (1960). Review of r. luce: Individual choice behavior, *American Economic Review* .

Deming, W. (1960). *Sample design in business research*, John Wiley, New-York.

Dennis, J. E. and Schnabel, R. B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations*, Society for Industrial and Applied Mathematics (SIAM).

Domencich, T. and McFadden, D. (1975). *Urban Travel Demand-A Behavioral Analysis*, North Holland, Amsterdam.

Donnelly, T. (1963). Algorithm 462: Bivariate normal distribution, *Communications of the ACM* **16**: 289–294.

Drezner, Z. and Wesolowsky, G. (1989). On the computation of the bivariate normal integral, *Journal of Statist. Comput. Simul.* **35**: 101–107.

Dubin, J. A. and McFadden, D. (1984). An econometric analysis of residential electricity appliance holdings and consumption, *Econometrica* **52**(2).

Dugundji, E. and Walker, J. (2005). Discrete choice with social and spatial network interdependencies, *Transportation Research Record* **1921**: 70–78.

Durbin, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities, *J. Roy. Stat. Soc. B* **15**: 262–269.

Electric Power Research Institute (1977). Methodology for predicting the demand for new electricity-using goods, *Final report. Project 488-1 EA-593*, Electric Power Research Institute, Palo Alto, Ca.

Farooq, B., Bierlaire, M., Hurtubia, R. and Flötteröd, G. (2013). Simulation based synthesis of population, *Transportation Research Part B: Methodological* **58**: 243–263.

Fink, A. (1995). *How to sample in surveys*, SAGE Publications.

Finney, D. (1971). *Probit Analysis*, 3rd edn, Cambridge University Press, Cambridge, England.

Fisher, R. (1926). The arrangement of field experiments, *Journal of the Ministry of Agriculture of Great Britain* **33**: 503–513.
URL: *http://digital.library.adelaide.edu.au/coll/special//fisher/48.pdf*

Fisher, R. (1935). *The Design of Experiments*, Oliver and Boyd, Edinburgh.

Fogg, B. J. (2009). A behavioral model for persuasive design. working paper, Stanford University.

Fosgerau, M. and Bierlaire, M. (2009). Discrete choice models with multiplicative error terms, *Transportation Research Part B: Methodological* **43**(5): 494–505.

Fosgerau, M., McFadden, D. and Bierlaire, M. (2013). Choice probability generating functions, *Journal of Choice Modelling* pp. 1–18.

Fowler, F. J. J. (1995). *Improving Survey Questions, Design and Evaluation*, Vol. 38 of *Applied Social Research Method Series*, Sage Publications, Thousand Oaks, Ca.

Fowler, J. (2008). *Case Study Research: Survey Research Methods,*, fourth edition edn, Sage Publications.

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum, *Ann. Soc. Polon. Math. Cracovie* **6**: 93–116.

Fundenberg, D. and Tirole, J. (1991). *Game Theory*, MIT Press.
URL: *http://books.google.ch/books?id=pFPHKwXro3QC*

Garrow, L. A. (2010). *Discrete choice modelling and air travel demand*, Ashgate, Surrey, England.

Garrow, L., Jones, S. and Parker, R. (2006). How much airline customers are willing to pay: an analysis of price sensitivity in online distribution channels, *Journal of Revenue and Pricing Management* **5**(4): 271–290.

Gaudry, M. and Dagenais, M. (1979). The dogit model, *Transportation Research Part B: Methodological* **13**: 105–111.

Genz, A. (2004). Numerical computation of rectangular bivariate and trivariate normal and t probabilities, *Statistics and Computing* **14**: 151–160.

Glerum, A., Stankovikj, L., Thémans, M. and Bierlaire, M. (forthcoming). Forecasting the demand for electric vehicles: accounting for attitudes and perceptions, *Transportation Science* . Accepted for publication.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association* **102**(477): 359–378.

