Choice modelling

An overview of theory and development in individual choice behaviour modelling

L.T. Wittink
BMI Paper
August 2011
Supervised by Alwin Haensel

Table of contents

Introduction **4**

Choice Theory **6**

Framework **6**

Rational Behaviour **8**

Discrete and Probabilistic Choice Theory **8**

Discrete Choice Theory **8**

Probabilistic Choice Theory **9**

Utility Theory **10**

Cardinal Utility **10**

Ordinal Utility **10**

Constant Utility **10**

Random Utility **11**

Expected Utility **11**

Stated and Revealed Preference **12**

Stated Preference **12**

Revealed Preference **13**

Exogenous-, Locational- and Utility-Based Choice Models **13**

Exogenous-based Models **13**

Locational-based Models **14**

Binary Choice Models **14**

Logit and Probit **18**

Linear Probability Models **18**

Probit **18**

Logit **19**

Estimation **19**

Alternative estimation models **20**

Multinomial Logit **20**

Multinomial choice **21**

Multinomial logit **21**

Estimation **22**

Nested Logit **23**
Multidimensional choice sets **24**
Nested Logit **24**
Estimation of Nested Logit **26**
Higher level Nested Logit and expansion on the Nested Logit Model **27**
Cross-Nested Logit **27**
Estimation of Cross-Nested Logit **28**
Mixed Logit **29**
Estimation **30**
Repeated Choice **31**
Latent Class Logit **31**
Estimation **33**
Variations: different choice models **36**
The Generalized Extreme Value Model **36**
Joint Logit **36**
Multinomial Probit **37**
Mixed Probit **37**
Further Research **37**
Summary **37**
Acknowledgement **38**
References **38**

“As far as the laws of math refer to reality, they are not certain; and as far as they are certain, they do not refer to reality.”
Albert Einstein (1879 – 1955)

Introduction

To some degree, all decisions or even most of the actions we take in life, involve choice. When we go to the supermarket, we have to decide on the way to travel. When we are at the supermarket we have to choose from a selection of vegetables for example. When we are at home we have to decide on what to cook etcetera. A day in our life is full of sequences of choices we have to make. But not just us: our entire fellow society goes through similar thought processes.

The fact that a whole population goes through such processes makes it worth investigating. Namely if it would be possible to make an indication on the behaviour of a population on certain processes, these processes could be adjusted likewise. If it would be possible to discover some pattern in behaviour or even better, to discover a certain demand in a process, it would be able to adjust to these discoveries. This could be of great help of course. Thurstone (1927) is often said to be one of the first people to do research in individual choice behaviour into food preferences. He is considered to be the first to describe this preference with some sort of utility function. Nowadays choice models are used in various areas: for example psychology, transport, energy, housing, marketing, voting and actually many more.

Since Thurstone's research there has been quite some development considering choice models. As in most fields of research, a new topic often causes more research and more elaborate research. Since the late 1920s new models have been developed, theories have been adjusted and original assumptions can be avoided. As newer models were developed, not all of these models were applicable due to computational constraints. Together with technological advancements in society, choice models that were unusable became suddenly became usable. Currently models that were not usable thirty years ago are usable and computationally possibilities only increase.

In this paper an attempt is made to describe the theory behind choice models and after that the actual models. The theory behind the models can be considered a framework, some sort of foundation of the models. To be able to understand the models and understand where they come from and what assumptions are made, this will be discussed first. In this first section on individual choice behaviour this framework will be discussed and explained. Most of the definitions needed for the choice models will be given here. After that comes the section on choice models, where the most important and through time most used and referenced models will be discussed. Here no derivations will be given – for derivations of the models the reader is referenced to more extensive literature as Ben-Akiva and Lerman (1985) and Train (2003). Finally the paper will be concluded with a section called Summary, comments and acknowledgement. Included in this section is a chapter with variations on the section before. Unfortunately it was not possible to include all these models in a more elaborate way, but choices had to be made on what literature to discuss. The models that are discussed follow on each other and are instrumental either because they are so often referenced and important in the development of new models or because these models are used currently. I hope the reader finds this paper informative and insightful and in the end has a better understanding of how choice models work and how they have developed over the years.

Individual Choice Behaviour: Framework

"Go down deep into anything and you will find mathematics."
Charles Schlichter (unknown)

Choice Theory

Observing the choices of one individual is interesting, but when statements can be made about a larger group of individuals, or even a whole population, then really something can be achieved. We could therefore state that we are not just interested in the choices of a single person, but rather of large groups of individuals. Think of market demand for some kind of service or commodity. Predicting the demand can be done by modelling individual choice behaviour, thus with the use of choice models. This chapter will be used mostly to describe principles of choice theories and to give a framework, which will be useful when formulating the different discrete choice models.

When examining the behaviour of individuals, in theory we look for behaviour that is according to Ben-Akiva and Lerman (1985) descriptive, abstract and operational. Descriptive so that the theory describes how individuals actually behave and not how we expect individuals to behave. We would also like to formalize their behaviour, independent of specific circumstances, therefore abstract. At last we look for operational behaviour, meaning that it results in actual models with measurable parameters and variables, or at least parameters and variables that can be estimated.

However there is no choice theory that satisfies all these requirements. There are choice theories that have these requirements as ideology, though different models differ in the level of detail in which they idealize the thought process behind observed behaviour. There are some common assumptions though, which are used for the different models. These assumptions will be described as a framework for the models that will be described later on.

Framework

Ben-Akiva and Lerman (1985) state that 'a choice can be viewed as an outcome of a sequential decision-making process that includes the following steps:'

1. Definition of the choice problem
2. Generation of alternatives
3. Evaluation of attributes of the alternatives
4. Choice
5. Implementation

This means that a choice is not viewed as a single choice at a specific time, but rather as a process. An example would be the way someone would travel to work. He could take the bus, go by car, take the bike or walk. Here the definition of the problem would be: how to get to work? The alternatives are stated above. Now the choice is not dependent on the alternatives themselves, but rather on their characteristics, or attributes: how expensive is every alternative? How much time would it take for every alternative? Is it really feasible to walk, meaning what level of comfort does it provide? Eventually the decision maker applies some decision rule, which is some sort of calculation to select the best alternative. In order to define the process above, we need to define the elements decision maker, alternatives, attributes of alternatives and decision rule. Note that we consider actual decision-making process here – choices following from habit, intuition or imitation or any other form of behaviour where there is no rational process are represented as a choice process with only one alternative. Rational behaviour will be discussed later.

The decision maker can be an individual, but also a household, a family or an organization. Because in this case we consider the organization as an individual we

abstract the internal interactions. In this case we are not as much interested in different individual choices, because every individual or family has different interests and backgrounds. We are more interested in predicting aggregate demand, and we must still treat the differences in decision-making behaviour explicitly.

Luce (1959) defined different choices in a situation as alternative choices, or simply *alternatives* and every choice is made from a set of alternatives. The environment of the decision maker determines the universal set of alternatives, but single decision makers do not consider all alternatives. For example, when one goes to work the alternative to take the car to work could be excluded, because the decision maker does not own a car. Therefore, each decision maker considers not the universal set of alternatives, but a subset hereof. This subset is called a choice set. This set includes the alternatives that are feasible and known during the decision process. Finally all alternatives should be mutually exclusive, the choice set needs to be exhaustive – meaning that all possible alternatives are included in the choice set – and the number of alternatives in the choice set must be finite.

As stated before, choices are not made on based on the alternatives apart, but rather on the characteristics or *attributes* of the alternatives. Ben-Akiva and Lerman (1985) state that 'the attractiveness of an alternative is evaluated in terms of a verctor of attribute values.' Attributes can be ordinal or cardinal.

If there are multiple alternatives in a choice set, the decision maker needs a *decision rule* to choose. The decision rule describes the internal process used by a decision maker to process the available information and make a unique choice. Slovic (1977) and Svenson (1979) give us rules that can be classified in four categories:

1. Dominance means that an alternative is better than another alternative when at least one attribute is better and all other attributes are no worse. In most situations this does not lead to a unique choice. It is more often used to exclude the worse alternatives from the choice set. It can be made more complex by using a threshold: one attribute is only better if the difference between both alternatives exceeds a certain threshold.
2. Another decision rule concerns a level of *satisfaction.* This means that every attribute of an alternative must assume a level of satisfaction. This level is set by the decision maker and should be attainable.
3. The third type of decision rule is called *lexicographical rules*. This means that the attributes are ordered by importance. The decision maker chooses the attributes he values the most. If attributes are qualitative, all alternatives that do not possess the desired quality will be excluded from the choice set. In the case the decision maker cannot make a decision based on the most important attribute, he will continue to try and make a decision based on the second most important attribute.
4. The last type of decision rule assumes that a vector that defines an objective function expressing the attractiveness of the attributes of an alternative expresses the attractiveness of an alternative. This attractiveness is referred to as the *utility*. The utility is a measure that the decision maker tries to maximize. In the section utility theory this will be discussed more elaborately.

Of these four categories, the utility has been used most in recent models. We mostly refer to the utility as a function (existing of a vector), the utility function.

Rational Behaviour

The term rational behaviour is based on the beliefs of an observer of what the outcome of a decision of a decision maker should be. It only seems natural that different observers have different beliefs, therefore it seems that there cannot be one universal type of rational behaviour. Thus rationality is not really a useful concept when applied to individual choice behaviour. The concept described in literature is one opposing impulsiveness, which means that individuals do not make decisions based on their variable psychological state at the time the decision has to be made. It means that it follows a consistent, calculated decision process. This does not mean that the individual cannot follow his or her own objectives.

In 1957 Simon described the distinction between what he called a perfect and a bounded reality. In a perfect world an all-knowing individual exists, capable of gathering and storing large quantities of data and perform complex calculations on these data and is able to make consequent decisions based on the data. Bounded reality recognizes the bounded capacity of humankind as problem solvers with limited information-processing capabilities.

This means that rationality is a quite ambiguous concept. It is therefore necessary to introduce a specific set of rules, to be able to use the concept. Simply said this means that we assume that a decision maker, if alternative A is more feasible than alternative B will choose A every time he faces that same decision. And if alternative A is more feasible than B and B is more feasible than C, the decision maker will prefer A to C.

Discrete and Probabilistic Choice Theory

This section can be seen as an expansion of the section about the framework of choice theory. The concepts used in choice theory are similar to the concepts in Economic Consumer Theory, which will not be treated in this paper. The view on demand in Economic Consumer Theory is well applicable when the feasible choices have continuous variables, but this might not always be the case. In discrete choice theory types of problems are better described as discrete bundles of attributes. Furthermore, in probabilistic choice theory the probability that a decision maker chooses a certain alternative can be provided, which makes it a powerful framework when working with discrete choice situations.

