Probabilistic Models for Qualitative Choice Behavior

Handout

by

John K. Dagsvik

Statistics Norway

P.O.Box 8131, Dep.

N-0033 Oslo

Norway

Email: john.dagsvik@ssb.no

Contents

1. Introduction

The traditional theory for individual choice behavior, such as it usually is presented in textbooks of consumer theory, presupposes that the goods offered in the market are infinitely divisible. However, many important economic decisions involve choice among qualitative—or discrete alternatives. Examples are choice among transportation alternatives, labor force participation, family size, residential location, type and level of education, brand of automobile, etc. In transportation analyses, for example, one is typically interested in estimating price and income elasticities to evalutate the effect from changes in alternative-specific attributes such as fuel prices and user-cost for automobiles. In addition, it is of interest to be able to predict the changes in the aggregate distribution of commuters that follow from introducing a new transportation alternative, or closing down an old one.

The set of alternatives may be “structurally” discrete or only “observationally” discrete. The set of feasible transportation alternatives is an example of a structurally categorical setting while different levels of labor supply such as “part time”, and “full time” employment may be interpreted as only observationally discrete since the underlying set of feasible alternatives, “hours of work”, is a continuum.

In several applications the interest is to model choice behavior for so-called discrete/continuous settings. Typical examples of phenomena where the response is discrete/continuous are variants of consumer demand models with corner solutions. Here the discrete choice consists in whether or not to purchase a positive quantity of a specific commodity, and the continuous choice is how much to purchase, given that the discrete decision is to purchase a positive amount. Another type of application is the demand for durables combined with the intensity of use. For example, a consumer that purchases an automobile has preferences over the intensity of use, and a household that purchases an electric appliance is also concerned with the intensity of use of the equipment.

The recent theory of probabilistic, or discrete/continuous choice is designed to model these kind of choice settings, and to provide the corresponding econometric methodology for empirical analyses. Due to variables that are unobservable to the econometrician (and possibly also to the individual agents themselves), the observations from a sample of agents' discrete choices can be viewed as outcomes generated by a stochastic model. Statistically, these observations can be considered as outcomes of multinomial experiments, since the alternatives typically are mutually exclusive. In the context of choice behavior, the probabilities in the multinomial model are to be interpreted as the probability of choosing the respective alternatives (choice probabilities), and the purpose of the theory of discrete choice is to provide a structure of the probabilities that can be justified from behavioral arguments. Specifically, one is, analogously to the standard textbook theory

of consumer behavior, interested in expressing the choice probabilities as functions of the agents' preferences and the choice constraints. The choice constraints are represented by the usual economic budget constraint and in addition, the choice set (possibly individual specific), which is the set of alternatives that are feasible to the agent. For example, in transportation modelling some commuters may have access to railway transportation while others may not.

In the last 25 years there has been an almost explosive development in the theoretical and methodological literature within the field of discrete choice. Originally, much of the theory was develop by psychologists, and it was not until the mid-sixties that economists started to adopt and adjust the theory with the purpose of analyzing discrete choice problems. In the present compendium we shall discuss central parts of the theory of discrete/continuous choice as well as some of the econometric methods that apply.

There exist by now a few textbooks that only consider discrete and discrete/continuous choice, such as Maddala (1983), Train (1986), Ben Akiva and Lerman (1985), and Train (2003). There are also several good survey articles, such as Amemiya (1981) and McFadden (1984), to mention just a few. Dagsvik (1985,) are two survey articles in Norwegian. In addition several textbooks contain one or several chapters on discrete and discrete/continuous econometric models. See for example Amemiya (1985, ch. 9, 10), Cameron and Trivedi (2005, ch. 14-16), Greene (1993, ch. 21, 22), Lattin, Carroll and Green (2003, ch. 13), Wooldridge (2002, ch.15, 16). In contrast to standard textbooks and surveys in econometric modeling of discrete choice such as Maddala (1983), Train (1986), Amemiya (1981), McFadden (1984) and Ben-Akiva and Lerman (1985), the focus of the present treatment is more on the theoretical developments than on statistical methodology. The reason for this is two-fold. First, it is believed that it is of substantial interest to bring forward some of the recent theoretical results that otherwise would not be easily accessible for the non-expert student. Second, the statistical methodology for estimation, testing and diagnostic analysis is rather well covered by the textbooks and surveys mentioned above.

This survey is organized as follows: In Section 2 I give a brief overview of reduced form type specifications of models with discrete response. In Section 3 I discuss some important elements of probabilistic choice theory, and in Section 4 I discuss the modeling of a few selected applications of discrete choice analysis. In Section 5 the estimation and testing based on the maximum likelihood method are discussed. In Section 6 I consider briefly the specification and estimation of Tobit models (nonstructural).

2. Statistical analysis when the dependent variable is discrete

As mentioned in the introduction there are many interesting phenomenons that naturally can be modelled with a dependent variable being qualitative (discrete) or where the dependent variable may be both discrete and continuous.

While most of the subsequent chapters will discuss theoretical aspects of discrete/continuous choice, we shall in this chapter give a brief summary of the most common statistical models which are useful for analyzing phenomena when the dependent variable is discrete, without assuming that the underlying response variables necessarily are generated by agents that make decisions. A more detailed exposition is found in Maddala (1983), chapter one and two. However, the statistical methodology we discuss is of relevance for estimating the choice models for agents (consumers, firms, workers, etc.), and will be further discussed in subsequent chapters.

2.1. Models for binary outcomes

In this section we shall consider models where the dependent variable is a *Binomial* variable. Recall that in statistics, the Binomial model is designed to represent random "experiments" in which the outcomes are independent across experiments, and in each experiment there are only two outcomes; either an event occurs or the event does not occur. For example, our experiment may consist in drawing independently a sample of n individuals and recording the labor force status of each of them ("participation" or "not participation"). Thus, we may represent the outcome in this case by a dummy variable Y_i, defined by

$$Y_i = \begin{cases} 1 & \text{if individual i participates in the labor market} \\ 0 & \text{otherwise}. \end{cases}$$

In the general case, Y_i equals one if a particular event—or outcome in question occurs, and zero otherwise. We may write

$$Y_i = E\,Y_i + \eta_i \tag{2.1}$$

where η_i is a random error term with zero mean. Since Y_i is a dummy variable with only two outcomes, it follows that

$$E\,Y_i = \sum_{y \geq 0} y\,P\left(Y_i = y\right) = 0 \cdot P\left(Y_i = 0\right) + 1 \cdot P\left(Y_i = 1\right) = P\left(Y_i = 1\right). \tag{2.2}$$

Thus, in this case EY_i has the interpretation as the probability that $Y_i = 1$. In general EY_i will depend on an exogenous variable just as in the classical regression model considered above. Let $\mathbf{X}_i$ denote a vector of exogenous variable and assume that

(2.3) $$E\left(Y_i \mid \mathbf{X}_i\right) = h\left(\mathbf{X}_i, \beta\right)$$

where $h\left(\mathbf{X}_i, \beta\right)$ is a function of $\mathbf{X}_i$, that is fully specified apart from a vector of unknown parameters, β. Hence we can in the general case write

(2.4) $$Y_i = h\left(\mathbf{X}_i, \beta\right) + \eta_i .$$

Assumption (2.3) implies that

(2.5) $$E\left(\varepsilon_i \mid \mathbf{X}_i\right) = 0 .$$

Also, due to the fact that the dependent variable is binary, we obtain

(2.6)

$$Var\left(\varepsilon_i \mid \mathbf{X}_i\right) = Var\left(Y_i \mid \mathbf{X}_i\right) = E\left(Y_i^2 \mid \mathbf{X}_i\right) - \left(E\left(Y_i \mid \mathbf{X}_i\right)\right)^2 = E\left(Y_i \mid \mathbf{X}_i\right) - \left(E\left(Y_i \mid \mathbf{X}_i\right)\right)^2 = h\left(\mathbf{X}_i, \beta\right) - h\left(\mathbf{X}_i, \beta\right)^2 .$$

Consequently, the model (2.4) differs from the classical regression model above in that

(2.7) $$0 \le h\left(\mathbf{X}_i, \beta\right) \le 1$$

and that the conditional variance (2.6) is a function of the conditional mean of Y_i expressed in (2.3). The restriction in (2.7) follows from the fact that similarly to (2.2), $h\left(\mathbf{X}_i, \beta\right)$ has the interpretation as a conditional probability, namely

(2.8) $$h\left(\mathbf{X}_i, \beta\right) = P\left(Y_i = 1 \mid \mathbf{X}_i\right).$$

It is therefore problematic to specify $h\left(\mathbf{X}_i, \beta\right)$ as a linear function in $\mathbf{X}_i$ because a linear specification will not necessarily satisfy (2.7), and consequently we may risk to get predictions from the model that are negative, or greater than one. This is the reason why the linear specification is seldom used in settings with discrete dependent variables. (Linear probabiliy model.) Instead it is common to specify $h\left(\mathbf{X}_i, \beta\right)$ as

(2.9) $$h\left(\mathbf{X}_i, \beta\right) = F\left(\mathbf{X}_i \beta\right)$$

where F(y) is an increasing function in y that satisfies $0 \le F(y) \le 1,$ and

$$\mathbf{X}_i \boldsymbol{\beta} = \beta_0 + \sum_{k=1}^{m} X_{ik} \beta_k . \tag{2.10}$$

Thus, apart from the nonlinear transformation, F(·), (2.9) has the structure of a linear regression model, and the unknown parameter vector equals the "regression" coefficients, $\boldsymbol{\beta}$.

The binary Probit model

In the Probit model F(y) is equal to the standard cumulative Normal distribution function, i.e.,

$$F(y) = \Phi(y) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-x^2/2}\, dx. \tag{2.11}$$

The binary Logit model

In the Logit model F(y) is equal to the cumulative Logistic distribution function,

$$F(y) = \frac{1}{1+e^{-y}}. \tag{2.12}$$

Clearly, $0 \le F(y) \le 1,$ since F(y) is increasing, $F(y) \to 1$ when $y \to \infty$ and $F(y) \to 0$ when $y \to -\infty$. It turns out that unless the explanatory variables take extreme values, the Logit and the Probit models are almost indistinguishable.

Example 2.1

Consider again the modelling of labor force participation. In this case the vector **X** is often assumed to contain variables such as age, marital status, number of small children, education. If one could estimate the unknown parameters of the model one would for example be possible to assess the marginal effect of education on labor force participation.