Gönül, F. and Srinivasan, K. (1993). Modeling multiple sources of heterogeneity in multinomial logit models: Methodological and managerial issues, *Marketing Science* **12**(3): 213–229.

Gopinath, D. (1995). *Modeling Heterogeneity in Discrete Choice Processes: Application to Travel Demand*, PhD thesis, Department of Civil and Envronmental Engineering, Massachusetts Institute of Technology.

Green, P., Carroll, D., Greem, P. and Goldberg, S. (1981). A general approach to product design and optimization via conjoint analysis, *Journal of Marketing* **45**: 17–37.

Green, P., Krieger, A. and Wind, Y. (2001). Thirty years of conjoint analysis: reflections and prospects, *Interfaces* **31**: S56–S73.

Greene, W. and Hensher, D. (2003). A latent class model for discrete choice analysis: Contrasts with mixed logit, *Transportation Research Part B: Methodological* **37**: 681–698.

Greene, W., Hensher, D. and Rose, J. (2006). Accounting for heterogeneity in the variance of unobserved effects in mixed logit models, *Transportation Research Part B: Methodological* **40**: 75–92.

Gumbel, E. J. (1958). *Statistics of Extremes*, Columbia University Press, New York.

Gurland, J., Lee, I. and Dahm, P. A. (1960). Polychotomous quantal response in biological assay, *Biometrics* **16**(3): 382–398.

Hahn, G. and Shapiro, S. (1966). A catalog and computer program for the design and analysis of orthogonal symmetric and asymmetric fractional factorial experiments, *Technical Report 66-C 165*, General Electric Research andDevelopment Center, Schenectady, NY, USA.

Hall, P. (1980). *Search Behavior in Urban Housing Markets*, PhD thesis, Department of Civil Engineering, MIT, Cambridge, Mass.

Halton, J. H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals, *Numerische Mathematik* **2**(1): 84–90.
URL: *http://dx.doi.org/10.1007/BF01386213*

Hammersley, J. and Handscomb, D. (1965). *Monte Carlo Methods*, Methuen, London, UK.

Hanemann, W. (1984). Welfare evaluations in contingent valuation experiments with discrete responses, *American journal of agricultural economics* **66**(3): 332–341.

Harris, A. J. and Tanner, J. C. (1974). Transport demand models based on personal characteristics, *in* D. J. Buckley (ed.), *Transportation and Traffic Theory*, Elsevier, Amsterdam.

Hausman, J. and McFadden, D. (1984). Specification tests for the multinomial logit model, *Econometrica: Journal of the Econometric Society* pp. 1219–1240.

Hausman, J. and Wise, D. (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences, *Econometrica* **46**: 403–426.

Hendrickson, C. and Sheffi, Y. (1978). A disaggregate model of trip generation by elderly individuals, 1.202 term paper. Department of Civil Engineering, MIT, Cambridge, Mass.

Hensher, D. and Greene, W. (2003). The mixed logit model: the state of practice, *Transportation* **30**(2): 133–176.

Hess, S., Bierlaire, M. and Polak, J. (2005a). Estimation of value of travel-time savings using mixed logit models, *Transportation Research Part A* **39**(3): 221–236.

Hess, S., Bierlaire, M. and Polak, J. (2005b). Estimation of value of travel-time savings using mixed logit models, *Transportation Research Part A: Policy and Practice* **39**(2-3): 221–236.

Hsieh, D., Manski, C. and McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations, *Journal of the American Statistical Association* **80**(391): 651–662.

Imbens, G. W. (1992). An efficient method of moments estimator for discrete choice models with choice-based sampling, *Econometrica* **60**(5): 1187–1214.
URL: *http://www.jstor.org/stable/2951544*

Inc., S. I. (2010). Jmp® 9 design of experiments guide, *Technical report*, SAS Institute Inc., Cary, NC.

Jenkinson, A. F. (1955). Frequency distribution of the anjual maximum (or minimum) values of meteorological elements, *Quarterly journal of the Royal Meteorological Society* **81**: 158–171.