Discrete Choice Theory

Table 1 Choice in travelling to work

	Attributes		
Alternatives	Travel Time (t)	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

For the section on discrete choice theory, consider a simple example that is used in literature (Ben-Akiva & Lerman 1985, Train 2003) more often. Consider a decision maker that has to travel to work. He has three options: take the car, take the bus or walk. The attributes of the alternatives are travel time, travel cost and comfort (see table 1).

The choice will have the utility function $U = U(q_1, q_2, q_3)$ with q is the alternative chosen. Obviously the decision maker can only choose one alternative, this $q_i = 1$ if mode *i* is chosen and $q_i = 0$ if otherwise, for all i in the choice set and $q_1q_2 = q_2q_3 = q_1q_3 = 0$. Because the possibilities U(1,0,0), U(0,0,1) and U(0,1,0) are not differentiable we apply a

maximization on the function with the attributes as parameters: $U_i = U(t_i, c_i, o_i)$. Now we can see that $U_1 > U_2$ and $U_1 > U_3$ have to be true for alternative 1, taking the car, to be chosen.

As for the form of the utility function, in most literature an additive utility function is assumed:

$U_i = - \beta_1 t_i - \beta_2 c_i + \beta_3 o_i$,

for all i in the choice set and $\beta_i > 0$ for all i. With this formula we can try to predict changes in U for different numerical values for the parameters. This approach to the utility function is called *revealed preference* and will be discussed more elaborately in the next chapter.

Lancaster (1966) defined the utility function $U_{in} = U(\boldsymbol{x}_{in})$, where $\boldsymbol{x}_{in}$ is the vector of the attribute values, for every alternative i by every decision maker n. Ben-Akiva and Lerman expand this formula a bit, due to variability in population: $U_{in} = U(\boldsymbol{x}_{in}, \boldsymbol{s}_n)$, where $\boldsymbol{s}_n$ is a vector of different characteristics of the decision maker, for example income, age, education and ethnical background.

Probabilistic Choice Theory

In probabilistic choice theory, it is argued that we cannot approximate human behaviour by deterministic parameters. It seems plausible to state that human behaviour has a probabilistic nature. Furthermore, it can be argued that whilst the decision maker has knowledge of his or her utility function, the researcher or analyst does not know the exact form. Therefore Train (2003) explains about the term representative utility. In the section about the framework the utility function of the form U was introduced. The decision maker chooses the alternative if $U_{in} > U_{jn}$ $\forall$ $j \neq i$, where j are the different choices from the choice set (C_n) and n is the labelled decision maker. Since there are aspects of utility function of a decision maker that the researcher does not know, we introduce the representative utility function $V_{jn} = V(x_{jn}, s_n)$ with x_{nj} $\forall$ j, again the attributes of the alternatives and s_n some attributes of the decision maker.

Because V depends on characteristics the researcher cannot know, it makes sense that $V_{jn} \neq U_{jn}$. Train states that the utility can be decomposed as $U_{jn} = V_{jn} + \varepsilon_{jn}$, where ε_{jn} captures the factors that affect utility but are not known to the researcher and therefore are not included in V_{jn}. Simply said ε_{jn} is the difference between U_{jn} and V_{jn} and could be considered an error term.

It seems logical that if ε_{jn} are factors that affect the utility, but are not known by the researcher, the form of ε_{jn} is unknown as well. Therefore these terms are treated as random. The joint density of the vector of these 'errors' is denoted $f(\varepsilon_n)$.

$$P_{in} = \Pr(\varepsilon_{jn} - \varepsilon_{in} < V_{in} - V_{jn} \forall j \neq i)$$

$$= \int_{\varepsilon} I(\varepsilon_{jn} - \varepsilon_{in} < V_{in} - V_{jn} \forall j \neq i) f(\varepsilon_n) d\varepsilon_n.$$

The first part is due to the probability that the decision maker chooses alternative i. In this part $I(\ldots)$ is the indicator function, which is equal to 1 if the statement between the parentheses is true and 0 if not. This is a multidimensional integral and only takes a closed form for specific forms of the density function f. For example logit and nested logit are models that will be discussed later on, but those have a closed form for this integral. Probit and mixed logit are derived differently and do not have an open form for this integral. They are not calculated exactly but they are evaluated numerically.

An example might help to clarify. Consider the example used in the last section, but this time without the attribute comfort. We can define $V_i = - \beta_1 t_i - \beta_2 c_i$ with i = {car,

bus, walk}. Now suppose after analysis the researcher finds that V_{car} = 4, V_{bus} = 3 and V_{walk} = 3. This does not mean that the decision maker will choose to go to work by car. It simply means that by observed factors it seems best to go by car, but there are still factors that are unobserved to the researcher. The probability that the decision maker walks to work instead of taking the car is the probability that the unobserved factors for walking are sufficiently better than those for taking the car. As stated in the formula above it would be:

$P_{walk} = \Pr(\varepsilon_{car} - \varepsilon_{walk} < V_{walk} - V_{car})$.

Utility theory

Anand (1993) states that decision theory is about 'choosing the act that is best with respect to the beliefs and desires that an agent holds.' He states that utility theory helps achieve this. Loosely said decision theory is about maximizing utility, given the 'attributes' of an agent or rather decision maker. Many theories exist about utility. Ben-Akiva and Lerman divide utility theory into two possible types: constant utility and random utility. In economics, usually the difference between cardinal and ordinal utility is used, which are in fact not very different from the former mentioned types.

Cardinal utility

Cardinal utility is usually considered out-dated though, as is the constant utility approach, because it is not really in line with consumer theory. In cardinal theory the magnitude of difference between utility values is treated as behaviourally significant. For example in the case of the example that has been used: if taking the car to work takes 10 minutes, taking the bus takes 20 minutes and walking takes 30, one can say that taking the car is as much better than taking the bus as is taking the bus than walking. But one cannot say that taking the car is twice is good as taking the bus. From this example comes forward that comparisons between alternatives is meaningless, because there is no good way to interpret differences between them. Nowadays instead of cardinal utility often preferences are used. Stated and revealed preference will be discussed in the next chapter.

Ordinal utility

When using ordinal utility and considering the example above, one can say that taking the car is preferred to the bus to walking, but one cannot state anything about the strengths of these preferences. Thus using ordinal utility it is possible to capture ranking, but not relative strengths. In ordinal utility theory, there is usage of a utility function to capture rank, as is the case with constant and random utility.

Constant utility

The values for the utilities of the different alternatives are fixed in this approach. Here it is not the case that the decision maker chooses the alternative with the highest utility, but it is assumed that there are choice probabilities involved. These probabilities are defined by a probability density function (PDF) over the different alternatives, with the utilities as parameters. Selecting a certain PDF in this approach can only be based on specific assumptions with respect to the properties of choice probabilities. An important property of this approach is the independence from irrelevant alternatives (IIA) that is noted as follows:

$$\frac{P(i \mid \overline{C}_n)}{P(j \mid \overline{C}_n)} = \frac{P(i \mid C_n)}{P(j \mid C_n)}, \qquad \text{with } i, j \in \overline{C_n} \subseteq C_n$$

which simply said means that removing irrelevant alternatives from the choice set C_n, resulting in subset $\overline{C_n}$, has no influence on the choice probabilities.

Random utility

Manksi (1977) formalized this approach, which is more in line with consumer theory than the constant utility approach. The observed inconsistencies, or errors, are now viewed as to be result of observational inaccuracies on the researcher's side. In this approach we again assume that a decision maker tries to maximize his or hers utility, as is in line with economic consumer theory. But – as stated in the section about probabilistic choice theory – the researcher does not know the utility of a decision maker with full certainty and therefore they are treated as random variables. We can say that the researcher defines the choice for a specific alternative *i* in the choice set as $P(i \mid C_n) = \Pr(U_{in} \geq U_{jn}, \forall j \in C_n)$,

as stated by Ben-Akiva and Lerman (1985). Here we assume a joint PDF for the set of (random) utilities, because a logical argument can be made about the underlying source for randomness in the utilities. Manski (1973) identified four sources:

1. *Unobserved attributes*: the vector of attributes that affects the decision is incomplete to the researcher. Therefore there is an element included in the utility function that is observationally random, thus it follows that the utility is random as well.
2. *Unobserved taste variations*: the researcher knows all attributes, but there is an unobserved argument that is unknown to the researcher. This can be explained as the researcher being unknown of the specific taste, of preference, of the decision maker. The variation of this argument is unknown, making the utility random.
3. *Measurement errors*: the attributes of the alternatives are not observable. That is why the researcher estimates the attributes, with a measurement error accounting for probable inaccuracy in the measurement. This error term is unknown, resulting in the utility becoming random.
4. *Instrumental variables*: the true utility function is known, but some elements in the vector of attributes are not observable. The researcher approaches these variables by a function that is derived by the relation of known variables. This means the utility functions contains instrumental variables, which are in nature an expression of an imperfect relation between estimate and actual attribute. Again this term contains a random error, making the utility random.

Expected utility

Expected utility is also known, especially because it is one of the underlying assumptions in game theory. This approach deals with analysis of choices in risky projects. This is one of the oldest utility approaches, as it was formulated in 1713 by Nicholas Bernoulli and solved in 1738 by Daniel Bernoulli. Savage (1954) formulated the subjective expected utility theory, which is a more up-to-date work and was reviewed by Anand (1993). If an uncertain event has a number of possible outcomes z_i all have utilities $U(z_i)$, and there is a subjective probability $P(z_i)$, then the subjective expected utility would be:

$\sum_i U(z_i)\mathrm{P}(z_i)$.

Savage also stated eight axioms that also suit well with the other utility approaches. These axioms are:

1. *Completeness*: if x and y are two alternatives, either x or y is preferred. Also, x and y are equally desirable.
2. *Transitivity*: if x is preferred to y and y to z, then x is preferred to z.
3. *Independence*: x and y should be independent of each other.
4. *Resolution independence*: Preference for an alternative x or y only depends on the attributes of the alternative ex ante.
5. *Expected wealth independence*: The preference for an alternative depends on the chance of winning and not on the size of the stakes: if there is a lottery and one can choose between lottery A, which has a 15% chance of winning and a 100 euro reward and lottery B, which has a 5% chance of winning and a 1000 euro reward, the decision maker will choose to participate in lottery A.
6. *Minimal strict preference*: there is at least one vector of attributes that is strictly preferred to the other vectors of attributes.
7. *Continuity in probability*: very unlikely events should be regarded as having zero probability.
8. *Partial resolution independence:* if the attributes of x are preferable to the attributes of y for different states, then x is preferred to y if one of the states is obtained.

Stated and Revealed Preference

Now the main framework behind choice models has been treated, we will turn to When considering data involved in choice models, it is possible to divide these data in two distinct types: *stated preference* (SP) data and *revealed preference* (RP) data. In the following two sections I will give a description of the models and point out strengths and weaknesses of the both data types. These sections will not contain specific methods of both approaches, but it will give insight into these approaches and how they differ from each other.