2.2. Estimation

The maximum likelihood method (MLE)

The maximum likelihood method is the most common method although it is possible to use other methods. Assume now that the model is given by (2.9). Suppose we have a sample of n observations. Then, conditional on the exogenous variables $(\mathbf{X}_i)$, the likelihood of the observations equal

$$L(\boldsymbol{\beta}) = \prod_{i \in S_1} F(\mathbf{X}_i \boldsymbol{\beta}) \cdot \prod_{i \in S_0} \left(1 - F(\mathbf{X}_i \boldsymbol{\beta})\right) \tag{2.13}$$

where S_1 is the subsample for which $Y_i = 1, i \in S_1$, while S_0 is the subsample for which $Y_i = 0, i \in S_0$. Thus the loglikelihood can be written as

$$\ln L(\boldsymbol{\beta}) = \sum_{i \in S_1} \ln F(\mathbf{X}_i \boldsymbol{\beta}) + \sum_{i \in S_0} \ln\left(1 - F(\mathbf{X}_i \boldsymbol{\beta})\right). \tag{2.14}$$

Alternatively, (2.14) can be expressed as

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \ln F(\mathbf{X}_i \boldsymbol{\beta}) + \sum_{i=1}^{n} (1 - Y_i) \ln\left(1 - F(\mathbf{X}_i \boldsymbol{\beta})\right). \tag{2.15}$$ (10.27)

From (2.15) we obtain that

(2.16)

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^{n} \frac{Y_i F'(\mathbf{X}_i \boldsymbol{\beta}) X_{ik}}{F(\mathbf{X}_i \boldsymbol{\beta})} - \sum_{i=1}^{n} \frac{(1 - Y_i) F'(\mathbf{X}_i \boldsymbol{\beta}) X_{ik}}{1 - F(\mathbf{X}_i \boldsymbol{\beta})} = \sum_{i=1}^{n} \frac{\left(Y_i - F(\mathbf{X}_i \boldsymbol{\beta})\right) F'(\mathbf{X}_i \boldsymbol{\beta}) X_{ik}}{F(\mathbf{X}_i \boldsymbol{\beta})\left(1 - F(\mathbf{X}_i \boldsymbol{\beta})\right)},$$

for $k = 0,1,\ldots,m$. Therefore, the maximum likelihood estimator, $\hat{\boldsymbol{\beta}}$, is determined by

$$\sum_{i=1}^{n} \frac{\left(Y_i - F\left(\mathbf{X}_i \hat{\boldsymbol{\beta}}\right)\right) F'\left(\mathbf{X}_i \hat{\boldsymbol{\beta}}\right) X_{ik}}{F\left(\mathbf{X}_i \hat{\boldsymbol{\beta}}\right)\left(1 - F\left(\mathbf{X}_i \hat{\boldsymbol{\beta}}\right)\right)} = 0, \tag{2.17}$$

for $k = 0,1,\ldots,m,$ where $X_{i0} = 1$. The system of equation (2.17) must of course be solved for $\hat{\boldsymbol{\beta}}$ by iteration methods. If the model is a Logit model where F is given by (2.12) then (2.17) reduces to

$$\sum_{i=1}^{n} \left(Y_i - \frac{1}{1 + \exp\left(-\mathbf{X}_i \hat{\boldsymbol{\beta}}\right)} \right) X_{ik} = 0 \tag{2.18}$$

for $k = 0,1,\ldots,m$.

Also (2.18) is nonlinear in $\hat{\boldsymbol{\beta}}$, and must similarly to the general case (2.17) be solved by iteration methods. It can be demonstrated that for the Probit and the Logit models the loglikelihood

function is globaly concave and consequently a unique maximum of the likelihood function is guaranteed.

The MLE has the following main properties:

(i) it is consistent, i.e. $\operatorname{p\lim}_{n\to\infty}\hat{\beta}=\beta$

(ii) it is asymptotically efficient, i.e. it attains the smallest variance among all consistent, asymptotically normal estimators

(iii) it is asymptotically normally distributed according to:

(2.19) $$\sqrt{n}\left(\hat{\beta}-\beta\right)\sim N(0,V)$$

where V is the asymptotic covariance matrix.

The covariance matrix V is determined by the likelihood function. It is equal to

(2.20) $$V=\left(-E\left\{\frac{\partial^2 \ln L(\beta)}{\partial\beta\partial\beta'}\right\}\right)^{-1}$$

where

$$\frac{\partial^2 \ln L(\beta)}{\partial\beta\partial\beta'}$$

means the covariance matrix with elements

$$\left\{\frac{\partial^2 \ln L(\beta)}{\partial\beta_i\partial\beta_j}\right\}.$$

Thus,

$$\text{Asympt. Var}\,\hat{\beta} \quad = \quad V/n.$$

In practice the covariance matrix V can be estimated consistently by replacing the expectation operator by the sample average and the unknown β-coefficients by their ML estimators.

Finally, the MLE is asymptotically efficient because it attains asymptotically the so-called Cramér-Rao lower bound.

When we apply the above model to some data set, the computer program will estimate the unknown β's by ML. Usually these programs will also give the *t*-values for each parameter. Hence, simple hypotheses can be tested in the "usual way". If we wish to test more composite hypotheses we have to resort to test procedures like Wald's test or the Likelihood ratio test.

2.3. Binary random utility models

Often the model with a discrete dependent variable is derived from a random utility representation. That is, to each alternative in a choice setting is associated a random index which represents the utility of the alternative. Specifically, assume that the individual decision-maker faces a choice set consisting of two alternatives, indexed by zero and one, respectively. Let U_{ij} be the individual i's utility of alternative j, $j = 0,1$. Assume that

(2.21) $$U_{ij} = v\left(\mathbf{X}_{ij}, \boldsymbol{\theta}\right) + \varepsilon_{ij}$$

where $v\left(\mathbf{X}_{ij}, \boldsymbol{\theta}\right)$ is a deterministic term that may depend on explanatory variables $\mathbf{X}_{ij}$, an unknown vector of parameters $\boldsymbol{\theta}$, and ε_{ij} is a random term. A utility-maximizing individual i will choose alternative j if $U_{ij} = \max\left(U_{i1}, U_{i2}\right)$ which means that

(2.22) $$Y_i = \begin{cases} 1 & \text{if} \quad U_{i1} > U_{i0} \\ 0 & \text{if} \quad U_{i1} < U_{i0} \end{cases}.$$

Let F(y) be the cumulative distribution function of $\varepsilon_{i0} - \varepsilon_{i1}$, i.e.

(2.23) $$F(y) = P\left(\varepsilon_{i0} - \varepsilon_{i1} \le y\right).$$

Then it follows that

(2.24)
$$E\left(Y_i \middle| \mathbf{X}_{i1}, \mathbf{X}_{i0}\right) = P\left(U_{i1} > U_{i0} \middle| \mathbf{X}_{i1}, \mathbf{X}_{i0}\right) = P\left(\varepsilon_{i0} - \varepsilon_{i1} < v\left(\mathbf{X}_{i1}, \boldsymbol{\theta}\right) - v\left(\mathbf{X}_{i0}, \boldsymbol{\theta}\right)\right) = F\left(v\left(\mathbf{X}_{i1}, \boldsymbol{\theta}\right) - v\left(\mathbf{X}_{i0}, \boldsymbol{\theta}\right)\right).$$

In applications the function $v\left(\mathbf{X}_{ij}, \boldsymbol{\theta}\right)$ is often assumed linear in parameters, i.e.,

(2.25) $$v\left(\mathbf{X}_{ij}, \boldsymbol{\theta}\right) = \mathbf{X}_{ij}\,\boldsymbol{\beta} \equiv \beta_0 + \sum_{k=1}^{m} X_{ijk}\,\beta_k$$

where $\boldsymbol{\theta} = \boldsymbol{\beta}$. If (2.25) holds, (2.24) one can write

(2.26) $$h\left(\mathbf{X}_{i1}, \mathbf{X}_{i0}, \boldsymbol{\theta}\right) \equiv E\left(Y_i \mid \mathbf{X}_{i1}, \mathbf{X}_{i0}\right) = F\left(\mathbf{X}_i \boldsymbol{\beta}\right)$$

where $\mathbf{X}_i = \mathbf{X}_{i1} - \mathbf{X}_{i0}$.

The Probit model

Suppose ε_{i1} and ε_{i0} are independent and normally distributed with

(2.27) $$\mathrm{Var}\left(\varepsilon_{ij} \mid \mathbf{X}_{i1}, \mathbf{X}_{i0}\right) = \tau_j^2 .$$

Then, conditional on $\left(X_{i1}, X_{i0}\right)$,

$$\varepsilon_{i1} - \varepsilon_{i0} \sim N\left(0, \tau^2\right)$$

where $\tau^2 = \tau_1^2 + \tau_2^2$. Hence we obtain in this case that

(2.28) $$F(y) = \Phi\left(\frac{y}{\tau}\right),$$

and consequently we obtain the *Probit model*,

(2.29) $$h\left(\mathbf{X}_{i1}, \mathbf{X}_{i0}, \boldsymbol{\theta}\right) = \Phi\left(\mathbf{X}_i \boldsymbol{\beta}^*\right)$$

where $\boldsymbol{\beta}^* = \boldsymbol{\beta}/\tau$. We cannot identify the parameter τ in this model, and we need not either, since the model is fully determined through $\mathbf{X}_i\boldsymbol{\beta}^*$.

The Logit model

Suppose that the error terms ε_{i1} and ε_{i0} are independent extreme value distributed (type III), i.e.,

(2.30) $$P\left(\varepsilon_{ij} \le y\right) = \exp\left(-e^{-y}\right), \; y \in R .$$

Then it follows easily that

(2.31) $$P\left(\varepsilon_{i0} - \varepsilon_{i1} \le y\right) = \frac{1}{1+e^{-y}}$$

which is the Logistic distribution introduced in (2.12). If (2.31) holds we therefore get the Logit model;

(2.32) $$h\left(\mathbf{X}_{i1},\mathbf{X}_{i0},\boldsymbol{\theta}\right)=\frac{1}{1+\exp\left(-\mathbf{X}_i\,\beta\right)}.$$

2.4.The multinomial Logit model

In many instances it is of interest to analyze data that are outcomes of multinomial experiments, regardless or not these are generated by discrete choice behavior. This means that the "outcomes" fall into one out of m (say) categories, where m may be greater than two. For example, when analyzing traffic accidents it may be useful to operate with several type of accidents.

Let Y_{ij} be equal to one if outcome j occurs for individual I and zero otherwise. Let P_{ij} = $P(Y_{ij} = 1)$. Then one must have that $0 \le P_j \le 1$, and $\sum_j P_j = 1$. One type of specification that fulfills these requirements is the multinomial logit model. One version of the multinomial logit model has the structure

(2.34) $$P_j = H_j(X;\boldsymbol{\beta}) \equiv \frac{\exp\left(X\beta_j\right)}{\sum_{k=1}^{m}\exp\left(X\beta_k\right)}$$

where X is, typically, a vector of agent-specific variables $\beta_j,\ j=1,2,\ldots,m,$ are vectors of unknown parameters, and $\boldsymbol{\beta}=\left(\beta_1,\beta_2,\ldots,\beta_m\right)$. This specification is also convenient for estimation purposes as we shall discuss in Section 6.

From (2.34) it follows that

(2.35) $$\log\left(\frac{H_j(X;\boldsymbol{\beta})}{H_1(X;\boldsymbol{\beta})}\right)=X\left(\beta_j-\beta_1\right).$$

Eq. (2.35) demonstrates that at most $\beta_j-\beta_1$ can be identified. To realize this, suppose $\beta_j^*,$ are parameter vectors such that $\beta_j^*\neq\beta_j,\ j=1,2,\ldots,m$. If

$$\beta_j^*=\beta_j-\beta_1+\beta_1^*$$

for $j=2,\ldots,m,$ then $\left\{\beta_j^*\right\}$ will satisfy (2.35), and consequently $\left\{\beta_j\right\}$ are not identified. We can therefore, without loss of generality, put $\beta_1=0$, and write

(2.36a) $$H_1(X;\boldsymbol{\beta}) = \frac{1}{1+\sum_{k=2}^{m} \exp(X\beta_k)}$$

and

(2.36b) $$H_j(X;\boldsymbol{\beta}) = \frac{\exp(X\beta_j)}{1+\sum_{k=2}^{m} \exp(X\beta_k)}$$

for $j = 2,3,\ldots,m$. Evidently, with sufficient variation in the X-vector, β_j, $j = 2,3,\ldots,m$, will be identified.

Example 2.2

Consider the choice of tourist destination. Suppose there are m actual destinations. We assume that actual variables that influence this choice are age, income, education, marital status, family size, etc. Let X be the vector of these variables. The probability of choosing destination j can be modelled as in (2.36).

3. Theoretical developments of probabilistic choice models

3.1. Random utility models

As indicated above, the basic problem confronted by discrete choice theory is the modelling of choice from a set of mutually exclusive and collectively exhaustive alternatives. In principle, one could apply the conventional microeconomic approach for divisible commodities to model these phenomena but a moment's reflection reveals that this would be rather ackward. This is due to the fact that when the alternatives are discrete, it is not possible to base the modelling of the agent's chosen quantities by evaluating marginal rates of substitution (marginal calculus), simply because the utility function will not be differentiable. In other words, the standard marginal calculus approach does not work in this case. Consequently, discrete choice analysis calls for a different approach.