Joe, H. (1997). *Multivariate models and dependence concepts*, Chapman & Hall.
URL: *http://books.google.com/books?id=iJbRZL2QzMAC*

Johnson, N. and Kotz, S. (1970). *Continuous univariate distributions*, Vol. 1 & 2, Wiley, New-York.

Johnson, R. (1974). Trade-off analysis of consumer values, *Journal of Marketing Research* **11**: 121–127.

Johnston, J. (1972). *Econometric methods*, 2 edn, McGraw-Hill.

Joreskog, K. G. (1973). A general method for estimating a linear structural equation system, *in* A. S. Goldberger and O. D. Duncan (eds), *Structural models in the social sciences*, Academic Press.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality, *American Psychologist* **58**(9): 697–720.

Kahneman, D. (2011). *Thinking, Fast and Slow*, Farrar, Straus and Giroux.

Kahneman, D., Knetsch, J. L. and Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias, *The journal of economic perspectives* **5**(1): 193–206.

Kahneman, D. and Tversky, A. (2000). *Choices, Values, and Frames*, Cambridge University Press.

Kamakura, W. and Russell, G. (1989). A probabilistic choice model for market segmentation and elasticity structure, *Journal of Marketing Research* **25**(379-390).

Kauermann, G. and Carroll, R. (2001). A note on the efficiency of sandwich covariance matrix estimation, *Journal of the American Statistical Association* **96**(456).

Keesling, J. W. (1972). *Maximum likelihood approaches to causal analysis*, PhD thesis, University of Chicago.

Kitamura, R., Kostyniuk, L. and Ting, K. L. (1979). Aggregation in spatial choice modelling, *Transportation Science* **13**: 325–342.

Koenig, A., Abay, G. and Axhausen, K. (2003). Time is money: the valuation of travel time savings in switzerland, *Proceedings of the 3rd Swiss Transportation Research Conference*. `http://www.strc.ch/Paper/Koenig.pdf`.

Koenig, A., Abay, G. and Axhausen, K. (2004). Zeitkostenansaetze im personenverkehr, *Schriftenreihe 1065*. final report for SVI 2001/534.

Koppelman, F. (1975). *Travel Prediction with Models of Individualistic Choice Behavior*, PhD thesis, Department of Civil Engineering, MIT, Cambridge, Mass.

Koppelman, F. and Hauser, J. (1978). Destination choice behavior for non-grocery shopping trips, *Transportation Research Record* (673): 157–165.

Kotz, S., Balakrishnan, N. and Johnson, N. L. (2000). *Continuous multivariate distributions*, Vol. 1: Models and Applications of *Wiley Series in Probability and Statistics*, 2nd edn, John Wiley and Sons.

Kotz, S. and Nadarajah, S. (2001). *Extreme Value Distributions: Theory and Applications*, Imperial College Press.

Kozel, V. and Swait, J. (1982). *Maceio Travel Demand Model System Calibration Results*, Vol. 2 of *Studies of Urban Travel Behavior and Policy Analysis in Maceio*, Center for Transportation Studies, MIT, Cambridge, Ma.

Kuhfeld, W. F. (2005). Marketing research methods in sas: Experimental design, choice, conjoint, and graphical techniques, *SAS 9.1 Edition TS-722*, SAS Institute Inc., Cary, NC, USA.

Lancaster, K. (1966). A new approach to consumer theory, *J. Pol. Econ.* **74**: 132–157.

Lerman, S. (1975). *A Disaggregate Behavioral Model of Urban Mobility Decisions*, PhD thesis, Department of Civil Engineering, MIT, Cambridge, Mass.

Lerman, S. and Gonzalez, S. (1980). Poisson regression analysis under alternate sampling strategies, *Transportation Science* **14**(4): 346–364.

Lerman, S. and Mahmassani, H. S. (1985). The econometrics of search, *Environment and Planning A* **17**(8): 1009–1024.