Stated preference

According to Kroes and Sheldon (1988) SP methods refer to 'a family of techniques which use statements of individual respondents about their preferences' in a set of alternatives to estimate utility functions. SP data are collected through experimental situations or surveys where the respondents are faced with hypothetical choice problems. For example, the respondent is asked to choose between five bikes. In this hypothetical situation only these five bikes exist. The response is the stated choice.

Another way to describe the SP approach would be the direct approach, because the data comes directly from the respondents, the hypothetical decision makers. Due to this approach the data does not describe actual behaviour, but it describes how decision makers state they would behave alike. A strong point from this data is that it can give an indication how respondents would behave in a situation that currently not (yet) exists. So if it is the researcher's objective to examine behaviour for example in a product that does not yet exists, the SP approach would be suitable. Also SP data works very well in data that contain little or no variation, because the questionnaire can be designed in a way that will result in the data having the desired variation.

The main disadvantage of SP data seems obvious: the way respondents expect themselves to behave, moreover the way respondents say they will behave, is not the way they actually will behave. This phenomenon may arise because the respondent actually does not know how they would respond or because he or she feels it is expected of them to respond in a specific way.

Revealed preference

In contrast to SP data, RP data relate to actual behaviour. It is called RP because decision makers reveal their preference through the choices they make. In the example used in the SP section and considering the RP approach, the respondent would be asked what bike he or she bought last, instead of choosing from a selected set in a hypothetical situation. Therefore we can state that purchasing or choosing habits reveals preference. In this approach utility functions are defined by observing behaviour.

Where the SP approach is called the direct approach, the RP approach is called the indirect approach. In the RP approach actual behaviour is observed, instead of confronting the respondent with a hypothetical situation. The largest advantage of RP data is that the data represents actual choices.

The downside of RP data representing actual choices is that it not suitable for situations that currently not exist. Because we observe behaviour, there is too much uncertainty in stating behaviour in new situations. Of course approximations can be made, but SP data is simply better suitable for these situations. Also in situations with little or no variation RP data is not suitable, because relations between different cannot be estimated well without variation.

By using an estimation procedure that allows the relative importance of the attributes to be estimated through primarily SP data and at the same time allows the alternative-specific constants and overall scale of the parameters to be determined by RP data (Train, 2003), the strengths of both approaches can be used. This will not be discussed into detail in this paper, but Hensher et al (1999) and Brownstone et al (2000) describe this approach for respectively logit models and mixed logit models.

Exogenous, Locational and Utility-based Choice Models

Until now only choice models based on utility and especially random utility have been considered. The utility models are models that involve a set of alternatives, a decision maker and some utility function that describes how the decision maker chooses the most attractive alternative to them. In other words, the decision maker bases his or her decision on the attributes of the alternatives and chooses the most preferable alternative through some decision process described by the utility function. Now there are different types of models besides utility-based models. These models will not be discussed elaborately, but it will be good for the reader to be aware that other types of models exist.

Exogenous-based Models

Paul Waddell (1993) investigates whether the assumption that the choice of workplace is exogenously determined in models of residential location is true. So this research builds very much on McFadden's (1978) research. Exogenous-based models state that choices are driven by outside factors and are therefore very different from utility-based models, where the decision made is dependent on characteristics of the decision maker

as well as attributes of the alternatives; one could say all endogenous variables. There has been debate whether or not locational-based models or somewhat exogenous, since the paper by Waddell. Before the 1990s locational-based models were assumed to be exogenous, as residential location was assumed to driven by mostly workplace. Even now one could argue if this makes a model exogenous, as workplace is not the only deciding factor and other factors might not be exogenous. In his paper, Waddell reaffirms many of the influences assumed in urban economic theory, the same accounts for the assumption on the relation between workplace and residential location.

Locational-based Choice Models

This type of model is not that different from utility-based models as one might think. The former models developed in the 1960s (Alonso 1964, Muth 1969) originate from a model called the monocentric model or are derived from gravity model (Lowry 1964) derivatives. These models will not be discussed here, but they do have a very important assumption in common: workplace choice is exogenous in determining the residential location choice of households. So there is a link with exogenous-based models. Still very often researchers reference to the work of McFadden (1978) where he described how utility-maximizing consumers are assumed when considering residential housing, thus locational choice. In this type of choice model there is also some random part in the utility. McFadden points out that the MNL and NL model, which will be discussed later in this paper, are very usable in locational-based choice modelling.

A difference to some extent in locational-based choice modelling is that the characteristics of alternatives are not just decided by their own attributes, but also by some external attributes. Think of climate, image of the location or employment in the area. It is possible that alternatives should be placed in a Nested Logit model which allows for overlap (Cross-Nested Logit) but also allows for classes to overlap, as houses belong to different classes but also are defined by exogenous-driven factors. McFadden (1978) concludes that the problem of modelling disaggregate choice of housing location is impractically large. So to a certain extent locational-based choice models do not differ as much from utility-based choice models as the foundation is similar.

Head et al. (1995) discussed location choice through the example of Japanese manufacturing investments in the United States. They state that firms in the same industry are drawn to similar locations because proximity causes positive externalities. This is very much in line with McFadden who stated that with location choice much more exogenous-driven variables play a role. Head et al. also state that chance events can have a lasting influence on the geographical pattern of manufacturing in that case. This is also a big difference with utility-based modelling. There the assumption is made that in the largest sense rational processes take place and therefore if the same decision maker conducts the same decision processes, similar outcomes will take place. In the case of location choice this apparently is not the case, as choices made also have an effect on the actual alternatives.

Binary Choice Models

In this chapter a general background of binary choice models will be given, that will be used in the following chapters when specific models will be discussed.

As stated in the chapter on choice theory, we have a decision maker facing a set of feasible discrete choice alternatives and he or she will select the alternative with the

greatest utility, with the utility a random variable (r.v.). As in random utility theory, the probability that a decision maker will select a certain alternative is
$P(i \mid C_n) = \Pr(U_{in} \geq U_{jn}, \forall j \in C_n)$.

If the choice set C_n exists of only two alternatives, *i* and *j*, we have a so called binary choice model. In this case we can state the probability that decision maker *n* will choose alternative *i* or *j* is:

$P_n(i) = \Pr(U_{in} \geq U_{jn})$ and $P_n(j) = 1 - P_n(i)$.

Ben-Akiva and Lerman (1985) describe how theory described in the last chapter, random utility theory, can be made operational:

1. Separating the total utility into deterministic and random components of the utility function
2. Specify the deterministic component
3. Specify the random component

Remember that in the section of probabilistic choice theory we stated that for the utility we have an observed part and an unobserved part, we also called disturbance:
$U_{in} = V_{in} + \varepsilon_{in}$,
$U_{jn} = V_{jn} + \varepsilon_{jn}$
Here V_{in} and V_{jn} are the systematic components. In the chapter on choice theory *V* is described as the part of the utility that can be observed by the researcher. These components are assumed to be deterministic. *V* can be thought of as the means of *U*. We can shift the scale of measurement by transforming both U_{in} and U_{jn} by any strictly monotonic increasing function. Ben-Akiva and Lerman show that adding a constant to both utilities has no affect on the choice probabilities. It does change V_{in} and V_{jn}, but eventually that is no problem. The absolute levels of *V* and ε do not matter, what does matter is that $V_{in} - V_{jn} < \varepsilon_{in} - \varepsilon_{jn}$. Though by specifying just the differences instead of individual components could develop binary choice models, usually each utility function is specified separately for the sake of continuity. There exist choice models with more than two alternatives; therefore the same notation is used for binary choice models.

After separating the utility into a deterministic and a random part, we specify both parts, starting with the deterministic or systematic part. As *V* does not just depend on the underlying attributes, but also on attributes of the decision maker, we can define *V* as $V(\boldsymbol{z}_{in}, \boldsymbol{S}_n)$, as is similar to a description in the chapter on choice theory. Seeing as these two vectors $\boldsymbol{z}$ and $\boldsymbol{S}$ are actually combined to describe *V*, we define a new vector $\mathbf{x}_{in} = \boldsymbol{h}(\boldsymbol{z}_{in}, \boldsymbol{S}_n)$, with **h** is some vector-valued function. Now we can write $V_{in} = V(\mathbf{x_{in}})$ and $V_{jn} = V(\mathbf{x_{jn}})$. Secondly, a functional form for *V* is needed. Because we would like it if the function was to reflect theory about how the elements in **x** influence utility and we want the function's parameters to be estimated easily, most researchers have chosen to use functions that are linear in the parameters. If we define $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_K]$ as the vector of K unknown parameters, we can write
$V_{in} = \beta_1 x_{in1} + \beta_2 x_{in2} + \ldots + \beta_K x_{inK}$,
$V_{jn} = \beta_1 x_{jn1} + \beta_2 x_{jn2} + \ldots + \beta_K x_{jnK}$.

for both observed utilities. One important characteristic is that linearity in parameters does not mean linearity in the attributes **z** and **S**. This totally depends on the form of **h**. Finally, here it is assumed that the parameters $\beta_1, \beta_2, \ldots, \beta_K$ are the same for the whole population. It is possible however, that market segmentation is present. Then $\beta_1, \beta_2, \ldots, \beta_K$ are treated as r.v.'s distributed across the population.

Finally we need to specify the disturbances to obtain an operational binary choice model. Where in the last paragraph the functions for V were depicted separately, now the functions for the disturbances ε will be depicted in the most convenient way. So the differences $\varepsilon_{jn} - \varepsilon_{in}$ are discussed. As stated before, the choice probabilities are unaffected if we add a constant to both disturbances. Besides this it also will not make any difference if the mean of the disturbance is shifted, as long as the systematic component is shifted by the same amount. From this follows that the means of the disturbances can be fully represented by any constant without loss of generality. Usually it is assumed that all disturbances have zero means. In addition to the mean, the scale of the disturbances should be consistent with the scale of the functions V. As for the functional form of the distribution of the disturbances, it does not make sense to think of the distribution of the ε's to be different from the V's distribution. Especially since the disturbances reflect the different sources of observational error, different specifications of V will lead to different, fitting, distributions for ε. Because there are many different unobserved factors that affect the overall distribution, it is hard to make strong statements about this distribution. However, nowadays more and more we gain insight about what is included into the disturbances.

Choice models

"If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is."
John von Neumann (1903-1957)

Logit and Probit

Now that the framework of choice theory in general and binary choice models has been set, we are now able to talk about specific models. There are three common binary models: the linear probability model, the (binary) logit model and the (binary) probit model. These models were both discussed by Thurstone (1927) to some extent.

The differences between these models are based on the assumption that is made about the distribution of the disturbances, or the difference between the disturbances. To obtain the eventual model, the choice probabilities can be derived under the assumption about the disturbances.