3.1.1. The Thurstone model

Historically, discrete choice analysis was initiated by psychologists. Thurstone (1927) proposed the Thurstone model to explain the results from psychological and psychophysical experiments. These

experiments involved asking students to compare intensities of physical stimuli. For example, a student could be asked to rank objects in terms of weights, or tones in terms of loudness. The data from these experiments revealed that there seemed to be the case that some students would make different rankings when the choice experiments were replicated. To account for the variability in responses, Thurstone proposed a model based on the idea that a stimulus induces a "psychological state" that is a realization of a random variable. Specifically, he represented the preferences over the alternatives by random variables, so that the individual decision-maker would choose the alternative with the highest value of the random variable. The interpretation is two-fold: First, the utilities may vary across individuals due to variables that are not observable to the analyst. Second, the utility of a given alternative may also vary from one moment to the next, for the same individual, due to fluctuations in the individual's psychological state. As a result, the observed decisions may vary across identical experiments even for the same individual.

In many experiments Thurstone asked each individual to make several binary comparisons, and he represented the utility of each alternative by a normally distributed random variable. Let U_1^i and U_2^i denote the utilities a specific individual associates with the alternatives in replication no. i, $i = 1,2,\ldots,n$. Thurstone assumed that

$$U_j^i = v_j + \varepsilon_j^i$$

where ε_j^i, $j = 1,2, i = 1,2,\ldots,n,$ are independent and normally distributed where ε_j^i has zero mean and standard deviation equal to σ_j. Thus according to the decision rule the individual would choose alternative one in replication i if U_1^i is greater than U_2^i. Due to the "error term", ε_j^i, the individual may make different judgments in replications of the same experiment. Let $Y_j^i = 1$ if alternative j is chosen in replication i and zero otherwise. The relative number of times the individual chooses alternative j, $\hat{P}_j$, equals

$$\hat{P}_j \equiv \sum_{i=1}^{n} Y_j^i / n,$$

$j = 1,2.$ When the number of replications increases, then it follows from the law of large numbers that $\hat{P}_1$ tends towards the theoretical probability;

(3.1) $$P_1 \equiv P\left(U_1^i > U_2^i\right) = \Phi\left(\frac{v_1 - v_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)$$

where $\Phi(\cdot)$ is the standard cumulative normal distribution. The last equality in (3.1) follows from the assumption that the error terms are normally distributed random variables. The probability in (3.1) represents the propensity of choosing alternative j and it is a function of the standard deviations and the means, v_1 and v_2. While v_j repesents the "average" utility of alternative j the respective standard deviations account for the degree of instability in the individuals preferences across replicated experiments. We recognize (3.1) as a version of the binary probit model.

Although Thurstone suggested that the above approach could be extended to the multinomial choice setting, and with other distribution functions than the normal one, the statistical theory at that time was not sufficiently developed to make such extensions practical.

3.1.2. The neoclassisist's approach

The tradition in economics is somewhat different from the psychologist's approach. Specifically, the econometrician usually is concerned with analyzing discrete data obtained from a *sample* of individuals. With a neoclassical point of departure, the tradition is that preferences are typically assumed to be deterministic from the agent' point of view, in the sense that if the experiment were replicated, the agent would make identical decisions. In practice, however, one may observe that observationally identical agents make different choices. This is explained as resulting from variables that affect the choice process and are unobservable to the econometrician. The unobservables are, however, assumed to be perfectly known to the individual agents. Consequently, the utility function is modeled as random from the observing econometricians point of view, while it is interpreted as deterministic to the agent himself. Thus the randomness is due to the lack of information available to the observer. Thus, in contrast to the psychologist, the neoclassical economist seems usually reluctant to interpret the random variables in the utility function as random to the agent himself. Since the economist often does not have access to data from replicated experiments, he is not readily forced to modify his point of view either. There are, however, exceptions, see for example Quandt (1956) and Georgescu-Roegen (1958).

3.1.3. General systems of choice probabilities

Formally, we shall define a system of choice probabilities as follows:

Definition 1; System of choice probabilities

(i) A univers of choice alternatives, S. Each alternative in S may be characterized by a set of variables which we shall call attributes.

(ii) Possibly a set of agent-specific characteristics.

(iii) A family of choice probabilities $\{P_j(B),\ j \in B \subseteq S\}$*, where* $P_j(B)$ *is the probability of choosing alternative j when B is the set (choice set) of feasible alternatives presented to the agent. The choice probabilities are possible dependent on individual characteristics of the agent and of attributes of the alternatives within the choice set.*

Evidently, for each given $B \subseteq S$, $\sum_{j \in B} P_j(B)=1$, since for given B, $P_j(B)$ are "multinomial" probabilities.

Definition 2

A system of choice probabilities constitutes a random utility model if there exists a set of (latent) random variables $\{U_j,\ j \in S\}$ *such that*

(3.2)
$$P_j(B) = P\left(U_j = \max_{k \in B} U_k\right).$$

The random variable U_j is called the utility of alternative j. If the joint distribution function of the utilities has been specified it is possible to derive the structure of the choice probabilities by means of (3.2) as a function of the joint distribution of the utilities. However, in most cases the resulting expression will be rather complicated. As explained above, the empirical counterpart of $P_j(B)$ is the fraction of individuals with observationally identical characteristics that have chosen alternative j from B.

Often , the random utilities are assumed to have an additively separable structure,

(3.3)
$$U_j = v_j + \varepsilon_j,$$

where v_j is a deterministic term and ε_j is a random variable. The joint distribution of the terms $(\varepsilon_1, \varepsilon_2, \ldots)$ is assumed to be independent of $\{v_j\}$. In empirical applications the deterministic terms are specified as functions of observable attributes and individual characteristics.

Similarly to Manski (1977) we may identify the following sources of uncertainty that contribute to the randomness in the preferences:

(i) *Unobservable attributes*: The vector of attributes that characterize the alternatives may only partly be observable to the econometrician.

(ii) *Unobservable individual-specific characteristics:*Some of the variables that influence the variation in the agents tastes may partly be unobservable to the econometrician.

(iii) *Measurement errors:* There may be measurement errors in the attributes, choice sets and individual characteristics.

(iv) *Functional misspecification:* The functional form of the utility function and the distribution of the random terms are not fully known by the observer. In practice, he must specify a parametric form of the utility function as well as the distribution function which at best are crude approximations to the true underlying functional forms.

(v) *Bounded rationality:* One might go along with the psychologists point of view in allowing the utilities to be random to the agent himself. In addition to the assessment made by Thurstone, there is an increasing body of empirical evidence, as well as common daily life experience, suggesting that agents in the decision-process seem to have difficulty with assessing the precise value of each alternative. Consequently, their preferences may change from one moment to the next in a manner that is unpredictable (to the agents themselves).

To summarize, it is possible to interpret the randomness of the agents utility functions as partly an effect of unobservable taste variation and partly an effect that stem from the agents difficulty of dealing with the complexity of assessing the proper value to the alternatives. In other words, it seems plausible to interpret the utilities as random variables both to the observer as well as to the agent himself. In practice, it will seldom be possible to identify the contribution from the different sources to the uncertainty in preferences. For example, if the data at hand consists of observations from a cross-section of consumers, we will not be able to distinguish between seemingly inconsistent choice behavior that results from unobservables versus preferences that are uncertain to the agents themselves.

Before we discuss the random utility approach further we shall next turn to a very important contribution in the theory of discrete choice.

3.2. Independence from Irrelevant Alternatives and the Luce model

Luce (1959) introduced a class of probabilistic discrete choice model that has become very important in many fields of choice analyses. Instead of Thurstone's random utility approach, Luce postulated a structure on the choice probabilities directly without assuming the existence of any underlying

(random) utility function. Recall that $P_j(B)$ means the probability that the agent shall choose alternative j from B when B is the choice set. Statistically, for each given B, recall that these are the probabilities in a multinomial model, (due to the fact that the choices are mutually exclusive), which sum up to one. However, the question remains how these probabilities should be specified as a function of the attributes and how the choice probabilities should depend on the choice set, i.e., in other words, how should $\{P_j(B)\}$ and $\{P_j(A)\}$ be related when $j \in B \cap A$? To deal with this challenge, Luce proposed his famous Choice Axiom, which has later been known as the IIA property; "Independence from Irrelevant Alternatives". To describe IIA we think of the agent as if he is organizing his decision-process in two (or several) stages: In the first stage he selects a subset A from B, where A contains alternatives that are preferable to the alternatives in B\A. In the second stage the agent subsequently chooses his preferred alternative from A. So far this entails no essential loss of generality, since it is usually always possible to think of the decision process in this manner. The crucial assumption Luce made is that, on average, the choice from A in the last stage does not depend on alternatives outside A; the alternatives discarded in the first stage has been completely "forgotten" by the agent. In other words, the alternatives outside A are irrelevant. A probabilistic statement of this property is as follows: Let $P_A(B)$ denote the probability of selecting a subset A from B, defined by

$$P_A(B) = \sum_{j \in A} P_j(B).$$

Specifically, $P_A(B)$ means the probability of selecting a set of alternatives A which are at least as attractive as the alternatives B\A.

To state IIA formally, let J(B) denote the agent's choice from B. Thus, we can express the choice probability alternatively as $P_j(B) = P(J(B) = j)$.

Definition 3; Independence from Irrelevant Alternatives (IIA)

Let $\{P_j(B)\}$ *be a system of choice probabilities with probabilities that are different from zero and one. This system satisfies IIA if and only if for any* $A,B \subseteq S$ *such that* $j \in A \subset B \subseteq S$

(3.4) $$P\left(J(B) = j \middle| J(B) \in A\right) = P\left(J(A) = j\right).$$

Eq. (3.4) states that the choice from B, given that the chosen alternative belongs to A is the same as if A were the "original" choice set. We can rewrite (3.4) as follows. The left hand side of (3.4) can be expressed as

$$P\left(J(B)=j \middle| J(B)\in A\right)=\frac{P\left(\left(J(B)=j\right)\cap\left(J(B)\in A\right)\right)}{P\left(J(B)\in A\right)}=\frac{P\left(J(B)=j\right)}{P\left(J(B)\in A\right)}=\frac{P_j(B)}{P_A(B)}.$$

Hence, (3.4) is equivalent to

(3.5) $$P_j(B)=P_A(B)P_j(A).$$

Eq. (3.4) states that the probability of choosing alternative j from B equals the probability that A is a subset of the "best" alternatives which is selected in stage one times the probability of selecting alternative j from A in the second stage. Notice that the second stage probability, $P_j(A)$, has the same structure as $P_j(B)$, i.e., it does not depend on alternatives outside the (current) choice set A. Note that since this is a probabilistic statement it does not mean that IIA should hold in every single experiment. It only means that it should hold on average, when the choice experiment is replicated a large number of times, or alternatively, it should hold on average in a large sample of "identical" agents. (In the sense of agents with identically distributed tastes.) We may therefore think of IIA as an assumption of "probabilistic rationality". Another way of expressing IIA is that the rank ordering within any subset of the choice set is, on average, independent of alternatives outside the subset.

Definition 4; The Constant-Ratio Rule

A system of choice probabilities, $\{P_j(B)\}$, satisfies the constant-ratio rule if and only if for all j, k, B such that $j,k\in B\subseteq S$,

(3.5) $$P_j\left(\{k,j\}\right)/P_k\left(\{k,j\}\right)=P_j(B)/P_k(B)$$

provided the denominators do not vanish.

The following results are due to Luce (1959):

Theorem 1

Suppose $\{P_j(B)\}$ *is a system of choice probabilities and assume that* $P_j(\{j,k\}) \in (0,1)$ *for all* $j,k \in S$. *Then part (i) of the IIA assumption holds if and only if there exist positive scalars, a(j),* $j \in S$, *such that the choice probabilities equal*

(3.6)
$$P_j(B) = \frac{a(j)}{\sum_{k \in B} a(k)}.$$

Moreover, the scalars {a(j)} are unique apart from multiplication by a positive constant.