Lisco, T. (1967). *The Value ofCommuter's Travel Time: A Study in Urban Transportation*, PhD thesis, Department of Economics, University of Chicago, Chicago, Ill.

Lohr, S. L. (1999). *Sampling: Design and Analysis*, Duxbury Press.

Louviere, J. J., Hensher, D. A. and Swait, J. D. (2000). *Stated choice methods: analysis and applications*, Cambridge University Press.

Louviere, J. and Woodworth, G. (1983). Design and analysis of simulated consumer choice or allocation experiments, *Journal of Marketing Research* **20**(4): 350–367.

Lovelace, R. and Ballas, D. (2013). 'truncate, replicate, sample': A method for creating integer weights for spatial microsimulation, *Computers, Environment and Urban Systems* **41**(0): 1 – 11.
URL: *http://www.sciencedirect.com/science/article/pii/S0198971513000240*

Luce, R. (1959). *Individual Choice Behavior: A Theoretical Analysis*, Wiley, New York.

Luce, R. and Suppes, P. (1965). Preference, utility and subjective probability, *in* R. Luce, R. Bush and E. Galanter (eds), *Handbook of Mathematical Psychology*, Vol. 3, Wiley, New York.

Luce, R. and Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement, *Journal of Mathematical Psychology* **1**(1): 1 – 27.
URL: *http://www.sciencedirect.com/science/article/B6WK3-4DTKF6W-B5/2/de74cad0c9f558e577f99ebc294c3c09*

Manski, C. (1975). Maximum score estimation of the stochastic utility model of choice, *J. Econometrics* **3**: 205–228.

Manski, C. (1977). The structure of random utility models, *Theory and Decision* **8**: 229–254.

Manski, C. and Lerman, S. (1977). The estimation of choice probabilities from choice-based samples, *Econometrica* **45**(8): 1977–1988.

Manski, C. and McFadden, D. (1981a). Alternative estimators and sample designs for discrete choice analysis, *in* C. Manski and D. McFadden (eds), *Structural analysis of discrete data with econometric applications*, MIT Press, Cambridge, Ma.

Manski, C. and McFadden, D. (1981b). Alternative estimators and sample designs for discrete choice analysis, *in* C. Manski and D. McFadden (eds), *Structural analysis of discrete data with econometric application*, MIT Press, Cambridge, Mass.

Marschak, J. (1960). Binary choice constraints on random utility indicators, *in* K. Arrow (ed.), *Stanford Symposium on Mathematical Methods in the Social Sciences*, Stanford University Press, Stanford, Calif.

Martinez, F., Aguila, F. and Hurtubia, R. (2009). The constrained multinomial logit: A semi-compensatory choice model, *Transportation Research Part B: Methodological* **43**(3): 365 – 377.

URL: *http://www.sciencedirect.com/science/article/B6V99-4T4JDM1-1/2/2556f32b64ea0e0d53b80b8d00a5a3e6*

McConnell, K. E. (1995). Consumer surplus from discrete choice models, *Journal of Environmental Economics and Management* **29**(3): 263 – 270.
URL: *http://www.sciencedirect.com/science/article/B6WJ6-45S92WM-N/2/a252034061e51b518591926a6914d672*

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior, *in* P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press, New York, pp. 105–142.

McFadden, D. (1978). Modelling the choice of residential location, *in* A. Karlquist *et al.* (ed.), *Spatial interaction theory and residential location*, North-Holland, Amsterdam, pp. 75–96.

McFadden, D. (1981). Econometric models of probabilistic choice, *in* C. Manski and D. McFadden (eds), *Structural Analysis of Discrete Data With Econometric Applications*, MIT Press.

McFadden, D. (1986). The choice theory approach to market research, *Marketing Science* **5**(4): 275–297.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration, *Econometrica* **57**(5): 995–1026.
URL: *http://www.jstor.org/stable/1913621*

McFadden, D. (1999). Rationality for economists?, *Journal of Risk and Uncertainty* **19**(1-3): 73–105.