Linear Probability Model

The easiest of the three models is the linear probability model. In this model the difference in the disturbances is uniformly distributed: $\varepsilon_{jn} - \varepsilon_{in} \sim Unif(-L, L)$, where $L >$ 0. The difference between the disturbances is defined as $\varepsilon_{jn} - \varepsilon_{in} = \varepsilon_n$, with density function $f(\varepsilon_n)$. Here

$P_n(i) = \Pr(\varepsilon_n \leq V_{in} - V_{jn})$.

The choice probability is given by the cumulative distribution function of ε_n. When V is linear in its parameters, the probability function is linear as well between $-L$ and L.

According to Cox (1970) this model has a major drawback: unless restrictions are placed on the β's (which are again used to estimate V), the estimated coefficients can imply probabilities outside the interval [$-L$, L]. Therefore the logit and probit models are used more often. Besides this drawback, it is unrealistic to assume the interval [$-L$, L], and zero probabilities outside this interval.

Probit

Another way to view the disturbances is as being the sum of a large number of unobserved, independent constituents. Due to the large number and the central limit theorem the disturbances tend to be normally distributed.

Now we can state that ε_{in} and ε_{jn} both have a normal distribution with mean zero and variance σ_i^2 and σ_j^2 respectively. Now the difference between the disturbances also has a normal distribution with mean zero and variance $\sigma_i^2 + \sigma_j^2 - 2\sigma_{ij} = \sigma^2$. When $V_{in} = \beta' x_{in}$ and $V_{jn} = \beta' x_{jn}$

we can state for the choice probabilities:

$$P_n(i) = \Phi\left(\frac{\beta'(x_{in} - x_{jn})}{\sigma}\right),$$

where Φ denotes the standardized cumulative normal distribution.

The choice probability here only depends on σ, not on the variance of either disturbance or the covariance. Moreover, the choice for σ is arbitrary, as rescaling σ or β by any positive constant will not affect the choice probability. Usually σ = 1 is chosen.

Of course the assumption on normality is a very convenient assumption, as it improves possibilities considering calculations compared to the linear probability model, but it can also be a limitation. Now a normal distribution is required for all unobserved components in the utility. Also the integral for the choice probabilities has an open form for probit models. This is not a big problem, but is not considered convenient analytically.

Logit

Logit models are very much alike probit models, but a big difference is that the integral for the choice probability has a closed form, which makes these types of models analytically more convenient. In the logit model it is assumed that $\varepsilon_{jn} - \varepsilon_{in} = \varepsilon_n$ is logistically distributed. The logistic distribution approximates the normal distribution, but has fatter tails. Under this assumption, the choice probabilities are:

$$P_n(i) = \frac{e^{\mu V_{in}}}{e^{\mu V_{in}} + e^{\mu V_{jn}}}.$$

If the *V* functions are linear in their parameters, the choice probabilities can be derived onto

$$P_n(i) = \frac{1}{1 + e^{-\mu\beta'(x_{in} - x_{jn})}}.$$

Here μ is a positive scale parameter. For convenience it is assumed, as has been done similarly for probit, that μ = 1. But for probit σ = 1 is chosen, which corresponds with var($\varepsilon_{jn} - \varepsilon_{in}$) = 1. This also implies that the scaled logit coefficients are $\frac{\pi}{\sqrt{3}}$ times larger than scaled probit coefficients.

Train (2000) describes a couple of characteristics of the logit model, which prove to be as well the power of the model, but in some sense also the limitations. Firstly the logit model is able to represent systematic taste variation very well, but the flipside is that it cannot represent random taste variation. Secondly, if the unobserved factors are independent over time in repeated choice situations, the model can capture the dynamics of repeated choice. On the other hand this seems restrictive as it exhibits substitution patterns.

There are some limiting cases for all three models. If $\mu \to \infty, \sigma \to 0 \; or \; L \to 0$, $P_n(i)$ will become 1 if $V_{in} - V_{jn} > 0$ and 0 otherwise. If $\mu \to 0, \sigma \to \infty \; or \; L \to \infty$, there is exactly 0.5 probability for both alternatives.

Estimation

For both logit and probit models, usually maximum likelihood estimators are used to estimate the parameters $\beta_1, \beta_2, \ldots, \beta_K$ from a (random) sample of observations from the population. An indicator variable y_{in} is constructed and defined as 1 if person *n* chose alternative *i* and 0 if that decision maker chose alternative *j*. Also two vectors $\mathbf{x_{in}}$ and $\mathbf{x_{jn}}$ are constructed, both contain all *K* values of the relevant variables. Now given a sample of *N* observations, we now have to find estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K$.

Now we consider the likelihood, which is eventually equal to the probability of the observed outcomes given the parameter values $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K$. Since the assumption is that they are drawn at random from the whole population, we can state that the likelihood of the sample in total is the product of the likelihood of all individual observations. Analytically it is more convenient to consider the logarithm of the likelihood function, denoted as log *L*. Now we can write the likelihood as:

$$\log L(\beta_1, \beta_2, \ldots, \beta_K) = \sum_{n=1}^{N} [y_{in} \log P_n(i) + y_{jn} \log P_n(j)],$$

where $P_n(i)$ is a function of $\beta_1, \beta_2, \ldots, \beta_K$. Now the log *L* function is solved to maximize it by differentiation with respect to the β's and then setting the partial derivatives to zero,

thus we solve max log $L(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K)$, while we seek the estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K$ that solve this function.

Often if a solution to the first-order conditions exists, it is a unique solution. However it is quite possible that there will be multiple solutions to the first-order conditions. Just one of these solutions constitutes the maximum likelihood estimate. The estimates are consistent and asymptotically normal. The estimates are given in a matrix of the second derivatives of the logarithmic likelihood function with respect to the parameters, which are evaluated at the true parameters. The estimate in the *k*th row and the *l*th column is $\frac{\partial^2 L}{\partial \beta_k \partial \beta_l}$. Since we do not know the actual values of the parameters where we need to evaluate the second derivatives or the distribution of the **x** vectors, usually an estimated variance-covariance matrix that is estimated at the estimated parameters $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K$ and the sample distributions of the vectors **x** to estimate their actual distribution. Therefore

$$\sum_{n=1}^{N}\left[\frac{\partial^2[y_{in}\log P_n(i) + y_{jn}\log P_n(j)}{\partial \beta_k \partial \beta_l}\right]_{\beta=\hat{\beta}}$$

is used as a consistent estimator of the actual value.

As for the computational aspect of this problem, it is known that the solutions of the first-order conditions are typically non-linear and the use of a computer is needed to solve, even for two-variable problems. Ben-Akiva and Lerman (1985) describe how the Newton-Raphson algorithm can be used. Briefly described the algorithm works as follows: firstly an initial guess for the parameters is made. Then the function of second derivatives around the parameters is made. Then the linearized form received after approximating after first-order condition is solved and finally we look at the difference in steps between the new approximations. If it is small enough (Ben-Akiva and Lerman describe different criteria) then these approximations are used. Most other procedures are similar, but with different steps for the second and third step.

Alternative estimation models

The maximum likelihood estimation procedure is used mostly for the logit and probit model. For the linear probability model the least squares method, as is more common for regression models, or Berkson's procedure is used more often. As probit and logit are the main models to be discussed here, least squares and Berkson's procedure will not be discussed in this chapter.

Multinomial Logit

The last two chapters the main focus has revolved around binary choice models and the estimation technique behind these models. In most decision processes however, the number of alternatives in the choice set is not limited to two. This type of model is called a multinomial choice model. Again, the choice set is different for every individual, as each individual has their own index of attributes and a different subset of the universal set. In this case where more that two alternatives can be chosen the derivation of choice models and estimation models are more complex that those for binary choice models. Instead of using just the difference between the disturbances, we now need to characterize the whole joint distribution of all disturbances.

Different types of multinomial choice models exist. It is possible to expand the binary logit and binary probit model to multinomial models. Dow and Endersby (2003) compare multinomial logit (MNL) and multinomial probit (MNP) for voting research. They state that the MNL model is preferable to the MNP model. As explained in the last chapter, the logit model has a closed form integral whilst the probit model has an open form. Therefore MNP is more complex than MNL and it could give some estimation problems. Burda et al. (2008) present a model that is a mix between MNL and MNP where estimation is conducted by using a Bayesian Markov Chain Monte Carlo technique. However in this chapter we focus specifically on the multinomial logit model, thus we will elaborate and expand on the theory treated in the last chapter.

First some sort of background will be painted concerning multinomial choice theory. Then some definition of the MNL model and its characteristics, strengths and weaknesses will be given before considering estimation models.

Multinomial choice

As stated before, every decision maker has as his or hers choice set some subset of the universal set, and every decision maker can have a different subset. Manski (1977) calls this process of generating a subset from the universal set "the choice set generation process". However for the researcher it makes the model a lot more complex if every individual decision maker can choose a different choice set than other decision makers. It is defined that every choice set C_n has $J_n \leq J$ feasible choices. Now we state that the probability that an alternative i is chosen follows directly from the probability described in the section on random utility theory:

$$P_n(i) = \Pr(U_{in} \geq U_{jn}, \forall j \in C_n, j \neq i).$$

Here we can distinct a deterministic and a random component in the utility:

$$P_n(i) = (V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}, \forall j \in C_n, j \neq i).$$

Define $f(\varepsilon_{1n}, \ldots, \varepsilon_{J_n n})$ as the joint density function of the disturbances. Now there are different ways to express the choice probabilities described in literature. Ben-Akiva and Lerman offer three ways of deriving $P_n(i)$. The most insightful way is to reduce the multinomial problem to a binary problem, as we discussed this earlier. We can state that

$$U_{in} \geq U_{jn,} \ \forall j \in C_n, j \neq i$$

and from this follows that if alternative i is preferred over all other alternatives j, that also:

$$U_{in} \geq \max_{\substack{j \in C_n \\ j \neq i}} U_{jn}.$$

Therefore we can write these last formulas combined:

$$P_n(i) = \Pr\left[V_{in} + \varepsilon_{in} \geq \max_{\substack{j \in C_n \\ j \neq i}} (V_{jn} + \varepsilon_{jn}) \right].$$

As U_{jn} is a random variable the maximum of U_{jn} will be a random variable as well. Now the distribution of this maximum has to be derived from the underlying distribution of the disturbances. Where for the MNP model for example this is quite a task, there it is doable for the MNL model. This is a reason that MNL is a much-used model.