Proof: Assume first that (3.6) holds. Then it follows immediately that (3.4) holds. Assume next that (3.4) holds. Define $a(j) = c\,P_j(S)$, where c is an arbitrary positive constant. Then by (3.4) with $B = S$ and $A = B$, we obtain

$$P_j(B) = \frac{P_j(S)}{P_B(S)} = \frac{a(j)\,c}{\sum_{k \in B} a(k)\,c} = \frac{a(j)}{\sum_{k \in B} a(k)}$$

where $B \subseteq S$. This shows that $P_j(B)$ has the structure (3.6).

To show uniqueness (apart from multiplication by a constant), let $\tilde{a}(j)$ be positive scalars such that (3.6) holds with a(j) replaced by $\tilde{a}(j)$. Then with $B = S$ we get

$$\frac{P_j(S)}{P_1(S)} = \frac{a(j)}{a(1)} = \frac{\tilde{a}(j)}{\tilde{a}(1)}$$

which implies that

$$\tilde{a}(j) = a(j) \cdot \frac{\tilde{a}(1)}{a(1)}.$$

Thus we have proved that IIA implies the existence of scalars $\{a(j),\ j \in S\}$, such that (3.6) holds and these scalars are unique apart from multiplication by a constant.

Q.E.D.

Theorem 2

Let $\{P_j(B)\}$ *be a system of choice probabilities. The Constant-Ratio Rule holds if and only if IIA holds.*

Proof: The constant ratio rule implies that for $j,k \in A \subset B \subset S$

$$\frac{P_j(B)}{P_k(B)} = \frac{P_j(\{j,k\})}{P_k(\{j,k\})} = \frac{P_j(A)}{P_k(A)}.$$

Hence, since

$$P_j(B)\,P_k(A) = P_j(A)\,P_k(B)$$

and

$$\sum_{k\in A} P_k(A) = 1,$$

we obtain

$$P_j(B) = P_j(B)\sum_{k\in A} P_k(A) = P_j(A)\sum_{k\in A} P_k(B) = P_j(A)\,P_A(B).$$

Conversely, if IIA holds we realize immediately that the constant ratio rule will hold.

Q.E.D.

The results above are very powerful in that they establish statements that are equivalent to the IIA assumption, and they yield a simple structure of the choice probabilities. For example, if the univers S consists of four alternatives, S = {1,2,3,4}, there will be at most 11 different choice sets, namely {1,2}, {1,3}, {2,3}, {1,4}, {2,4}, {3,4}, {1,2,3}, {1,2,4}, {1,3,4}, {2,3,4}, {1,2,3,4}. This yields altogether 28 probabilities. Since the probabilities sum to one for each choice set we can reduce the number of "free" probabilities to 17. However, when IIA holds we can express all the choice probabilities by only three scale values, a_2, a_3 and a_4 (since we can choose a_1=1, or equal to any other positive value). We therefore realize that the Luce model implies strong restrictions on the system of choice probabilities.

There is another interesting feature that follows from the Luce model, expressed in the next Corollary.

Corollary 1

If IIA, part (i) holds it follows that for distinct i, j and $k \in S$

(3.7) $$P_i(\{i,j\})P_j(\{j,k\})P_k(\{k,i\}) = P_i(\{i,k\})P_k(\{k,j\})P_j(\{j,i\}).$$

The proof of this result is immediate.

Recall that IIA only implies rationality "in the long run", or at the aggregate level. Thus the probability of intransitive sequences (chains) is positive. The result in Corollary 1 is a statement about intransitive chains beause the interpretation of (3.7) is that

$$\mathrm{P}(\mathrm{i} \succ \mathrm{j} \succ \mathrm{k} \succ \mathrm{i}) = \mathrm{P}(\mathrm{i} \succ \mathrm{k} \succ \mathrm{j} \succ \mathrm{i})$$

where $\succ$ means "preferred to". In other words, the intransitive chains $\mathrm{i} \succ \mathrm{j} \succ \mathrm{k} \succ \mathrm{i}$ and $\mathrm{i} \succ \mathrm{k} \succ \mathrm{j} \succ \mathrm{i}$ have the same probability. This shows that although intransitive "chains" can occur with positive probability there is no systematic violation of transitivity. In fact, it can also be proved that if (3.7) holds then the binary choice probabilities must have the form

(3.8) $$\mathrm{P_j}(\{\mathrm{i},\mathrm{j}\}) = \frac{\mathrm{a(j)}}{\mathrm{a(i)} + \mathrm{a(j)}}$$

where $\{\mathrm{a(j)}, \mathrm{j} \in \mathrm{S}\}$ are unique up to multiplication by a constant, cf. Luce and Suppes (1965). However, (3.7) does not imply IIA. Equation (3.7) is often called the *Product rule.*

3.3. The relationship between IIA and the random utility formulation

After Luce had introduced the IIA property and the corresponding Luce model, Luce (1959), the question whether there exists a random utility model that is consistent with IIA was raised. A first answer to this problem was given by Holman and Marley in an unpublished paper (cf. Luce and Suppes, 1965, p. 338).

Theorem 3

Assume a random utility model, $U_j = \mathrm{v}_j + \varepsilon_j$ *, where* ε_j *,* $j \in S$*, are independent random variables with standard type III extreme value distribution*[1]

(3.9) $$P(\varepsilon_j \le x \mid \mathrm{v}_k, k \in S) = exp(-e^{-x}).$$

Then, for $j \in B \subseteq S$*,*

[1] In the following the distribution function (3.9) will be called the standard extreme value distribution.

(3.10) $$P_j(B) \equiv P\left(U_j = \max_{k\in B} U_k\right) = \frac{e^{v_j}}{\sum_{k\in B} e^{v_k}}.$$

We realize that (3.10) is a Luce model with $v_j = \log a(j)$. Thus, by Theorem 3 there exists a random utility model that rationalizes the Luce model.

Proof: Let us first derive the cumulative distribution for $V_j \equiv \max_{k\in B\setminus\{j\}} U_k$. We have

(3.11) $$P\left(V_j \le y\right) = \prod_{k\in B\setminus\{j\}} P\left(\varepsilon_k \le y - v_k\right) = \prod_{k\in B\setminus\{j\}} \exp\left(-e^{v_k - y}\right) = \exp\left(-e^{-y} D_j\right)$$

where

(3.12) $$D_j = \sum_{k\in B\setminus\{j\}} e^{v_k}.$$

Hence

(3.13) $$P\left(U_j = \max_{k\in B} U_k\right) = P\left(U_j > V_j\right) = P\left(\varepsilon_j + v_j > V_j\right) = \int_{-\infty}^{\infty} P\left(y > V_j\right) P\left(\varepsilon_j + v_j \in (y, y+dy)\right).$$

Note next that since by (3.9)

$$P\left(U_j \le y\right) = P\left(\varepsilon_j + v_j < y\right) = \exp\left(-e^{v_j - y}\right)$$

it follows that

$$P\left(\varepsilon_j + v_j \in (y, y+dy)\right) = \exp\left(-e^{v_j - y}\right) e^{v_j - y}\, dy.$$

Hence

(3.14)
$$\begin{aligned}
&\int_{-\infty}^{\infty} P\left(y > V_j\right) P\left(\varepsilon_j + v_j \in (y, y+dy)\right) = \int_{-\infty}^{\infty} \exp\left(-D_j e^{-y}\right) \exp\left(-e^{v_j - y}\right) e^{v_j - y}\, dy \\
&= e^{v_j} \int_{-\infty}^{\infty} \exp\left(-\left(D_j + e^{v_j}\right) e^{-y}\right) e^{-y}\, dy \\
&= \frac{e^{v_j}}{D_j + e^{v_j}} \Big|_{-\infty}^{\infty} \exp\left(-\left(D_j + e^{v_j}\right) e^{-y}\right) = \frac{e^{v_j}}{D_j + e^{v_j}}.
\end{aligned}$$

Since

$$D_j + e^{v_j} = \sum_{k \in B} e^{v_k}$$

the result of the Theorem follows from (3.13) and (3.14).

Q.E.D.

An interesting question is whether or not there exists other distribution functions than (3.9) which imply the Luce model. McFadden (1973) proved that under particular assumptions the answer is no. Later Yellott (1977) and Strauss (1979) gave proofs of this result under weaker conditions. Yellott (1977) proved the following result.

Theorem 4

Assume that S contains more than two alternatives, and $U_j = v_j + \varepsilon_j$, *where* ε_j, $j \in S$, *are i.i.d. with cumulative distribution function that is independent of* $\{v_j, j \in S\}$ *and is strictly increasing on the real line. Then (3.10) holds if and only if* ε_j *has the standard extreme value distribution function.*

Example 3.1

Consider the choice between m brands of cornflakes. The price of brand j is Z_j. We assume that the utility function of the consumer has the form

(3.15) $$U_j = Z_j \tilde{\beta} + \varepsilon_j \sigma$$

where $\beta < 0$ and $\sigma > 0$ are unknown parameters, ε_j, $j = 1,2,\ldots,m$, are i.i. extreme value distributed. Without loss of generality we can write the utility function as

(3.16) $$\tilde{U}_j = Z_j \tilde{\beta}/\sigma + \varepsilon_j \equiv Z_j \beta + \varepsilon_j .$$

From Theorem 3 it follows that the choice probabilities can be written as

(3.17) $$P_j = \frac{\exp(Z_j \beta)}{\sum_{k=1}^{m} \exp(Z_k \beta)}.$$

Clearly, β is identified, since

$$\log\left(\frac{P_j}{P_1}\right) = \left(Z_j - Z_1\right)\beta .$$

However, σ is not identified. Note that the variance of the error term in the utility function is large when σ is large, which in formulation (3.16) corresponds to a small β.

When β has been estimated one can compute the aggregate own- and cross-price elasticities according to the formulae

(3.18) $$\frac{\partial \log P_j}{\partial \log Z_j} = \beta Z_j \left(1 - P_j\right)$$

and

(3.19) $$\frac{\partial \log P_j}{\partial \log Z_k} = -\beta Z_k P_k$$

for $k \neq j$.

Example 3.2

Consider a transportation choice problem. There are two feasible alternatives, namely driving own car (Alternative 1), or riding a bus (Alternative 2).

Let i index the commuter and let

$$Z_{ij1} = \begin{cases} 1 & \text{if} \quad j = 1 \\ 0 & \text{otherwise}, \end{cases}$$

Z_{ij2} = In-vehicle time, alternative j,

Z_{ij3} = Out-of-vehicle time, alternative j,

Z_{ij4} = Transportation cost, alternative j.

The variable Z_{ij1} is supposed to represent the intrinsic preference for driving own car. The utility function is assumed to have the structure

$$U_{ij} = Z_{ij}\beta + \varepsilon_{ij}$$

where $Z_{ij} = \left(Z_{ij1}, Z_{ij2}, Z_{ij3}, Z_{ij4}\right)$, ε_{i1} and ε_{i2} are i.i. extreme value distributed, and β is a vector of unknown coefficients. From these assumptions it follows that the probability that commuter i shall choose alternative j is given by

(3.20) $$P_{ij} = \frac{\exp\left(Z_{ij}\beta\right)}{\sum_{k=1}^{2} \exp\left(Z_{ik}\beta\right)}.$$

From a sample of observations of individual choices and attribute variables one can estimate β by the maximum likelihood procedure.

Let us consider how the model above can be applied in policy simulations once β has been estimated. Consider a group of individuals facing some attribute vector Z_j, $j = 1,2$. The corresponding choice probability equals

(3.21) $$P_j = \frac{\exp\left(Z_j\beta\right)}{\sum_{k=1}^{2} \exp\left(Z_k\beta\right)}$$

for $j = 1,2$. From (3.21) it follows that

(3.22) $$\frac{\partial \log P_j}{\partial \log Z_{jr}} = \beta_r \, Z_{jr} \left(1 - P_j\right)$$

and

(3.23) $$\frac{\partial \log P_j}{\partial \log Z_{kr}} = -\beta_r \, Z_{kr} \, P_k$$

for $k \neq j$. Eq. (3.22) expresses the "own elasticities" while (3.23) expresses the "cross elasticities". Specifically, (3.22) yields the relative increase in the fraction of individuals that choose alternative j that follows from a relative increase in Z_{jr} by one unit.