McFadden, D. (2001). Economic choices, *The American Economic Review* **91**(3): 351–378.

McFadden, D. (2013). The new science of pleasure, *Working paper*, National Bureau of Economic Research.

McFadden, D. and Train, K. (2000). Mixed mnl models of discrete response, *Journal of Applied Econometrics* **15**: 447–470.

McFadden, D., Train, K. and Tye, W. (1977). An application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model, *Transportation Research Record* (637).

McKelvey, R. D. and Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables, *Journal of Mathematical Sociology* **4**: 103–120.

Montgomery, D. C. (2008). *Design and analysis of experiments*, John Wiley and Sons.

Morichi, S., Ishida, H. and Yai., T. (1984). Comparisons of various utility functions for behavioral travel demand models, *Proceedings of WCTR*, SNV, Hamburg.

Morikawa, T. (1996). A hybrid probabilistic choice set model with compensatory and noncompensatory choice rules, *Proceedings of the 7th World Conference on Transport Research*, Vol. 1, pp. 317–325.

Morikawa, T., Ben-Akiva, M. and McFadden, D. (2002). Discrete choice models incorporating revealed preferences and psychometric data, *in* P. Franses and A. Montgomery (eds), *Econometric models in marketing*, Elsevier Science, pp. 29–55.

Moscati, I. (2007). Early experiments in consumer demand theory: 1930-1970, *History of Policital Economy* **39**(359-401).

Muth, R. (1966). Household production and consumer demand functions, *Econometrica* **34**: 699–708.

Natarajan, R., McCulloch, C. E. and Kiefer, N. M. (2000). A monte carlo EM method for estimating multinomial probit models, *Computational Statistics and Data Analysis* **34**(1): 33 – 50.
URL: *http://www.sciencedirect.com/science/article/B6V8V-40SFH08-3/2/8dc1d6352644a7f2a9a7f36b15ad1aec*

Nelsen, R. B. (2006). *An introduction to copulas*, Springer.
URL: *http://books.google.com/books?id=B3ONT5rBv0wC*

Netherlands Ministry of Transport (1977). SIGMO final reports, Vols. 1-4. Projectbureau IVVS, The Hague, The Netherlands.

Neuburger, H. (1971). User benefit in the evaluation of transport and land use plans, *Journal of Transport Economics and Policy* **5**(1): 52–75.

Newman, J. P. (2008). Normalization of network generalized extreme value models, *Technical report*, Department of Civil Engineering, Northwestern University.

Nicholson, W. and Snyder, C. M. (2007). *Microeconomic Theory: Basic Principles and Extensions*, 10th edn, South-Western College Pub.

Nikoloulopoulos, A. K. and Karlis, D. (2008). Multivariate logit copula model with an application to dental data, *Statistics in Medicine* **27**(30): 6393–6406.
URL: *http://dx.doi.org/10.1002/sim.3449*

Nocedal, J. and Wright, S. J. (1999). *Numerical optimization*, Operations Research, Springer Verlag, New-York.

Payne, J., Bettman, J. and Johnson, E. (1992). Behavioral decision research: a constructive process perspective, *Annual review of psychology* **43**: 87–131.

Perugini, M. and Bagozzi, R. P. (2001). The role of desires and anticipated emotions in goal-direction behaviors: Broadening and deepenng the theory of planned behaviour, *British Journal of Social Psychology* **40**(1): 79–98.

Pickands, J. (1981). Multivariate extreme value distributions, *Proceedings 43rd Session International Statistical Institute*, pp. 859–878.

Pindyck, R. and Rubinfeld, D. (2008). *Microeconomics*, 7th edn, Prentice Hall.

Prelec, D. (1991). Values and principles: some limitations on traditional economic analysis, *in* A. Etzioni and P. Lawrence (eds), *Perspectives on socioeconomics*, M.E. Sharpe.

Prochaska, J. O. and Velicer, W. F. (1997). The transtheoretical model of health behavior change, *American journal of health promotion* **12**(1): 38–48.