Multinomial logit

In the section on binary logit the choice probability was described as follows:

$$P_n(i) = \frac{e^{\mu V_{in}}}{e^{\mu V_{in}} + e^{\mu V_{jn}}}.$$

For the MNL model the choice probability is similar:

$$P_n(i) = \frac{e^{\mu V_{in}}}{\sum_{j \in C_n} e^{\mu V_{jn}}}.$$

These formulas are equal if $J_n = 2$ and $\mu = 1$. Here it is evident that the MNL model in an extension, a development on the binary logit model. Also in the function it can be seen that it is a proper probability mass function, as $0 \leq P_n(i) \leq 1$ and its sum over all i is equal to 1. Here again $U_{in} = V_{in} + \varepsilon_{in}, \forall i \in C_n$, with the disturbances independently and identically distributed (iid) and Gumbel-distributed with location and scale parameters η and μ. As with binary logit, as long as all systematic terms of the utility include a constant, it is not restrictive to take $\eta = 0$. To calculate the probability of one of the alternatives in the choice set, order the alternatives so that alternative $i = 1$. Now we still have an unidentifiable parameter μ, but it is common to set this parameter to a convenient value, usually $\mu = 1$ is used.

Obviously a big advantage of the MNL model is that is able to analyze a choice set consisting of multiple alternatives, as is possible for the MNP model. Though in literature (Bunch 1991, Alvarez and Michael 2001) use of MNP in several fields is recommended, MNP cannot work optimally with a large number of observations. The MNL model however is able to work with larger datasets. Both models do not work effectively with small datasets. Also the MNL model is criticized because of the independence of irrelevant alternatives (IIA) property, which states that for a specific individual the ratio of the choice probabilities of any two alternatives is completely unaffected by systematic utilities of other alternatives. This is closely related to the assumption that all disturbances are mutually independent. Ben-Akiva and Lerman state that the problem does not per se lie with the IIA property, but rather every model that has an underlying assumption that the disturbances are mutually independent state similar results. Dow and Endersby (2003) state that for most applications the IIA property is not particularly restrictive and for most applications not even relevant.

Besides these strengths and weaknesses there are two limiting cases for the MNL model, as was the case for the binary models. Firstly if $\mu \to 0$ then:

$P_n(i) = \frac{1}{J_n}, \quad \forall i \in C_n$. This means that if $\mu \to 0$ the variance of the disturbances approaches infinity, so the model will not provide any information. That means that all alternatives are equally likeable. The other limiting case is when $\mu \to \infty$. Now the variance of the disturbances approaches zero and the model becomes deterministic.

Estimation

For the MNL model maximum likelihood is commonly used for estimation of the parameters as well. For the maximum likelihood estimation procedure for the MNL model there are no big differences with binary logit, but their computational burden grows with the number of alternatives. McFadden (1974) showed that for the MNL model has some special properties that can simplify estimation of its parameters under certain circumstances.

Again, most of this theory is expansion on the section on maximum likelihood estimation for the binary logit model. Again let y_{in} be 1 if decision maker or observation n chose alternative i and 0 otherwise. We write the likelihood function:

$$L = \prod_{n=1}^{N} \prod_{i \in C_n} P_n(i)^{y_{in}},$$

with for the choice parameters or as Ben-Akiva and Lerman (1985) state, the linear-in-parameters logit:

$$P_n(i) = \frac{e^{\beta' x_{in}}}{\sum_{j \in C_n} e^{\beta' x_{jn}}}.$$

As in the section on binary logit, we rewrite the likelihood function to a log likelihood function:

$$\log L = \sum_{n=1}^{N} \sum_{i \in C_n} y_{in} (\beta' x_{in} - \ln \sum_{j \in C_n} e^{\beta' x_{jn}}).$$

When setting the derivatives of the log likelihood function to zero, the first-order conditions can be obtained:

$$\sum_{n=1}^{N} \sum_{i \in C_n} [y_{in} - P_n(i)] x_{ink} = 0, \quad for\ k = 1, \ldots, K.$$

This can be rewritten as

$$\frac{1}{N} \sum_{n=1}^{N} \sum_{i \in C_n} y_{in} x_{ink} = \frac{1}{N} \sum_{n=1}^{N} \sum_{i \in C_n} P_n(i)\, x_{ink}, \quad k = 1, \ldots, K.$$

This shows that the average value of an attribute for the chosen alternatives equals the average value predicted by the estimated choice probabilities. Moreover, this means that if an alternative-specific constant is defined for the alternative i at the maximum likelihood estimates the sum of the choice probabilities is equal to the number in the sample that chose i. All properties of the maximum likelihood estimation of binary logit extend to the MNL model. This also applies for the computational methods that are used for solving the system of K equations.

Nested Logit

Hensher et al. (2005) state that the bulk of choice behaviour study applications do not go farther than the simple MNL model, because of the ease of computation and because of a wide availability of software packages. They also state this it does come with a price in the form of the IID assumption on the disturbances and the IIA property, which will are violated at times. The nested logit (NL) model includes a partial relaxation on both assumptions. As the MNL model, the NL model as a closed form solution in contrast to for example the multinomial probit (MNP) and mixed logit (ML) model. The ML model will be discussed in a later chapter.

The NL model is a so-called multidimensional choice model. Many choice situations are not just situations where a decision maker has to choose from some list of alternatives, but where the set of alternatives are combinations of underlying choice dimensions, for example as shown in figure 1, where a NL model is depicted with two dimensions.

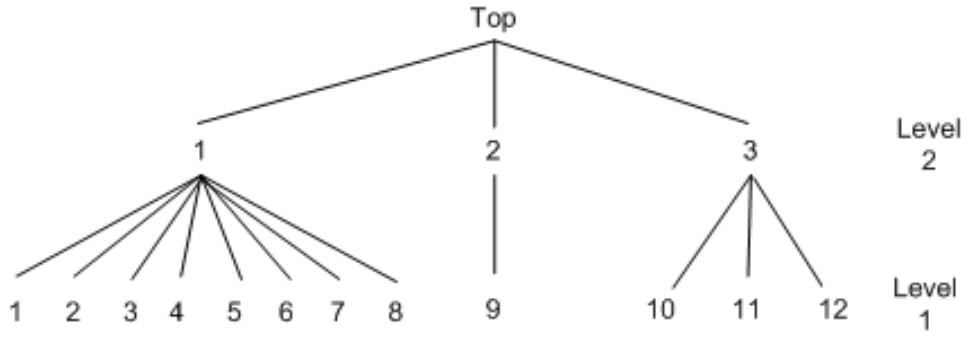

Figure 1. An example of the structure of a NL model.

It is not the case that the MNL model cannot be used as a multidimensional model. We can distinct two cases of multidimensional models: multidimensional choice sets with shared observed attributes and multidimensional choice sets with shared unobserved attributes. The multidimensional MNL model, called the joint logit (JL) model is an example of the former; the NL and MNP models are examples of the latter. Before it is able to discuss NL models properly, it is necessary to have a look at multidimensional choice sets in general first.

Multidimensional choice sets

In multidimensional choice theory, every decision process consists of more than one choice set. The most simple example is probably the scenario with two choice sets: C_1 and C_2. Both choice sets have J_1 and J_2 elements. Therefore the choice set $C_1 \times C_2 (Carthesian\ product)$ will consist of $J_1 \cdot J_2$ elements. The multinomial choice set for a decision maker *n* will be $C_n = C_1 \times C_2 - C_n^*$, where C_n^* is the set of elements that are not feasible for that decision maker. This obviously also goes for choice sets with higher levels of dimensionality. The example Ben-Akiva and Lerman give is a choice situation where **M** denotes possible modes for shopping and **D** denotes shopping destinations. The choice set becomes **M** $\times$ **D.** It is possible however to add extra dimensions as for example time of day or route.

The difference with multinomial choice sets is that elements are somewhat ordered, meaning that elements share common elements along one or more dimensions. This linkage between the elements makes analysis useful, because it implies that for a linkage to exist either some of the observed attributes of elements in the choice set me be equal across subsets of alternatives or this may be true for some of the unobserved attributes. Consequences of the former will result in the JL model; the latter will result in the NL model. Ben-Akiva and Lerman (1985) state that the results for multidimensional choice situations will not be different from multinomial situations, as long as elements of the choice set share either observed or unobserved attributes.

Nested logit

Train (2003) states that a nested logit model is appropriate when the choice set can be partitioned into subsets, or nests, in such a way that two properties hold. The first of these being that the IIA property holds within each nest. The second of these is that the IIA property does not hold in general for alternatives in different nests. The NL model can be derived from the General Extreme Value (GEV) model, but that will not be covered in this paper.

So when designing the model, one should note that the two properties should hold. A way to test this is by removing one of the alternatives from the choice set. If choice probabilities rise equally for certain alternatives, these would fit in one nest. Otherwise they would have to be in two different nests, because the IIA property does not hold. The IIA property should hold within each nest but not across nests.

The NL model is consistent with utility maximization (Daly and Zachary (1978), McFadden (1978) and Williams (1977)). Let the total choice set be partitioned in K non-overlapping subsets (nests) $B_1, \ldots, B_k$. The utility of alternative i in nest B_K is $U_{in} = V_{in} + \varepsilon_{in}$, again with V_{in} the observed part of the utility and ε_{in} the random, unobserved part. The NL model is obtained by assuming that the vector of disturbances has a cumulative distribution of a GEV type distribution:

$$\exp\left(-\sum_{k=1}^{K}\left(\sum_{j\in B_k} e^{-\varepsilon_{in}/\lambda_k}\right)^{\lambda_k}\right).$$

This is a generalization of the distribution that is underlying to the logit model. The unobserved utilities are correlated within nests. For any alternatives in different nests, the correlation is zero. In the function above the parameter λ_k is a measure of the degree of independence in the random part of the utility among the alternatives in nest k. A higher value causes greater independence and thus less correlation. When $\lambda_k = 1$ for all k, the GEV distribution becomes the product of extreme values terms that are independent, as $\lambda_k = 1$ represents independence among all alternatives. This means that the NL model reduces to the MNL model.

The distribution for the unobserved components proceed the choice probability for alternative i in nest B_k:

$$P_{in} = \frac{e^{V_{in}/\lambda_k}\left(\sum_{j\in B_k} e^{V_{jn}/\lambda_k}\right)^{\lambda_k - 1}}{\sum_{\ell=1}^{K}\left(\sum_{j\in B_\ell} e^{V_{jn}/\lambda_\ell}\right)^{\lambda_\ell}}.$$

If $k = \ell$, meaning two alternatives are in the same nest, the factors in parentheses cancel each other out and it shows that IIA holds. For $k \neq \ell$ the factors in parentheses do not cancel each other out and IIA does not hold. Train (2003) shows that across nests some other form of IIA holds, independence from irrelevant nests (IIN). Therefore in a NL model, IIA holds for alternatives within each nest and IIN holds over alternatives in different nests.

The parameter λ_k can be different within different nests, because correlation among unobserved factors can be different within different nests. A researcher can, however, constrain the λ_k's in different nests to be the same. This would indicate that the correlation is the same in each of these nests. Testing if this term is equal to 1 for all k would mean testing if the logit model were appropriate. For the model to be consistent with maximum-utility behaviour, the value of λ_k must be in some range. If λ_k lies between zero and one the model will be consistent with utility-maximizing behaviour. If λ_k is larger than one, the model is only consistent with this behaviour for some range of the explanatory variables but not for all values. A value smaller than zero is inconsistent with utility-maximizing behaviour. Kling and Herriges (1995) provide tests of consistency of NL with utility maximization in the case that $\lambda_k > 1$. In reality λ_k does not have to be a fixed parameter, as every decision maker has different correlations. Bhat

(1977) describes a way to calculate λ_k based on a vector of characteristics of the decision maker and a vector of parameters that need to be estimated.