Example 3.3. (Multinomial logit)
Assume that

(3.24) $$F_j(y) = \exp\left(-e^{-y}\right).$$

Then (3.24) yields

(3.25) $$P_j(B) = \frac{e^{v_j}}{\sum_{k \in B} e^{v_k}}.$$

Example 3.4. (Independent multinomial probit)

If

$$F_j'(y) = \Phi'(y) \equiv \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \tag{3.26}$$

then we obtain the socalled *Independent multinomial Probit model;*

$$P_j(B) = \int_{-\infty}^{\infty} \prod_{k \in B \setminus \{j\}} \Phi(y - v_k) \exp\left(-\frac{1}{2}(y - v_j)^2\right) \frac{dy}{\sqrt{2\pi}}. \tag{3.27}$$

It has been found through simulations and empirical applications that the independent probit model yields choice probabilities that are close to the multinomial logit choice probabilities.

Example 3.5. (Binary probit)

Assume that $B = \{1,2\}$ and $F_j(y) = \Phi\left(y\sqrt{2}\right)$. Then

$$P(U_1 > U_2) = \Phi(v_1 - v_2). \tag{3.28}$$

Example 3.6. (Binary Arcus-tangens)

Assume that $B = \{1,2\}$ and

$$F_j'(y) = \frac{2}{\pi\left(1 + 4y^2\right)}. \tag{3.29}$$

The density (3.29) is the density of a Cauchy distribution. Then

$$P(U_1 > U_2) = \frac{1}{2} + \frac{1}{\pi} \operatorname{Arctg}(v_1 - v_2). \tag{3.30}$$

The Arcus-tangens model differs essentially from the binary logit and probit models in that the tails of the Arcus-tangens model are much heavier than for the other two models.

3.4. Specification of the structural terms, examples

Let $Z_j = (Z_{j1}, Z_{j2}, \ldots, Z_{jK})$ denote a vector of attributes that characterize alternative j. In the absence of individual characteristics, a convenient functional form is

(3.31) $$v_j = Z_j \beta \equiv \sum_{k=1}^{K} Z_{jk} \beta_k .$$

A more general specification is

(3.32) $$v_j = \sum_{k=1}^{K} h_k (Z_j, X) \beta_k$$

where $h_k (Z_j, X), k = 1, \ldots, K,$ are known functions of the attribute vector and a vector variable X that characterizes the agent.

Example 3.7

Let $X = (X_1, X_2)$ and $Z_j = (Z_{j1}, Z_{j2})$. A type of specification that is often used is

(3.33) $$v_j = Z_{j1} \beta_1 + Z_{j2} \beta_2 + Z_{j1} X_1 \beta_3 + Z_{j1} X_2 \beta_4 + Z_{j2} X_1 \beta_5 + Z_{j2} X_2 \beta_6 .$$

In some applications the assumption of linear-in-parameter functional form may, however, be too restrictive.

Example 3.8. (Box-Cox transformation):

Let $Z_j = (Z_{j1}, Z_{j2}), Z_{jk} > 0, k = 1,2,$

and

(3.34) $$v_j = \left(\frac{Z_{j1}^{\alpha_1} - 1}{\alpha_1} \right) \beta_1 + \left(\frac{Z_{j2}^{\alpha_2} - 1}{\alpha_2} \right) \beta_2$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2$ are unknown parameters. The transformation

(3.35) $$\frac{y^{\alpha} - 1}{\alpha},$$

$y > 0,$ is called a Box-Cox transformation of y and it contains the linear function as a special case $(\alpha = 1)$. When $\alpha \to 0$ then

$$\frac{y^{\alpha}-1}{\alpha} \to \log y.$$

When $\alpha < 1$, $\left(y^{\alpha}-1\right)/\alpha$ is concave while it is convex when $\alpha > 1$. For any α, $\left(y^{\alpha}-1\right)/\alpha$ is increasing in y.

Example 3.9

A problem which is usually overlooked in discrete choice analyses is the fact that simultaneous equation problems can arise as a result of unobservable attributes. Consider the following example where the utility function has the structure

$$U_j = Z_j \beta_1 + Z_j X_1 \beta_2 + Z_j X_2 \beta_3 + \varepsilon_j$$

where Z_j is an attribute variable (scalar) and X_1, X_2 are individual characteristics. The random error term ε_j is assumed to be uncorrelated with Z_j, X_1 and X_2. Also Z_j is assumed uncorrelated with X_1 and X_2. However, X_2 is unobservable to the researcher. The researcher therefore specifies the utility function as

(3.36) $$U_j^* = Z_j \beta_1 + Z_j X_1 \beta_2 + \varepsilon_j^*.$$

Thus, the interpretation of ε_j^* is as

(3.37) $$\varepsilon_j^* = \varepsilon_j + Z_j X_2 \beta_3.$$

Then

$$E\left(\varepsilon_j^* \middle| X_1, Z_j\right) = Z_j \beta_3 E\left(X_2 \middle| X_1\right).$$

In this case we therefore get that the error terms are correlated with the structural terms when X_1 and X_2 are correlated. A completely similar argument applies in the case with unobservable attributes.

This simple example shows that simultaneous equation bias may be a serious problem in many cases where data contains limited information about population heterogeneity or/and relevant attributes. Note that even if we were able to observe the relevant explanatory variables, we may still face the risk of getting simultaneous equation bias as a result of misspesified functional form of the deterministic term of the utility function. This is easily demonstrated by a similar argument as the one above.

3.5. Stochastic models for ranking

So far we have only discussed models in which the interest is the agent's (most) preferred alternative. However, in several cases it is of interest to specify the joint probability of the rank ordering of alternatives that belong to S or to some subset of S. For example, in stated preference surveys, where the agents are presented with hypothetical choice experiments, one has the possibility of designing the questionaires so as to elicit information about the agents' rank ordering. This yields more information about preferences than data on solely the highest ranked alternatives, and it is therefore very useful for empirical analysis. This type of modeling approach has for example been applied to analyze the potential demand for products that may be introduced in the market, see Section 4.8.

The systematic development of stochastic models for ranking started with Luce (1959) and Block and Marschak (1960). Specifically, they provided a powerful theoretical rationale for the structure of the so-called ordered Luce model. The theoretical assumptions that underly the ordered Luce model can briefly be described as follows.

Let $\mathbf{R}(B)=\left(R_1(B), R_2(B),\ldots, R_m(B)\right)$ be the agent's rank ordering of the alternatives in B, where m is the number of alternatives in B, and $B \subseteq S$. This means that $R_i(B)$ denotes the element in B that has the i'th rank. As above let $P_j(B)$, $j \in B$, be the probability that the agent shall rank alternative j on top when B is the set of feasible alternatives. Recall that the empirical counterpart of these probabilities is the respective number of times the agent chooses a particular rank ordering to the total number of times the experiment is replicated, or alternatively, the fraction of (observationally identical) agents that choose a particular rank ordering. Let $\boldsymbol{\rho}(B)=\left(\rho_1,\rho_2,\ldots,\rho_m\right)$, where the components of the vector $\boldsymbol{\rho}(B)$ are distinct and $\rho_k \in B$ for all $k \le m$.

Similarly to Definition 1 one can define a system of ranking probabilities formally. Since the extension from Definition 1 to the case with ranking is rather obvious we shall not present the formal definition here.

Definition 5

A system of ranking probabilities constitute a random utility model if and only if

$$P\left(\boldsymbol{R}(B)=\boldsymbol{\rho}(B)\right)=P\left(U(\rho_1)>U(\rho_2)>\ldots>U(\rho_m)\right)$$

for $B \subseteq S$, *where* $\{U(j), j \in S\}$, *are random variables.*

The next definition is a generalization of IIA to the setting with rank ordering. For simplicity we rule out the case with degenerate choice probabilities equal to zero or one.

Definition 6: Generalized IIA (IIAR)

A system of ranking probabilities satisfies the Independence from Irrelevant Alternatives (IIAR) property if and only if for any $B \subseteq S$

(3.38) $$P\left(\boldsymbol{R}(B)=\boldsymbol{\rho}(B)\right) = P_{\rho_1}(B)\, P_{\rho_2}\left(B \setminus \{\rho_1\}\right) \ldots P_{\rho_{m-1}}\left(\{\rho_{m-1},\rho_m\}\right).$$

Definition 6 states that an agent's ranking behavior can (on average) be viewed as a multistage process in which he first selects the most preferred alternative, next he selects the second best among the remaining alternatives, etc. The crucial point here is that in each stage, the agent's ranking of the remaining alternatives is independent of the alternatives that were selected in earlier steps. In other words, they are viewed as "irrelevant".

We realize that Definition 3 is a special case of Definition 6.

Let

$$\Omega_j(B) = \{\boldsymbol{\rho}(B) : \rho_1(B) = j,\ j \in B\}.$$

The interpretation of $\Omega_j(B)$ is as the set of rank orderings among the alternatives within B, where alternative j is ranked highest.

Theorem 5

Let $\{P(\boldsymbol{\rho}(B))\}$ *be a system of ranking probabilities, defined by* $P(\boldsymbol{\rho}(B)) = P\left(\boldsymbol{R}(B)=\boldsymbol{\rho}(B)\right)$. *This system constitutes a random utility model if and only if*

$$P_j(B) = \sum_{\boldsymbol{\rho}(B) \in \Omega_j(B)} P(\boldsymbol{\rho}(B)).$$

A proof of Theorem 5 is given by Block and Marschak (1960, p. 107).

Theorem 6

Assume that a system of ranking probabilities is consistent with a random utility model and that IIAR holds. Then there exists positive scalars, $a(j),\ j \in S$, *such that the ranking probabilities are given by*

(3.39)
$$P\left(\boldsymbol{R}(B)=\boldsymbol{\rho}(B)\right)=\frac{a(\rho_1)}{\sum_{k\in B}a(k)}\cdot\frac{a(\rho_2)}{\sum_{k\in B\setminus\{\rho_1\}}a(k)}\cdots\frac{a(\rho_{m-1})}{a(\rho_{m-1})+a(\rho_m)}$$

for $B\subseteq S$. The scalars, $\{a(j)\}$, are uniquely determined up to multiplication by a positive constant. Conversely, the model (3.41) satisfies IIAR.

Block and Marschak (1960, p. 109) have proved Theorem 6, cf. Luce and Suppes (1965).

Example 3.10

Consider the rankings of different brands of beer. Let $\mathrm{B}=\{1,2,3\}$ where alternative 1 is Tuborg, alternative 2 is Budweiser and alternative 3 is Becks. Suppose one has data on consumers rank ordering of these brands of beer. If IIAR holds then the probability that for example $\rho_{\mathrm{B}}=(2,3,1)$, i.e., Budweiser is ranked on top and Becks second best. According to (3.39) we obtain that the probability of ρ_{B} equals

$$\mathrm{P}(\mathbf{R}(\mathrm{B})=(2,3,1))=\frac{\mathrm{a}(2)}{\mathrm{a}(1)+\mathrm{a}(2)+\mathrm{a}(3)}\cdot\frac{\mathrm{a}(3)}{\mathrm{a}(1)+\mathrm{a}(3)}.$$

The next result shows that (3.39) is consistent with a simple random utility representation.