Quandt, R. E. (1956). A probabilistic theory of consumer behavior, *The Quarterly Journal of Economics* **70**(4): 507–536.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, Wiley, New-York.

Rea, L. M. and Parker, R. A. (2005). *Designing and conducting survey research: a comprehensive guide*, 3rd edition edn, Jossey-Bass.

Reid, F. (1978). Minimizing error in aggregate predictions from disaggregate models, *Trans. Research Record* **673**: 59–65.

Revelt, D. and Train, K. (1998). Mixed logit with repeated choices: Households' choice of appliance efficiency level, *Review of Economics and Statistics* **80**(4): 647–657.

Richards, M. and Ben-Akiva, M. (1975). A disaggregate travel demand model, D.C. Heath, Lexington, Ma.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*, Springer texts in statistics, 2 edn, Springer.

Ross, S. (2012). *Simulation*, fifth edition edn, Academic Press.
URL: *http://books.google.ch/books?id=sZjDT6MQGF4C*

Savage, L. J. (1971). Elicitation of personal probabilities and expectations, *Journal of the American Statistical Association* **66**(336): 783–801.
URL: *http://www.jstor.org/stable/2284229*

Schervish, M. J. (1989). A general method for comparing probability assessor, *The Annals of Statistics* **17**: 1856–1879.

Schnabel, R. B. and Eskow, E. (1999). A revised modified Cholesky factorization, *SIAM Journal on Optimization* **9**: 1135–1148.

Sheffi, Y. (1979). Estimating choice probabilities among nested alternatives, *Transportation Research Part B* **13**: 113–205.

Siddarth, S., Bucklin, R. E. and Morrison, D. G. (1995). Making the cut: Modeling and analyzing choice set restriction in scanner panel data, *Journal of Marketing Research* **32**(3): 255–266.
URL: *http://www.jstor.org/stable/3151979*

Silman, L. (1980). Disaggregate travel demand models for short-term forecasting, *Information paper 81*, Institute of Transportation Planning and Research, Tel Aviv, Israel.

Simon, H. (1957). *Models of Man*, Wiley, New York.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris* **8**: 229–231.

Slovic, P., Fischoff, B. and Lichtenstein, L. (1977). Behavioral decision theory, *Ann. Rev. Psychology* .

Small, K. (1981). Ordered logit: a discrete choice model with proximate covariance among alternatives, *Working paper*, Department of Economics, Princeton University, Princeton, NY.

Small, K. A. and Rosen, H. S. (1981). Applied welfare economics with discrete choice models, *Econometrica* **49**(1): 105–30.
URL: *http://ideas.repec.org/a/ecm/emetrp/v49y1981i1p105-30.html*

Stern, S. (1992). A method for smoothing simulated moments of discrete probabilities in multinomial probit models, *Econometrica* **60**(4): 943–952.
URL: *http://www.jstor.org/stable/2951573*

Strotz, R. H. (1957). The empirical implications of a utility tree, *Econmerica* **25**(2): 269–280.

Svenson, O. (1979). Process descriptions of decision-making, *Org. Behavior Human Performance* **23**: 86–112.

Swait, J. (2001). Choice set generation within the generalized extreme value family of discrete choice models, *Transportation Research Part B: Methodological* **35**(7): 643 – 666.
URL: *http://www.sciencedirect.com/science/article/B6V99-437XR4M-2/2/8d02ca17d544e5672e08e02158288d7d*

Swait, J. and Ben-Akiva, M. (1987a). Empirical test of a constrained choice discrete model: model choice in sao paulo, brazil, *Transportation Research Part B* **21**(2): 103–115.

Swait, J. and Ben-Akiva, M. (1987b). Incorporating random constraints in discrete models of choice set generation, *Transportation Research Part B* **21**(2): 91–102.