The choice probability as given before is still quite a hard formula to grasp. It is possible to express this choice probability in a different fashion without loss of generality. The observed component of the utility function can be distinct in two parts:
$U_{in} = W_{nk} + Y_{in} + \varepsilon_{in}$.
Here W_{nk} is the part that is constant for all alternatives within a nest: this variable depends only on variables that describe nest *k*, therefore they differ over nests but not over the alternatives within a nest. Y_{in} depends on variables that describe alternative *i*, so they vary over alternatives within nest *k*. Finally ε_{in} is the unobserved part of the utility. Note that Y_{in} is simply defined as $V_{in} - W_{nk}$. Now the NL probability can be written as the product of two logit probabilities. The probability that an alternative is chosen can be written as the product of the probability that a certain nest is chosen multiplied with the probability that an alternative within that nest is chosen:
$P_{in} = P_{in|B_k} P_{nB_k}$,
where $P_{in|B_k}$ is the conditional probability that given an alternative in nest B_k is chosen an alternative *i* is chosen. P_{nB_k} is the marginal probability of choosing an alternative in nest B_k. Any probability can be written as the product of a marginal and conditional probability, therefore it is exactly the same as the situation before. Now both can take the form of logits:

$$P_{nB_k} = \frac{e^{W_{nk} + \lambda_k I_{nk}}}{\sum_{\ell=1}^{K} e^{W_{n\ell} + \lambda_\ell I_{n\ell}}},$$

$$P_{in|B_k} = \frac{e^{Y_{in}/\lambda_k}}{\sum_{j \in B_k} e^{Y_{jn}/\lambda_k}},$$

where

$$I_{nk} = \ln \sum_{i \in B_k} e^{Y_{in}/\lambda_k}.$$

These expressions are derived from the choice probabilities stated earlier. Train (2003) gives the derivation by algebraic rearrangement. It is customary to refer to the marginal probability as the upper model and to the conditional probability as the lower model. The quantity I_{nk} links the lower and upper model by transferring information from the lower model to the upper model (Ben-Akiva (1973)). This term is the logarithm of the denominator of the lower model, which means that $\lambda_k I_{nk}$ is the expected utility that the decision maker obtains from the choice among the alternatives in nest B_k. The formula for the expected utility is the same as the utility for logit, as the lower and upper model are both logit models. I_{nk} is often referred to as the inclusive utility of nest B_k. It includes the term W_{nk}, which is the utility the decision maker receives no matter what alternative in that nest he chooses. Added to that is the extra utility he or she obtains by choosing the alternative with highest utility in that nest, $\lambda_k I_{nk}$.

Estimation of nested logit
For the NL model the same applies as for the MNL: its parameters can be estimated by standard maximum likelihood techniques:

$$L = \prod_{n=1}^{N} \prod_{i \in B_k} \left(P_{in \setminus B_k} P_{nB_k} \right)^{y_{in}},$$

thus the log likelihood becomes:

$$\log L = \sum_{n=1}^{N} \sum_{i \in B_k} y_{in} \ln P_{in \setminus B_k} + \sum_{n=1}^{N} \sum_{k \in K} y_{nk} \ln P_{nB_k} .$$

Train describes that the NL model can also be estimated in a sequential fashion in a bottom up way: the lower models are estimated first. Then the inclusive utility is calculated for each lower model. Then the upper model is estimated with the inclusive utility as explanatory variables. However Train (2003) also describes that two difficulties come with using estimation in a sequential fashion. Firstly the standard errors of the upper model parameters are biased downward and some parameters appear in several sub models. Estimating the lower and upper model separately causes separate estimates of whatever common parameters appear in the model. Maximum likelihood estimation is conducted simultaneously for both models; therefore the common parameters are constrained to be the same wherever in the model they appear. It is stated that maximum likelihood estimation is the most efficient estimation technique for the NL model.

Higher-level nested logit and expansion on the NL model

Until now the NL model has been discussed at a two dimensional level, also known as a two-level nested logit. Here there are two levels of modelling: the marginal probabilities and the conditional probabilities. In some situations however, a higher-level NL model might be appropriate. The choice probabilities of three- or higher-level NL models can be expressed as a series of logit models. The top-level model describes the choice of a nest, then it describes choices of subnets to a certain extend and eventually the lowest level model is the choice of alternatives in a (sub) model. The top model includes an inclusive utility for each nest.

It also possible that an alternative is a member of more than one nest. The example that Train gives is an example of home-to-work travelling. Consider four alternatives: bus, train, drive alone and carpooling. Obviously bus and train can be considered public transport and driving alone and carpooling can be considered going by car. However, carpooling has some shared unobserved attributes that are similar to going by public transport: it has a lack of flexibility in scheduling. Now it is possible that this alternative is placed in both nests. This phenomenon is called overlapping nests. This model is called the Cross Nested Logit (CNL) model. Ben-Akiva and Bierlaire (1999) proposed the general formulation of the CNL model. The CNL model is also called the Generalized Nested Logit model (Wen and Koppelman (2001)).

Cross-Nested Logit

According to Papola (2003), the full specification consists of two phases: specification of the correlation structure and the identification of the parameters. Now both of these last variations (higher dimensions and overlap) can be included in the choice probabilities of the NL model, leading to CNL model (or Generalized NL model). The nests are labelled $B_1, B_2, \ldots, B_K$. Each alternative can be part of more than one nest and also to varying degrees. Therefore allocation parameter α_{ik} reflects the degree to which alternative *i* is a member of nest *k*. This parameter must be nonnegative and if it is zero it means it is not a member of nest *k.* The sum of over all nests for one alternative must be one. Again

we have parameter λ_k that indicates to what extent alternatives among a nest are independent and a higher value can be explained as greater independence and less correlation. Now the probability that decision maker *n* chooses alternative *i* is:

$$P_{in} = \frac{\sum_k \left(\alpha_{ik} e^{V_{in}}\right)^{1/\lambda_k} \left(\sum_{j \in B_k} \left(\alpha_{jk} e^{V_{jn}}\right)^{1/\lambda_k}\right)^{\lambda_k - 1}}{\sum_{\ell=1}^{K} \left(\sum_{j \in B_\ell} \left(\alpha_{j\ell} e^{V_{jn}}\right)^{1/\lambda_\ell}\right)^{\lambda_\ell}}.$$

Now if α_{ik} is equal to one for all alternatives in the choice set and enters only one nest, we get the same choice probability as for the two-level NL model. If λ_k is equal to one for all nests next to this, the model becomes a standard logit model. Also for higher-level, overlapping models it is possible to decompose the model:

$$P_{in} = \sum_k P_{in|B_k} P_{nk},$$

where marginal and conditional probabilities are:

$$P_{nk} = \frac{\sum_{j \in B_k} \left(\alpha_{jk} e^{V_{jn}}\right)^{1/\lambda_k}}{\sum_{\ell=1}^{K} \left(\sum_{j \in B_\ell} \left(\alpha_{j\ell} e^{V_{jn}}\right)^{1/\lambda_\ell}\right)^{\lambda_\ell}},$$

$$P_{in|B_k} = \frac{\left(\alpha_{ik} e^{V_{in}}\right)^{1/\lambda_k}}{\sum_{j \in B_k} \left(\alpha_{jk} e^{V_{jn}}\right)^{1/\lambda_k}}.$$

The model was first used by Vovsha (1997) who used the model for a mode choice survey in Israel. The model is appealing because it is able to capture a wide variety of correlation structures. The CNL model has a closed form, as it is derived form the GEV model and an expansion on the NL model. As shown before, it is in some ways a special case of the standard logit model. This makes the CNL model analytically interesting.

One of the most obvious merits is the ability to capture complex situations where the NL model cannot handle correlations, because it does not allow for overlap. Also the open form of the PDF makes the model analytically interesting.

A disadvantage of the model, maybe because it is quite a new model, is that the issue of identification still remains open. There are different estimation techniques and a maximum likelihood estimator can be identified. However if the model is over specified the speed of the algorithm may decrease significantly or not even perform well at all.

Estimation of cross-nested logit

The first estimation procedures for the CNL model were proposed by Small (1987) and Vovsha (1997) and are based on heuristics. However, currently maximum likelihood estimation techniques are used. Again, these techniques aim at identifying the set of parameters that maximize the probability that a given model perfectly reproduced the observations (Bierlare (2001)). The objective function of the maximum likelihood estimation problem for the CNL model is a nonlinear analytical function, as the PDF has a closed form. Most nonlinear programming algorithms are designed to identify local optima of the objective function. As the CNL model has a closed form, the log likelihood does as well:

$$\ln L = \sum_{n \in sample} \ln P_{in|C_n},$$

where this is the probability that alternative *i* is chosen by decision maker *n* and C_n is the choice set for that specific decision maker. Ben-Akiva and Bierlaire (1999) give a more elaborate derivation of the log likelihood function. Whatever algorithm is preferred, it is instrumental that different initial solutions are used. No meta-heuristics can provide a global optimum. Ben-Akiva and Bierlaire also give a number of steps that need to be taken. First, constraints to guarantee the model validity have to be defined. Then constraints imposing a correct intuitive interpretation could be important. Finally normalization constraints are necessary; otherwise the model would not be estimable.

Mixed Logit

According to McFadden and Train (2000), mixed logit (MXL) is a highly flexible that can approximate any random utility model. The limitations stated for the logit model are prevented because it allows for random taste variation, unrestricted patterns and correlation in unobserved factors over time. In contrast to the probit model, it is not restricted to normal distributions for the error terms and together with the probit model it has been known for years but has only become applicable since simulation become accessible.