Theorem 7

Assume a random utility model with $U(j)=\mathrm{v}(j)+\varepsilon_j$, where ε_j, $j\in S$, are i.i.d. with standard extreme value distribution function that is independent of $\{\mathrm{v}(j),\ j\in S\}$. Then

(3.40)
$$\begin{aligned}&P\left(\boldsymbol{R}(B)=\boldsymbol{\rho}(B)\right)=P\left(U(\rho_1)>U(\rho_2)>\ldots>U(\rho_m)\right)\\&=\frac{exp\left(\mathrm{v}(\rho_1)\right)}{\sum_{k\in B}exp\left(\mathrm{v}(k)\right)}\cdot\frac{exp\left(\mathrm{v}(\rho_2)\right)}{\sum_{k\in B\setminus\{\rho_1\}}exp\left(\mathrm{v}(k)\right)}\cdots\frac{exp\left(\mathrm{v}(\rho_{m-1})\right)}{exp\left(\mathrm{v}(\rho_{m-1})\right)+exp\left(\mathrm{v}(\rho_m)\right)}.\end{aligned}$$

Also here we realize that Theorem 1 is a special case of Theorem 6 and Theorem 3 is a special case of Theorem 7 because the choice probability $\mathrm{P_j(B)}$ is equal to the sum of all ranking probabilities with $\rho_1=\mathrm{j}$. A proof of Theorem 7 is given in Strauss (1979).

3.6. Stochastic dependent utilities across alternatives

In the random utility models discussed above we only focused on models with random terms that are independent across alternatives. In particular we noted that the independent extreme value random utility model is equivalent to the Luce model. It has been found that the independent multinomial probit model is "close" to the Luce model in the sense that the choice probabilities are close provided the structural terms of the two models have the same structure (see for example, Hausman and Wise, 1978). However, the assumption of independent random terms is rather restrictive in some cases, which the following example will demonstrate.

Example 3.11

Consider a consumer choice problem in which there are two soda alternatives, namely "Coca cola", (1), "Fanta", (2). The fractions of consumers that buy Coca cola and Fanta are 1/3 and 2/3, respectively. If we assume that Luce's model holds we have

$$P_1(\{1,2\}) = \frac{a_1}{a_1 + a_2} = \frac{1}{3}.$$

With $a_1 = 1$ it follows that $a_2 = 2$. Suppose now that another Fanta alternative is introduced (alternative 3) that is equal in all attributes to the existing one except that its bottles have a different color from the original one. Since the new Fanta alternative is essential equivalent to the existing one it must be true that the corresponding response strengths must be equal, i.e., $a_3 = a_2 = 2$. Consequently, since the choice set is now equal to $\{1,2,3\}$ we have according to (3.6) that

$$P_1(\{1,2,3\}) = \frac{a_1}{a_1 + a_2 + a_3} = \frac{1}{1+2+2} = \frac{1}{5}$$

which implies that

$$P_2(\{1,2,3\}) = P_3(\{1,2,3\}) = \frac{2}{5}.$$

But intuitively, this seems unrealistic because it is plausible to assume that the consumers will tend to treat the two alternatives as a single alternative so that

$$P_1(\{1,2,3\}) = \frac{1}{3}$$

and

$$P_2\left(\{1,2,3\}\right) = P_3\left(\{1,2,3\}\right) = \frac{1}{3}\,.$$

This example demonstrates that if alternatives are "similar" in some sense, then the Luce model is not appropriate. A version of this example is due to Debreu (1960).

Example 3.12

Let us return to the general theory, and try to list some of the reasons why the random terms of the utility function may be correlated across alternatives.

For expository simplicity consider the (true) utility specification

$$U_j = Z_{j1}\,\beta_1 + X_1\,Z_{j1}\,\beta_2 + X_2\,Z_{j2}\,\beta_3 + \varepsilon_j \tag{3.41}$$

and suppose that only Z_{j1} and X_1 are observable for all j. Thus, in practice we may therefore be tempted to resort to the misspecified version

$$U_j^* \equiv Z_{j1}\beta_1 + X_j Z_{j1}\beta_2 + \varepsilon_j^* \tag{3.42}$$

where

$$\varepsilon_j^* = \varepsilon_j + X_2 Z_{j2}\beta_3. \tag{3.43}$$

Let $\mathbf{Z}^1 = \left(Z_{11}, Z_{21}, \ldots, Z_{m1}\right)$. From (3.38) it follows that

$$\begin{aligned}
&\mathrm{Cov}\left(\varepsilon_j^*, \varepsilon_k^* \,\middle|\, X_1, \mathbf{Z}^1\right) = \mathrm{Cov}\left(X_2\,Z_{j2}\,\beta_3, X_2\,Z_{k2}\,\beta_3 \,\middle|\, X_1, \mathbf{Z}^1\right) \\
&= \beta_3^2\,\mathrm{E}\,\mathrm{Cov}\left(X_2\,Z_{j2}, X_2\,Z_{k2} \,\middle|\, X_1, \mathbf{Z}^1, X_2\right) \\
&+ \beta_3^2\,\mathrm{Cov}\left(\mathrm{E}\left(X_2\,Z_{j2} \,\middle|\, X_1, \mathbf{Z}^1, X_2\right), \mathrm{E}\left(X_2\,Z_{k2} \,\middle|\, X_1, \mathbf{Z}^1, X_2\right)\right) \\
&= \beta_3^2\,\mathrm{E}\left(X_2^2 \,\middle|\, X_1\right)\mathrm{Cov}\left(Z_{j2}, Z_{k2} \,\middle|\, \mathbf{Z}^1\right) + \beta_3^2\,\mathrm{Var}\left(X_2 \,\middle|\, X_1\right)\mathrm{E}\left(Z_{j2} \,\middle|\, \mathbf{Z}^1\right)\mathrm{E}\left(Z_{k2} \,\middle|\, \mathbf{Z}^1\right).
\end{aligned} \tag{3.44}$$

This shows that unobservable attributes and individual characteristics may lead to error terms that are correlated across alternatives. Suppose next that $\mathrm{Cov}\left(Z_{j2}, Z_{k2} \,\middle|\, \mathbf{Z}^1\right) = 0$. Then (3.44) reduces to

$$\mathrm{Cov}\left(\varepsilon_j^*, \varepsilon_k^* \,\middle|\, X_1, \mathbf{Z}^1\right) = \beta_3^2\,\mathrm{E}\left(Z_{j2} \,\middle|\, \mathbf{Z}^1\right)\mathrm{E}\left(Z_{k2} \,\middle|\, \mathbf{Z}^1\right)\mathrm{Var}\left(X_2 \,\middle|\, X_1\right). \tag{3.45}$$

Eq. (3.45) shows that even if the unobservable attributes are uncorrelated the error terms will still be correlated if $\mathrm{Var}\left(X_2 \,\middle|\, X_1\right) \neq 0$. (If $\mathrm{Var}\left(X_2 \,\middle|\, X_1\right) = 0$, X_2 is perfectly predicted by X_1.)

3.7. The multinomial Probit model

The best known multinomial random utility model with interdependent utilities is the multinomial probit model. In this model the random terms in the utility function are assumed to be multinormally distributed (with unknown covariance matrix). The concept of multinomial probit appeared already in the writings of Thurstone (1927), but due to its computational complexity it has not been practically useful for choice sets with more than five alternatives until quite recently. In recent years, however, there has been a number of studies that apply simulation methods in the estimation procedure, pioneered by McFadden (1989). Still the computational issue is far from being settled, since the current simulation methods are complicated to apply in practice. The following expression for the multinomial choice probabilities is suggestive for the complexity of the problem. Let $h(x;\Omega)$ denote the density of an m-dimensional multinormal zero mean vector-variable with covariance matrix Ω. We have

$$h(x;\Omega) = (2\pi)^{-m/2} |\Omega|^{-1/2} \exp\left(-\frac{1}{2} x' \Omega^{-1} x\right) \tag{3.46}$$

where $|\Omega|$ denotes the determinant of Ω. Furthermore

$$P\left(v_j + \varepsilon_j = \max_{k \le m}\left(v_k + \varepsilon_k\right)\right) = \int_{-\infty}^{v_j - v_1} \cdots \int_{-\infty}^{v_j - v_j} \cdots \int_{-\infty}^{v_j - v_n} h\left(x_1, \ldots, x_j, \ldots, x_m; \Omega\right) dx_1 \ldots dx_j \ldots dx_m . \tag{3.47}$$

From (3.47) we see that an m-dimensional integral must be evaluated to obtain the choice probabilities. Moreover, the integration limits also depend on the unknown parameters in the utility function. When the choice set contains more than five alternatives it is therefore necessary to use simulation methods to evaluate these choice probabilities.

3.8. The Generalized Extreme Value model

McFadden (1978) and (1981) introduced the class of GEV model which is a random utility model that contains the Luce model as a special case. He proved the following result:

Theorem 8

Let G be a non-negative function defined over R_+^m *that has the following properties:*

(i) *G is homogeneous of degree one,*

(ii) $\lim_{y_i \to \infty} G(y_1, \ldots, y_i, \ldots, y_m) = \infty,\ i = 1,2,\ldots,m,$

(iii) *the k^{th} partial derivative of G with respect to any combination of k distinct components exist, are continuous, non-negative if k is odd, and are non-positive if k is even.*

Then

(3.48) $$F(x) = exp\left(-G\left(e^{-x_1}, e^{-x_2}, \ldots, e^{-x_m}\right)\right)$$

is a well defined multivariate (type III) extreme value distribution function. Moreover, if $\left(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m\right)$ *has joint distribution function given by (3.51), then it follows that*

(3.49) $$P\left(v_j + \varepsilon_j = \max_{k \le m}\left(v_k + \varepsilon_k\right)\right) = \frac{\partial G\left(e^{v_1}, e^{v_2}, \ldots, e^{v_m}\right) / \partial v_j}{G\left(e^{v_1}, e^{v_2}, \ldots, e^{v_m}\right)}.$$

The proof of Theorem 8 is analogous to the proof of Lemma A2 in Appendix A.

Conditions (ii) and (iii) are necessary to ensure that F(x) is a well defined multivariate distribution function (with non-negative density), while condition (i) characterizes the multivariate extreme value distribution.

Above we have stated the choice probability for the case where all the choice alternatives in S belong to the choice set. Obviously, we get the joint cumulative distribution function of the random terms of the utilities that correspond to any choice set B by letting $x_i = \infty$, for all $i \notin B$. This corresponds to letting $v_i = -\infty$, for all $i \notin B$ in the right hand side of (3.49).

To see that the Luce model emerges as a special case, let

(3.50) $$G\left(y_1, \ldots, y_m\right) = \sum_{k=1}^{m} y_k$$

from which it follows by (3.49) that

$$P_j(B) = \frac{e^{v_j}}{\sum_{k \in B} e^{v_k}}.$$

Example 3.13

Let $S = \{1,2,3\}$ and assume that

(3.51) $$G(y_1, y_2, y_3) = y_1 + \left(y_2^{1/\theta} + y_3^{1/\theta}\right)^{\theta}$$

where $0 < \theta \leq 1$. It can be demonstrated that θ has the interpretation

(3.52) $$\operatorname{corr}(\varepsilon_2, \varepsilon_3) = 1 - \theta^2$$

and

$$\operatorname{corr}(\varepsilon_1, \varepsilon_j) = 0, \quad j = 2,3.$$

From Theorem 8 we obtain that

(3.53) $$P_1(S) = \frac{e^{v_1}}{e^{v_1} + \left(e^{v_2/\theta} + e^{v_3/\theta}\right)^{\theta}}$$

and

(3.54) $$P_j(S) = \frac{\left(e^{v_2/\theta} + e^{v_3/\theta}\right)^{\theta-1} e^{v_j/\theta}}{e^{v_1} + \left(e^{v_2/\theta} + e^{v_3/\theta}\right)^{\theta}},$$

for $j = 2,3$. If $B = \{1,2\}$, then

(3.55) $$P_1(\{1,2\}) = \frac{e^{v_1}}{e^{v_1} + e^{v_2}}.$$

When alternative 2 and alternative 3 are close substitutes θ should be close to zero. By applying l'Hôpital's rule we obtain

$$\lim_{\theta \to 0} \theta \log\left(e^{v_2/\theta} + e^{v_3/\theta}\right) = \max(v_2, v_3).$$

Consequently, when θ is close to zero the choice probabilities above are close to

(3.56) $$P_1(S) = \frac{e^{v_1}}{e^{v_1} + \exp\left(\max(v_2, v_3)\right)}$$

and

(3.57) $$P_2(S) = \frac{e^{v_2}}{e^{v_1} + e^{v_2}},$$

if $v_2 > v_3$, and zero otherwise, and similarly for $P_3(S)$. For $v_2 = v_3$ we obtain

(3.58) $$P_1(S) = \frac{e^{v_1}}{e^{v_1} + e^{v_2}}$$

and

(3.59) $$P_j(S) = \frac{e^{v_2}}{2\left(e^{v_1} + e^{v_2}\right)}$$

for $j = 2,3$.