Tarem, Z. (1982). *Evaluation of a sampling optimization method for discrete choice models*, Master's thesis, Department of Civil Engineering, MIT, Cambridge, Ma.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving Decisions About Health, Wealth, and Happiness*, Penguin Books.

Theil, H. (1969). A multinomial extension of the linear logit model, *Int. Econ. Rev.* **10**: 251–259.

Theil, H. (1971). *Principles of econometrics*, John Wiley and Sons.

Thurstone, L. (1927). A law of comparative judgement, *Psychological Rev.* **34**: 273–286.

Thurstone, L. (1931). The indifference function, *Journal of Social Psychology* **2**: 139–167.

Toledo, T. (2003). Integrating driver behavior modeling.

Touring Club Suisse (n.d.). Exemple de frais pour un véhicule. **URL:** *http://www.tcs.ch/fr/auto-mobilite/couts-de-la-voiture/exemple.php*

Train, K. (2003). *Discrete Choice Methods with Simulation*, Cambridge University Press. http://emlab.berkeley.edu/books/choice.html.

Train, K., McFadden, D. and Ben-Akiva, M. (1987). The demand for local telephone service: a fully discrete model of residential calling patterns and service choices, *The RAND Journal of Economics* **18**(1): 109–123.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*, Springer Series in Statistics, Springer.

Tversky, A. (1969). Intrasitivity of preferences, *Psychological Review* **76**(1): 31–48.

Tversky, A. (1972). Elimination by aspects: A theory of choice, *Psychological Rev.* **79**: 281–299.

Varian, H. R. (2009). *Intermediate Microeconomics. A Modern Approach*, eighth edn, W. W. Norton & Company.

Vichiensan, V., Miyamoto, K. and Tokunaga, Y. (2005). Mixed logit modeling framework with structuralized spatial effects: A test of applicability with area unit systems in location choice analysis, *Journal of the Eastern Asian Society for Transportation Studies* **6**: 3789–3802.

Vij, A. and Walker, J. (2014). Preference endogeneity in discrete choice models, *Transportation Research Part B: Methodological* **64**: 90–105.

Walker, J. and Ben-Akiva, M. (2011). Advances in discrete choice: Mixtures models, *in* E. Q. A. de Palma, R. Lindsey and R. Vickerman (eds), *Handbook in Transport Economics*, pp. 160–187.

Walker, J., Ben-Akiva, M. and Bolduc, D. (2007). Identification of parameters in normal error component logit-mixture (neclm) models, *Journal of Applied Econometrics* **22**: 1095–1125.

Walker, J. and Li, J. (2007). Latent lifestyle preferences and household location decisions, *Journal of Geographical Systems* **9**(1): 77–101.

Watanatada, T. and Ben-Akiva, M. (1977). Development of an aggregate model of urbanized area travel behavior, Final report prepared for US-DOT, Washington, D.C.

Watanatada, T. and Ben-Akiva, M. (1979). Forecasting urban travel demand for quick policy analysis with disaggregate choice models: A Monte Carlo simulation approach, *Transportation Research Part A* **13**: 241–248.

Wen, C.-H. and Koppelman, F. S. (2001). The generalized nested logit model, *Transportation Research Part B* **35**(7): 627–641.

Wetherill, G. B. and Glazebrook, K. D. (1986). *Sequential methods in statistics*, Chapman and Hall.

White, H. (1980). Nonlinear regression on cross-section data, *Econometrica* **48**: 721–746.

White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica* **50**: 1–25.

Wiley, D. E. (1973). The identification problem for structural equation models with unmeasured variables, *in* A. S. Goldberger and O. D. Duncan (eds), *Structural models in the social sciences*, Academic Press.

Williams, H. (1977). On the formation of travel demand models and economic measures of user benefit, *Environment and Planning* **9A**: 285–344.

Wolfe, P. (1969). Convergence conditions for ascent methods, *SIAM review* **11**: 226–235.

Wolfe, P. (1971). Convergence conditions for ascent methods II: some corrections, *SIAM review* **13**: 185–188.

Index