MXL models can be derived under different behavioural specifications and each derivation provides a different interpretation. The MXL model is defined on basis of the functional form for its choice probabilities. MXL probabilities are the integrals of logit probabilities over a density of parameters:

$$P_{in} = \int L_{in}(\beta) f(\beta) d\beta,$$

here L_{in} is the logit probability evaluated at parameters β, as described in the chapter on MNL and $f(\beta)$ is a density function. If the utility is linear in β, then $V_{in}(\beta) = \beta' x_{in}$. Then the MXL probability takes its usual form:

$$P_{in} = \int \left(\frac{e^{\beta' x_{in}}}{\sum_j e^{\beta' x_{jn}}} \right) f(\beta) d\beta.$$

This probability is a weighted average of the logit formula, but evaluated at different values of β, with weights given by $f(\beta)$. In literature a mixed function is a weighted average of several functions and the density that provides the weights is a mixing distribution. MXL is a mix of the logit function evaluated at different β's with $f(\beta)$ as the mixing distribution. Standard logit is a specific case where the mixing distribution is fixed for $f(\beta) = 1$. Then the formula becomes the normal choice probability for the MNL model. The mixing distribution can also be discrete, with β taking a finite set of distinct values. This results in the latent class model that will be discussed later. However in most cases the mixing distribution is specified as continuous. By specifying the explanatory variables and density appropriately, it is possible to represent any utility maximizing (and even some forms of non-utility-maximizing behaviour) by a MXL model. There is one issue though concerning notation. There are two sets of parameters in a MXL model: the parameters β that enter the logit formula and have density $f(\beta)$ and the second set parameters that describe that density. Denote the parameters that describe the density of β by θ, so the density is best denoted as $f(\beta|\theta)$. The mixed logit choice probabilities are not dependent on the values of β, but they are functions of θ. The parameters β can be integrated out. In that sense the β's are similar to the

disturbances in the sense that they are both random terms that are integrated out in order to obtain the choice probability.

One of the positive points about the MXL model is that it exhibits neither the IIA property nor the restrictive substitution patterns of the logit model. The ratio P_{in}/P_{jn} depends on all data, including alternatives and attributes other than *i* or *j*. The percentage change in probability depends on the relation between L_{in} and L_{jn}, the logit probabilities of alternatives *i* and *j*. Besides this, Greene and Hensher (2002) state that the MXL model is flexible, even though it is fully parametric. Therefore it can provide the researcher with a large range to specify the individual and unobserved heterogeneity. To some extent this flexibility even offsets the specificity of the distributional assumptions. They also state that with the MXL model it is possible for the researcher to harvest a rich variety of information about behaviour from a panel or repeated measures data set.

Estimation

Train (2003) states that MXL is well suited for simulation methods for estimation. Utility is again $U_{in} = \beta_n^{'} x_{in} + \varepsilon_{in}$ and the β coefficients are distributed with the mixing distribution $f(\beta|\theta)$ and θ denotes the parameters of this distribution (mean and covariance of β). Now the choice probabilities are

$$P_{in} = \int L_{in}(\beta) f(\beta|\theta) d\beta,$$

$$L_{in}(\beta) = \frac{e^{\beta' x_{in}}}{\sum_{j=1}^{J} e^{\beta' x_{jn}}}.$$

Here L_{in} again is the logit probability. The probabilities P_{in} are approximated through simulation for any given value of θ. According to Train (2003) this is conducted through three steps:

1. Draw a value of β from the mixing distribution $f(\beta|\theta)$ and label it β' with the superscript $r = 1$. This refers to the first draw.
2. Calculate the logit formula $L_{in}(\beta^r)$ with this draw.
3. Repeat step 1 and 2 often and average the result. This average is the simulated probability:

$$\breve{P}_{in} = \frac{1}{R} \sum_{r=1}^{R} L_{in}(\beta^r),$$

where R is the number of draws. Here $\breve{P}_{in}$ is an unbiased estimator of P_{in}. Its variance decreases as R increases and it is strictly positive, twice differentiable in the parameters θ and the variables x that facilitate the numerical search for the maximum likelihood function. Train denotes this simulated log likelihood (SLL) function as follows:

$$SLL = \sum_{n=1}^{N} \sum_{j=1}^{J} d_{jn} \ln \breve{P}_{jn}.$$

Here d_{jn} is an indicator function, it is equal to one if decision maker *n* chose alternative *j* and zero otherwise. The maximum simulated likelihood estimator (MSLE) is the value of θ that maximizes the SLL function. The exact properties of this simulated estimator will not be discussed in this paper, but can be found in Train (2003) and Greene (2001).

This method of estimation is related to accept-reject (AR) methods of simulation. This AR method will not be discussed extensively, as the MSLE method is more often

used. AR simulation can be applied generally. It is constructed as follows (Train (2003), Greene (2001)):

1. A draw of the random terms is taken.
2. The utility of each alternative is calculated from this draw and the alternative with the highest utility is identified.
3. Steps 2 and 3 are repeated often.
4. The simulated probability for an alternative is calculated as the proportion of draws for which that alternative has the highest utility.

The AR simulator is also unbiased by construction. It is however not strictly positive. Also it is not always twice differentiable. It seems to be some sort of step function. Therefore the MSLE method is more often used.

Repeated choice
Each sampled decision maker easily generalizes this model for repeated choices. The simplest specification treats the coefficients that enter utility as varying over different decision makers but constant over choice situations for each decision maker. Therefore utility from alternative *i* in choice situation *t* by decision maker *n* is $U_{itn} = \beta_n x_{itn} + \varepsilon_{itn}$, with ε_{itn} being iid extreme value over alternatives, time and decision makers. Now consider a sequence of alternative, one for each time period: $\mathbf{i} = \{i_1, \ldots, i_T\}$. The probability that the decision maker makes this sequence of choices is the product of the logit formulas, conditional on β:

$$L_{in}(\beta) = \prod_{t=1}^{T}\left[\frac{e^{\beta_n' x_{i_t tn}}}{\sum_j \beta_n' x_{jtn}}\right].$$

The ε_{itn} terms are independent over time. The unconditional probability is the integral of this product over all values of β:

$$P_{in} = \int L_{in}(\beta) f(\beta) d\beta.$$

The probability is simulated in a similar way to the probability with just one choice period:

1. A draw of β is taken from its distribution.
2. The logit formula is calculated for each period and the product of the logits is taken.
3. Repeat steps 1 and 2 often.
4. The MSLE is the average of the simulated probabilities.

Latent Class Logit
As stated in the last chapter, the MXL model and the Latent Class Logit (LCL) model are not that different. Remember the usual form that the mixed logit probability takes:

$$P_{in} = \int \left(\frac{e^{\beta' x_{in}}}{\sum_j e^{\beta' x_{jn}}}\right) f(\beta) d\beta \,.$$

This choice probability is the same for the LCL model, but instead of $f(\beta)$ being continuous, the mixing distribution is discrete with β taking a finite set of distinct values.

Here β takes M possible values labelled $b_1, \dots, b_M$ with probability s_m that $\beta = b_m$. This model has been popular in psychology and marketing for some time now (Kamakura and Russell (1989) and Chintagunta et al. (1991)).

Greene and Hensher (2002) state that the LCL model is similar to the MXL model by McFadden and Train (2001), but it relaxes the requirement that the researcher makes specific assumptions about the (continuous) distributions of parameters across each decision maker. Sagebiel (2011) states that the unconditional probability to choose a certain alternative i by decision maker n is the weighted average of the M b_m parameters:

$$P_{in} = \sum_{m=1}^{M} s_m P_{in|m},$$

with $P_{in|m}$ being the conditional logit (CL) probability to choose alternative i that belongs to class s:

$$P_{in|m} = \frac{e^{(b_{1n}x_{in}+\dots+b_{kn}x_{kn})}}{\sum_{n=1}^{N} e^{(b_{1n}x_{in}+\dots+b_{kn}x_{kn})}}.$$

The parameters h_s are unknown, but can be estimated with a MNL model with help of a case-specific variable vector that includes age, income or any other characteristic of the decision maker.

The underlying theory of the LCL model presumes that individual choice behaviour depends on the observable attributes and on latent heterogeneity that varies with the unobserved factors (Greene and Hensher (2002)). They analyse this heterogeneity through a model of discrete parameter variation. Here decision makers, as in the analysis by Sagebiel (2011), are sorted in a set of Q classes, but it is unknown to the researcher which class contains which decision makers, even if the decision makers themselves know it. There are b_n alternatives by decision maker n in T_n choice situations. As is clear, Greene and Hensher directly explain the model for repeated choice. Here the choice probability that alternative i is made by decision maker n is choice situation t, given that this occurs in class q:

$$P_{itn|q} = \frac{e^{x_{itn}'\beta_q}}{\sum_{j=1}^{J_n} e^{x_{itn}'\beta_q}} = F(i,t,n \mid q).$$

The size of the choice set and the number of observations may vary per decision maker. Moreover, the choice set should vary per choice situation as well. The probability for the specific choice by a decision maker can be formulated in different manners, so for convenience can be formulated in different manners, so for convenience y_{tn} denotes the specific choice made, so the model provides

$$P_{tn|q}(i) = \Pr(y_{tn} = i \mid class = q).$$

Given the class assignment, it is assumed that T_n are independent events and the contribution of decision maker n to the likelihood would be the joint probability of the sequence $y_i = [y_{1n}, \dots, y_{Tn}]$:

$$P_{n|q} = \prod_{t=1}^{T_n} P_{tn|q}.$$

Here the class assignment is unknown. Now H_{nq} is denoted as the prior probability for class q for decision maker n. For H_{nq} the form of the MNL is used:

$$H_{nq} = \frac{e^{z_n'\theta_q}}{\sum_{q=1}^{Q} e^{z_n'\theta_q}}, q = 1, \dots, Q, \theta_Q = 0,$$

where z_n is a set of observable characteristics, as described in the chapter on binary choice models. θ_q again denote the parameters for class *q.*

Greene and Hensher (2002) compared the MXL model and the LCL model for a transport situation, with the objective to seek understanding of the relative merits of both modelling strategies. They state that it is not possible to state that one of these models is unambiguously preferred to the other. The LCL model has the advantage that it is a semi parametric specification. This means that the researcher is freed from possibly strong or unwarranted distributional assumptions, in their case about individual heterogeneity. For the LCL model the same applies as for the MXL model, it allows the researcher to harvest a rich variety of information on behaviour from a panel or repeated measures data set.

Estimation

Now the probability, or now the likelihood is the expectation over classes of the class specific contributions:

$$P_n = \sum_{q=1}^{Q} H_{nq} P_{n|q} .$$

This is the likelihood for one decision maker. Now the log likelihood for the sample is

$$\ln L = \sum_{n=1}^{N} P_n = \sum_{n=1}^{N} \ln \left[\sum_{q=1}^{Q} H_{nq} \left(\prod_{t=1}^{T_n} P_{tn|q} \right) \right].$$

Maximization of this function with respect to the structural parameters and the latent class parameter vectors is a conventional problem in maximum likelihood estimation. Greene (2001) discusses the mechanics and other aspects of estimation, as it is a relatively difficult optimization problem in comparison to other situations. The choice of a good starting value for *Q* is crucial. Testing down to the appropriate *Q* by comparing the log likelihoods of smaller models is not a proper approach. Roeder et al. (1999) suggest that the Bayesian Information Criterion (BIC) model can be used:

$$\text{BIC(model)} = \ln L + \frac{(model\ size)\ln N}{N} .$$

Now we now parameter estimates θ_q, the prior estimates of the class probabilities are $\hat{H}_{nq}$. Using Bayes theorem (Bayes and Price, 1763) it is possible to obtain a posterior estimate of the latent class probabilities:

$$\hat{H}_{q|n} = \frac{\hat{P}_{n|q} \hat{H}_{nq}}{\sum_{q=1}^{Q} \hat{P}_{n|q} \hat{H}_{nq}} .$$

The maximum value of this formula would be associated by a strictly empirical estimator of the latent class within the class that the individual resides in. These results can be used to obtain posterior estimates of the individual specific parameter vector:

$$\hat{\beta}_n = \sum_{q=1}^{Q} \hat{H}_{q|n} \hat{\beta}_q .$$

Also this data can be used to estimate marginal effects in the logit model.