Consider again Example 3.11. With $v_2 = v_3$, $v_1 = 0$ and $e^{v_2} = 2$. Eq. (3.58) and (3.59) yield

$$P_1\left(\{1,2\}\right) = 1/3$$

and

$$P_2\left(\{1,2,3\}\right) = P_3\left(\{1,2,3\}\right) = 1/3.$$

Thus the model generated from (3.51) with θ close to zero is able to capture the underlying structure of Example 3.11.

3.8.1. The Nested multinomial logit model (nested logit model)

The nested logit model is an extension of the multinomial logit model which belongs to the GEV class. The nested logit framework is appropriate in a modelling situation where the decision problem has a "tree-structure". This means that the choice set can be partitioned into a hierarchical system of subsets that each group together alternatives having several observable characteristics in common. It is assumed that the agent chooses one of the subsets A_r (say) in the first stage from which he selects the preferred alternative. The choice problem in Example 3.11 has such a tree structure: Here the first stage concerns the choice between Coca cola and Fanta while the second stage alternatives are the two Fanta variants in case the first stage choice was Fanta.

Example 3.14

To illustrate further the typical choice situation, consider the choice of residential location. Specifically, suppose the agent is considering a move to one out of two cities, which includes

a specific location within the preferred city. Let U_{jk} denote the utility of location $k \in L_j$ within city j, $j = 1,2,$ where L_j is the set of relevant and available locations within city j. Let $U_{jk} = v_{jk} + \varepsilon_{jk}$, where

$$P\left(\bigcap_{k \in L_1} \left(\varepsilon_{1k} \le x_{1k}\right), \bigcap_{k \in L_2} \left(\varepsilon_{2k} \le x_{2k}\right) \right) = \exp\left(-G\left(e^{-x_{11}}, e^{-x_{12}}, \ldots, e^{-x_{21}}, e^{-x_{22}}, \ldots\right)\right) \tag{3.60}$$

and

$$G\left(y_{11}, y_{12}, \ldots, y_{21}, \ldots\right) = \sum_{j=1}^{2} \left(\sum_{k \in L_j} y_{jk}^{1/\theta_j} \right)^{\theta_j}. \tag{3.61}$$

The structure (3.61) implies that

$$\operatorname{corr}\left(\varepsilon_{jk}, \varepsilon_{jr}\right) = 1 - \theta_j^2, \text{ for } r \ne k, \tag{3.62}$$

and

$$\operatorname{corr}\left(\varepsilon_{jk}, \varepsilon_{ir}\right) = 0 \text{ for } j \ne i, \text{ and all } k \text{ and } r. \tag{3.63}$$

The interpretation of the correlation structure is that the alternatives within L_j are more "similar" than alternatives where one belongs to L_1 and the other belongs to L_2.

Let P_{jr} denote the joint probability of choosing location $r \in L_j$ and city j. Now from Theorem 8 we get that

$$\begin{aligned} P_{jr} &\equiv P\left(U_{jr} = \max_{i=1,2}\left(\max_{k \in L_k} U_{ik} \right)\right) = \frac{\partial G\left(e^{v_{11}}, e^{v_{12}}, \ldots\right) / \partial v_{jr}}{G\left(e^{v_{11}}, e^{v_{12}} \ldots\right)} \\ &= \frac{\left(\sum_{k \in L_j} e^{v_{jk}/\theta_j} \right)^{\theta_j - 1} e^{v_{jr}/\theta_j}}{\sum_{i=1}^{2} \left(\sum_{k \in L_i} e^{v_{ik}/\theta_i} \right)^{\theta_i}}. \end{aligned} \tag{3.64}$$

Note that we can rewrite (3.64) as

(3.65)
$$P_{jr} = \frac{\left(\sum_{k \in L_j} e^{v_{jk}/\theta_j} \right)^{\theta_j}}{\sum_{i=1}^{2} \left(\sum_{k \in L_i} e^{v_{ik}/\theta_i} \right)^{\theta_i}} \cdot \frac{e^{v_{jr}/\theta_j}}{\sum_{k \in L_j} e^{v_{jk}/\theta_j}} = P_j \cdot \frac{e^{v_{jr}/\theta_j}}{\sum_{k \in L_j} e^{v_{jk}/\theta_j}},$$

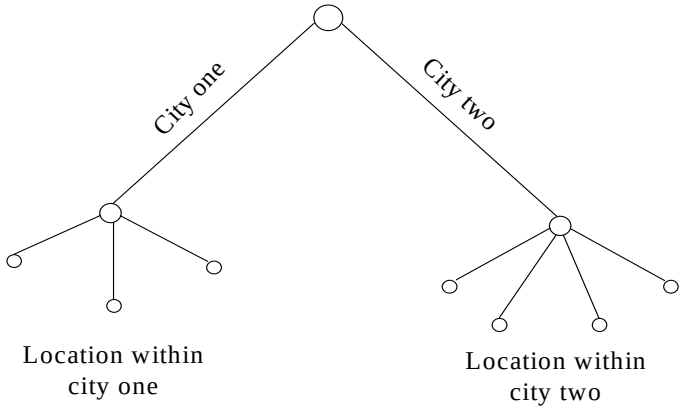
where

(3.66)
$$P_j = \sum_{k \in L_j} P_{jk} .$$

The probability P_j is the probability of choosing to move to city j (i.e. the optimal location lies within city j). Furthermore

(3.67)
$$\frac{P_{jr}}{P_j} = \frac{e^{v_{jr}/\theta_j}}{\sum_{k \in L_j} e^{v_{jk}/\theta_j}}$$

is the probability of choosing location $r \in L_j$, given that city j has been selected. We notice that P_{jr}/P_j does not depend on alternatives outside L_j. Thus the probability P_{jr} can be factored as a product consisting of the probability of choosing city j times the probability of choosing r from L_j, where the last probability has the same structure as the Luce model. However, this will not be the case if a subset different from L_1 and L_2 were selected in a first stage. Graphically, the above tree structure looks as follows:



So far no deep theoretical characterization of the GEV class of models has been given, apart from the property that it contains the Luce model as a special case. Specifically, and interesting

question is how restrictive the GEV class is. This issue has been addressed by Dagsvik (1994, 1995). He proves that any (additive) random utility model can be approximated arbitrarily closely by GEV models. In other words, one can approximate, as closely as desired, the choice probabilities of any (additive) random utility model by choice probabilities of a GEV model. This means that the GEV class represents no essential restrictions beyond being an additive random utility model.

3.9. The mixed logit model

Recently the so-called mixed logit model has become popular. This type of models is also known as random coefficient model. The idea of this approach is to allow the unknown parameters of the logit model be individual specific and distributed across the population according to some distribution function. The distribution function of the parameters may be specified parametrically or may be specified nonparametrically. McFadden and Train (2000) have shown that one can approximate any random utility model arbitrarily closely by mixed logit models.

To illustrate the idea explicitly, assume for example that one has specified the multinomial logit model conditional on the parameter vector β as in (3.17), that is

(3.68) $$P_j(\beta) = \frac{\exp(Z_j\beta)}{\sum_{k=1}^{m} \exp(Z_k\beta)}.$$

Then one obtains the unconditional choice probability by taking expectation with respect to the random vector β. That is, one "integrates out" with respect to the distribution of β. Thus, the resulting choice probability for choosing alternative j becomes

(3.69) $$P_j = E\{\frac{\exp(Z_j\beta)}{\sum_{k=1}^{m} \exp(Z_k\beta)}\}.$$

The econometrician's problem is now to estimate the unknown parameters in the distribution of β. Train (2003) discusses practical estimation techniques based on simulation methods.

4. Applications of discrete choice analysis

4.1. Labor supply

Consider the binary decision problem of choosing between the alternatives "working" and "not working". Take the standard neo-classical model as a point of departure. Let V(C,L) be the agent's utility in consumption, C, and annual leisure, L. The budget constraint equals

(4.1) $$C = hW + I$$

where W is the wage rate the agent faces in the market, h is annual hours of work and I is non-labor income (for example the income provided by the spouse). The time constraint equals

(4.2) $$h + L \le M\,(=8760).$$

According to this model utility maximization implies that the agent supplies labor if

(4.3) $$W > \frac{\partial_2 V(I,M)}{\partial_1 V(I,M)} \equiv W^*$$

where ∂_j denotes the partial derivative with respect to component j. If the inequality is reversed, then the agent will not wish to work. W^* is called the *reservation wage*. Suppose for example that the utility function has the form

(4.4) $$V(C,L) = \left(\frac{\left(C^{\alpha_1}-1\right)}{\alpha_1}\right)\beta_1 + \frac{\left(\left(\frac{L}{M}\right)^{\alpha_2}-1\right)}{\alpha_2}\beta_2 M,$$

where $\alpha_1 < 1,\ \alpha_2 < 1,\ \beta_1 > 0,\ \beta_2 > 0.$ Then V(C,L) is increasing and strictly concave in (C,L). The reservation wage equals

(4.5) $$W^* \equiv \frac{\partial_2 V(I,M)}{\partial_1 V(I,M)} = \frac{\beta_2}{\beta_1} I^{1-\alpha_1}.$$

After taking the logarithm on both sides of (4.3) and inserting (4.5) we get that the agent will supply labor if

(4.6) $$\log W > (1-\alpha_1)\log I + \log\left(\frac{\beta_2}{\beta_1}\right).$$

Suppose next that we wish to estimate the unknown parameters of this model from a sample of individuals of which some work and some do not work. Unfortunately, it is a problem with using (4.6) as a point of departure for estimation because the wage rate is not observed for those individuals that do not work. For all individuals in the sample we observe, say, age, non-labor income, length of education and number of small children. To deal with the fact that the wage rate is only observed for those agents who work, we shall next introduce a wage equation. Specifically, we assume that

$$\log W = X_1 a + \varepsilon_1 \tag{4.7}$$

where X_1 consists of length of education and age and a is the associate parameter vector. ε_1 is a random variable that accounts for unobserved factors that affect the wage rate, such as type of schooling, the effect of ability and family background, etc. We assume furthermore that the parameter β_2/β_1 depend on age and number of small children, X_2, such that

$$\log\left(\frac{\beta_2}{\beta_1}\right) = X_2 b + \varepsilon_2 \tag{4.8}$$

where ε_2 is a random term which accounts for unobserved variables that affect the preferences and b is a parameter vector. For simplicity we assume that α_1 is common to all agents. If ε_1 and ε_2 are independent and normally distributed with $E\,\varepsilon_j = 0,\ \ Var\,\varepsilon_j = \sigma_j^2,$ we get that the probability of working equals a probit model given by

$$P_2 \equiv P\left(W > W^*\right) = \Phi\left(\frac{Xs + (\alpha_1 - 1)\log I}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) \tag{4.9}$$

where $\Phi(\cdot)$ is the cumulative normal distribution function and s is a parameter vector such that $Xs = X_1 a - X_2 b$. From (4.9) we realize that only

$$\frac{s_j}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \text{and } \frac{\alpha_1 - 1}{\sqrt{\sigma_1^2 + \sigma_2^2}},\ k = 1,2,\ldots,$$

can be identified.

If the purpose of this model is to analyze the effect from changes in level of education, family size and non-labor income on the probability of supplying labor then we do not need to identify the remaining parameters. Let us write the model in a more convenient form;

$$P_2 = \Phi\left(Xs^* - c\log I\right), \tag{4.10}$$

where $c=(1-\alpha_1)/\sqrt{\sigma_1^2+\sigma_2^2}$ and $s_j^*=s_j/\sqrt{\sigma_1^2+\sigma_2^2}$. We have that

$$\frac{\partial \log P_2}{\partial \log I}=-c\frac{\Phi'(Xs^*-c\log I)}{\Phi(Xs^*-c\log I)}=-c\frac{\exp\left(-\frac{(Xs^*-c\log I)^2}{2}\right)}{\Phi(Xs^*-c\log I)\sqrt{2\pi}}. \tag{4.11}$$

Eq. (4.11) equals the elasticity of the probability of working with respect to in non-labor income.