Summary, comments and acknowledgement

"I abhor averages. I like the individual case. A man may have six meals one day and none the next, making an average of three meals per day, but that is not a good way to live."

Louis D. Brandeis (1856 – 1941)

Variations: different choice models

Until now, the most instrumental models have been discussed. However, this does not mean that there are no other choice models known. The models discussed are mostly models that are used very often and are instrumental in the development of new models. For example, the standard logit models are not used that often anymore, but serve more as reference models. The MNL model however, is still used in some fields. Dow and Endersby (2003) argue that in political voting research the MNL model is still usable, as it is easily understandable and the differences in results are very small and computationally it can be preferred over more complex models.

In this chapter some other models will be discussed briefly. The reader should be aware that the models discussed until now are not the only models and in different areas different models are used. Models as for example the MNP model are much used, but are analytically more complex and did not fit in the line of this paper. These models are worth mentioning however.

The Generalized Extreme Value (GEV) Model

This model has been mentioned before in this paper. The GEV model is actually more a class of models than a specific model. The MNL model and the NL model are both models that are part of the GEV family of models. McFadden (1978) gave a generalization of the MNL model that is in fact the GEV model. The unifying attribute of the models of the GEV family is that the unobserved components of the utility for all alternatives are jointly distributed as a generalized extreme value. This allows for correlations over alternatives. When all these correlations become zero, the GEV model becomes the standard logit model. Currently the most widely used member of the family are the NL and CNL models. These models are applied in for example energy, transportation, housing and telecommunication. Karlstrom (2001) showed that only a small portion of the possible models within the GEV classes has ever been implemented. This means that the full capability of this class of models has not yet been fully exploited and new research can be used to further investigate the potential of this group of models.

Joint logit

Regarding multidimensional choice mostly the NL and CNL models have been regarded, models that are based on multidimensional choice sets with shared unobserved attributes. The Joint logit (JL) model on the other hand is a model based on multidimensional choice sets with shared observed attributes. Here there are observed factors in the choice set that are shared, therefore overlap in some sense. As in the MXL model, marginal and conditional probabilities are derived. The distribution of the disturbances will affect the form of the conditional choice probabilities. A normal distribution is not always guaranteed. Currently joint models are mostly used for combining stated and revealed preference data.

Multinomial Probit

As the MNL model is an extension of the binary logit model, exactly that way is the MNP model an extension of the binary probit model. Now the vector of disturbances is extended from two to J_n. This requires for the model to have a solution of the $J_n - 1$ dimensional integral for evaluating the choice probabilities.

The concept of this model appeared a very long time ago in writings by Thurstone (1927) in applications of mathematical psychology. Because of the open form of the

integral in evaluating the choice probability, until early 1980s its computational difficulty made it unusable. Since the rise of simulation it has been possible to use the MNP model more extensively. Nowadays the MNP model is used in different areas. Dow and Endersby (2003) indicate that it is often used in political voting and in psychology (Train (2003)).

As for estimation, this is a bit more complicated than for the MNL model, as indicated before. Because the choice probabilities have an open form integral, simple estimation techniques are often insufficient. Geweke (1996) gives an explanation of quadrature methods, which approximate the integral by a weighted function of specially chosen evaluation points. However nowadays not many non-simulation techniques are used, for example because it is more general. Hajivassiliou et al. (1996) give an overview of different simulation techniques. The most straightforward and most used techniques are accept-reject (AR), which was briefly noted in the chapter on the MXL model, smoothed AR and GHK (after Geweke, Hajivassiliou and Keane) that can be combined with maximum likelihood or sampling. Train (2003) describes these models more elaborately.

Mixed Probit

Again mixed probit (MXP) can be seen as some development on the probit model, as was a similar case for MXL and MNL. A constraint for the MNP is that all random terms enter utility in a linear way and they are randomly distributed in a way so that the utility is normally distributed. This constraint is removed in MXP models. These models have, as do the MXL models conditional probabilities and the integral also has an open form. Therefore the model has long run times for the GHK simulator.

Train (2003) states that the MXP model provides a way to avoid some of the practical difficulties of the MXL model, as representing pure heteroskedasticity or fixed correlation patterns among alternatives. In the MXP estimation is made instead of specifying numerous error components in the MXL model. MXP is suitable some non-normal random terms, which is not possible in the MNP model. However MXP is more complex and computationally more burdensome.

Further research

As stated before, new models will continue to arise now possibilities have opened up with the advent of simulation. In 2002 Ben-Akiva et al. wrote a paper on hybrid choice models. Burda et al. (2008) for example present a Bayesian mixed logit-probit model for multinomial choice. It is expected that in the coming years more of the hybrid models will surface as computationally more will become possible.

Summary

In this paper the reader is given an overview of theory behind choice models. Choice models have many applications in society: psychology, transport, energy, housing, marketing, voting and many more areas make use of these models. Since the 1920s, when Thurstone (1927) introduced the binary logit and probit model, these models have developed significantly.

Firstly the theory behind the models, choice theory, probabilistic theory and utility theory have been discussed. Most models have the assumption of rational behaviour and utility-maximizing behaviour. Also the framework for binary models was given.

In the second part of the paper different models have been discussed. Binary logit and probit models can be considered predecessors of multinomial logit and probit models. In a similar way are nested logit and mixed logit models derived from the multinomial logit model. It is still discernable that the multinomial logit model is a special case of both the nested logit model and the mixed logit model, as with the right parameters the model is equal to the multinomial model. For the nested logit model it applies that both the marginal and conditional probabilities are logits. Also the cross-nested logit model can be seen as an extension of the nested logit model, as it allows for overlap between nests. The latent class model is derived from the general extreme value model, as applies for the mixed logit model. The difference between these models is the assumptions underlying and mostly the difference in mixing distribution. In this part for each model also estimation techniques were discussed. For two models that were discussed last, the latent class model and the mixed logit model, simulation techniques are necessary because of the open form of the integral in the choice probability. This makes these models more complex and computationally more burdensome. These models however do avoid some of the assumptions underlying at for example the multinomial model.

In the third and last section a short overview was given of models that were not discussed elaborately in the second section, but were worth mentioning due to their value and use in research. The general extreme value model is often referenced and the multinomial probit model is still very much used.

Acknowledgement

My sincere thanks goes out to Alwin Haensel, who besides helping me with my questions put me on the path to writing this paper by making this subject public on the website of the VU.

References

Ai C. and Norton E.C. (2003), Interaction terms in logit and probit models

Alvarez R.M. and Nagler J. (1995), Economics, Issues and the Perot Candidacy: Voter Choice in the 1992 Presidential Election

Alvarez R.M. and Nagler J. (2001), Correlated disturbances in discrete choice models: a comparison of multinomial probit and logit models

Adamowicz W., Louviere J. and Williams M. (1994), Combining Revealed and Stated Preference Methods for Valuing Environmental Amenities – *Journal of Environmental Economics and Management*

Anand P. (1993), Foundations of rational choice under risk

Bayes T. and Price, M. (1763), An Essay towards solving a problem in the doctrine of chances – *Philosophical Transactions of the Royal Society of London*

Ben-Akiva M.E. and Lerman S.R. (1985), Discrete Choice Analysis: Theory and Application to Travel Demand – *MIT Press*

Ben-Akiva M. E. and Bierlaire, M. (1999), Discrete choice methods and their applications to short-term travel decisions – *Handbook of Transportation Science*

Ben-Akiva, M. E., McFadden D., Train K., Walker J., Bhat C., Bierlaire M., Bolduc D., Boersch-Supan A., Brownstone D., Bunch D.S., Daly A., De Palma A., Gopinath D., Karlstrom A. and Munizaga M. (2002), Hybrid Choice Models: Progress and Challenges

Bierlaire M. (2001), A theoretical analysis of the cross-nested logit model

Collins L.M., and Lanza S.T. (2010), Latent class and latent transition analysis for the social, behavioural and health sciences

Cox D.R. (1970), Analysis of Binary Data

Dow J.K. and Endersby, J.W. (2003), Multinomial probit and multinomial logit: a comparison of choice models for voting research

Geweke J. (1996), Monte Carlo simulation and numerical integration – *Handbook of Computational Economics, Elsevier Science*

Glasgow G. (2001), Mixed logit models for multiparty elections

Greene W. (2001), Fixed and Random Effects in Nonlinear Models – *Working paper Stern School of Business, Department of Economics*

Greene W.H. and Hensher D.A. (2003), A latent class model for discrete choice analysis: contrasts with mixed logit

Hajivassiliou V., McFadden D. and Ruud P. (1996), Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results – *Journal of Econometrics*

Hensher D.A., Rose J.M. and Greene W.H. (2005), Applied Choice Analysis: A Primer

Karlstrom A. (2001), Developing generalized extreme value models using the Piekands representation theorem – *Working paper Infrastructure and Planning – Royal Institute of Technology Stockholm*

Kreps D.M. (1988), Notes on the theory of choice

Kroes E.P. and Sheldon R.J. (1988), Stated Preference Methods – *Journal of Transport Economics and Policy*

Luce, D. (1959), Individual Choice Behavior: A Theoretical Analysis

McCutcheon, A. L. (1987), Latent class analysis

McFadden, D. (1978), Modelling the choice of residential location – *Spatial interaction theory and residential location*

McFadden D. and Train K. (2000), Mixed MNL models for discrete response – *Journal of Applied Econometrics*

Papola A. (2003), Some developments on the cross-nested logit model - *Elsevier*

Stevens T. H. (2005), Can Stated Preference Valuations Help Improve Environmental Decision Making? – *Choices Magazine, A publication of the American Agricultural Economics Association*

Train K. (2003), Discrete Choice Methods with Simulation

Thurstone L. L. (1927), A law of comparative judgment – *Psychological Review*

Varian H.R. (2005), Revealed Preference

Vovsha, P. (1997), Cross-nested logit model: an application to mode choice in the Tel-Aviv metropolitan area

Wassenaar H.J. and Chen W. (2003), An Approach to decision-based design with discrete choice analysis for demand modeling

Wardman M. (1988) A comparison of revealed preference and stated preference models of travel behavior – *Journal of transport economics and policy*

Wooldridge J.M. (2002), Econometric Analysis of Cross Section and Panel Data, *MIT Press*