Suppose alternatively that $\sigma_1=\sigma_2$ and that the random terms $\theta\varepsilon_1$ and $\theta\varepsilon_2$ are i.i. standard extreme value distributed. This means that $\theta=\pi/\sigma\sqrt{6}$, cf. Lemma A1. Then it follows that P_2 becomes a binary logit model given by

$$P_2=\frac{\exp(\theta E\log W)}{\exp(\theta E\log W)+\exp(\theta E\log W^*)}=\frac{1}{1+\exp(-Xs\theta+(1-\alpha_1)\theta\log I)}. \tag{4.12}$$

From (4.12) we now obtain the elasticity with respect to I as

$$\frac{\partial \log P_2}{\partial \log I}=-(1-\alpha_1)\theta(1-P_2)=-\frac{(1-\alpha_1)\theta}{1+\exp(Xs\theta-(1-\alpha_1)\theta\log I)}. \tag{4.13}$$

A further discussion on the application of discrete choice models in the analysis of labor supply is given by Dagsvik (2004).

4.2. Transportation

Suppose that commuters have the choice between driving own car or taking a bus. One is interested in estimating a behavioral model to study, for example, how the introduction of a new subway line will affect the commuters' transportation choices. Consider a particular commuter (agent) and let $U_j(x)$ be the agent's joint utility of commodity vector x and transportation alternative j, $j=1,2$. Assume that the utility function has the structure

$$U_j(x)=U_{1j}+\tilde{U}(x). \tag{4.14}$$

The budget constraint is given by

$$p'x=y-q_j,\ x\geq 0, \tag{4.15}$$

where p is a vector of commodity prices and q_j is the per-unit-cost of transportation. By maximizing $U_j(x)$ with respect to x subject to (4.15) we obtain the conditional indirect utility, given j, as

(4.16) $$V_j\left(p, y - q_j\right) = U_{1j} + V^*\left(p, y - q_j\right)$$

where the function $V^*(p,y)$ is defined by

(4.17) $$V^*(p,y) = \max_{p'x=y} \tilde{U}(x).$$

Assume that

(4.18) $$U_{1j} = \beta T_j + \varepsilon_j$$

where T_j is the travelling time with alternative j, β is an unknown parameter and $\{\varepsilon_j\}$ are random terms that account for the effect of unobserved variables, such as walking distances and comfort. We assume that ε_1 and ε_2 are i.i. standard extreme value distributed. Assume furthermore that

(4.19) $$V^*\left(p, y - q_j\right) = \tilde{V}(p) + \theta \log\left(y - q_j\right)$$

where $\theta > 0$ is an unknown parameter. The assumptions above yield

(4.20) $$V_j\left(p, y - q_j\right) = \beta T_j + \theta \log\left(y - q_j\right) + \tilde{V}(p) + \varepsilon_j$$

which implies that

(4.21) $$P_j(\{1,2\}) = \frac{\exp\left(\beta T_j + \theta \log\left(y - q_j\right)\right)}{\sum_{k=1}^{2} \exp\left(\beta T_k + \theta \log\left(y - q_k\right)\right)}$$

for $j = 1,2$. After the unknown parameters β and θ have been estimated one can predict the fraction of commuters that will choose the subway alternative (alternative 3) given that T_3 and q_3 have been specified. Here, it is essential that one believes that T_j and q_j are the main attributes of importance. We thus get that the probability of choosing alternative j from {1,2,3} equals

(4.22) $$P_j(\{1,2,3\}) = \frac{\exp\left(\beta T_j + \theta \log\left(y - q_j\right)\right)}{\sum_{k=1}^{3} \exp\left(\beta T_k + \theta \log\left(y - q_k\right)\right)}.$$

4.3. Potential demand for alternative fuel vehicles

This example is taken from Dagsvik et al. (2002). To assess the potential demand for alternative fuel vehicles such as; "electric" (1), "liquid propane gas" (lpg) (2), and "hybrid" (3), vehicles, an ordered logit model was estimated on the basis of a "stated preference" survey. In this survey each responent in a randomly selected sample was exposed to 15 experiments. In each experiment the respondent was asked to rank three hypothetical vehicles characterized by specified attributes, according to the respondent's preferences. These attributes are: "Purchase price", "Top speed", "Driving range between refueling/recharging", and "Fuel consumption". The total sample size (after the non-respondent individuals are removed) consisted of 662 individuals. About one half of the sample (group A) received choice sets with the alternatives "electric", "lpg", and "gasoline" vehicles, while the other half (group B) received "hybrid", "lpg" and "gasoline" vehicles. In this study "hybrid" means a combination of electric and gasoline technology. The gasoline alternative is labeled alternative 4.

The individuals' utility function was specified as

$$U_j(t) = Z_j(t)\beta + \mu_j + \varepsilon_j(t) \tag{4.23}$$

where $Z_j(t)$ is a vector consisting of the four attributes of vehicle j in experiment t, $t = 1,2,\ldots,15$, and μ_j and β are unknown parameters. Without loss of generality, we set $\mu_4 = 0$. As mentioned above group A has choice set, $C_A = \{1,2,4\}$, while group B has choice set, $C_B = \{2,3,4\}$. Let $P_{ijt}(C)$ be the probability that an individual shall rank alternative i on top and j second best in experiment t, and let $Y_{ij}^h(t) = 1$ if individual h ranks i on top and j second best in experiment t, and zero otherwise. From Theorem 3 it follows that if $\{\varepsilon_j(t)\}$ are assumed to be i.i. standard extreme value distributed then

$$P_{ijt}(C) = \frac{\exp\left(Z_i(t)\beta + \mu_i\right)}{\sum_{r\in C}\exp\left(Z_r(t)\beta + \mu_r\right)} \cdot \frac{\exp\left(Z_j(t)\beta + \mu_j\right)}{\sum_{r\in C\setminus\{i\}}\exp\left(Z_r(t)\beta + \mu_r\right)} \tag{4.24}$$

where C is equal to C_A or C_B,. We also assume that the random terms $\{\varepsilon_j(t)\}$ are independent across experiments. Consequently, it follows that the loglikelihood function has the form

$$\ell = \sum_{t=1}^{15}\left(\sum_{h\in A}\sum_i\sum_j Y_{ij}^h(t)\log P_{ijt}(C_A) + \sum_{h\in B}\sum_i\sum_j Y_{ij}^h(t)\log P_{ijt}(C_B)\right). \tag{4.25}$$

The sample is further split into six age and gender groups, and Table 4.1 displays the estimation results for these groups.

Table 4.1. Parameter estimates[*)] for the age/gender specific utility function

	Age					
	18-29		30-49		50-	
Attribute	Females	Males	Females	Males	Females	Males
Purchase price (in 100 000 NOK)	-2.530 (-17.7)	-2.176 (-15.2)	-1.549 (-15.0)	-2.159 (-20.6)	-1.550 (-11.9)	-1.394 (-11.8)
Top speed (100 km/h)	-0.274 (-0.9)	0.488 (1.5)	-0.820 (-3.3)	-0.571 (-2.4)	-0.320 (-1.1)	-0.339 (-1.2)
Driving range (1 000 km)	1.861 (3.1)	2.130 (3.3)	1.018 (2.0)	1.465 (3.2)	0.140 (0.2)	1.000 (1.8)
Fuel consumption (liter per 10 km)	-0.902 (-3.0)	-1.692 (-5.1)	-0.624 (-2.5)	-1.509 (6.7)	-0.446 (-1.5)	-1.030 (-3.7)
Dummy, electric	0.890 (4.2)	-0.448 (-2.0)	0.627 (3.6)	-0.180 (-1.1)	0.765 (3.6)	-0.195 (-1.0)
Dummy, hybrid	1.185 (7.6)	0.461 (2.8)	1.380 (10.6)	0.649 (5.6)	1.216 (7.7)	0.666 (4.6)
Dummy, lpg	1.010 (8.2)	0.236 (1.9)	0.945 (9.2)	0.778 (8.5)	0.698 (5.7)	0.676 (5.6)
# of observations	1380	1110	2070	2325	1290	1455
# of respondents	92	74	138	150	86	96
log-likelihood	2015.1	1747.8	3140.8	3460.8	2040.9	2333.8
McFadden's ρ^2	0.19	0.12	0.15	0.17	0.12	0.10

[*)] t-values in parenthesis.

Table 4.1 displays the estimates when the model parameters differ by gender and age. We notice that the price parameter is very sharply determined and it is slightly declining by age in absolute value. Most of the other parameters also decline by age in absolute value. However, when we take the standard error into account this tendency seems rather weak. Further, the utility function does not differ much by gender, apart from the parameters associated with fuel-consumption and the dummies for alternative fuel-cars. Specifically, males seem to be more sceptic towards alternative-fuel than females.

To check how well the model performs, we have computed McFadden's ρ^2 and in addition we have applied the model to predict the individuals' rankings. The prediction results are displayed in Tables 4.2 and 4.3, while McFadden's ρ^2 is reported in Table 4.1. We see that McFadden's ρ^2 has the highest values for young females, and for males with age between 30-49 years.

Table 4.2. Prediction performance of the model for group A. Per cent

	First choice			Second choice			Third choice		
Gender	Electric	Lpg	Gaso-line	Electric	Lpg	Gaso-line	Electric	Lpg	Gaso-line
Females:									
Observed	52.1	26.1	21.9	22.3	46.5	31.2	25.6	27.4	46.9
Predicted	45.6	36.3	18.1	32.8	38.5	28.8	21.6	25.3	53.2
Males:									
Observed	40.0	34.5	25.5	20.3	43.5	36.2	39.7	22.0	38.3
Predicted	32.6	44.2	23.3	32.1	35.5	32.4	35.3	20.3	44.3

Table 4.3. Prediction performance of the model group B. Per cent

	First choice			Second choice			Third choice		
Gender	Hybrid	Lpg	Gaso-line	Hybrid	Lpg	Gaso-line	Hybrid	Lpg	Gaso-line
Females:									
Observed	45.0	42.0	13.0	33.0	44.9	22.1	22.0	13.1	64.9
Predicted	43.0	40.3	16.7	36.9	37.8	25.3	20.1	21.9	58.0
Males:									
Observed	38.1	46.2	15.7	32.9	41.0	26.2	29.0	12.8	58.1
Predicted	35.3	45.2	19.5	37.4	35.0	27.6	27.3	19.8	52.9

The results in Table 4.3 show that for those individuals who receive choice sets that include the hybrid vehicle alternative (group B) the model fits the data reasonably well. For the other half of the sample for which the electric vehicle alternative is feasible (group A), Table 4.2 shows that the predictions fail by about 10 per cent points in four cases. Thus the model performs better for group B than for group A.

4.4. Oligopolistic competition with product differentiation

This example is taken from Anderson et al. (1992). Consider m firms which each produces a variant of a differentiated product. The firms' decision problem is to determine optimal prices of the different variants.

Assume that firm j produces at fixed marginal costs c_j and has fixed costs K_j. There are N consumers in the economy and consumer i has utility

(4.26) $$U_{ij} = y_i + a_j - w_j + \sigma \varepsilon_{ij} .$$

for variant j, where y_i is the consumers income, a_j is an index that captures the mean value of non-pecuniary attributes (quality) of variant j, w_j is the price of variant j, ε_{ij} is an individual-specific random taste-shifter that captures unobservable product attributes as well as unobservable individual-specific characteristics and $\sigma > 0$ is a parameter (unknown). If we assume that ε_{ij}, $j = 1,2,\ldots,m$, $i = 1,2,\ldots,N$, are i.i. standard extreme value distributed we get that the aggregate demand for variant j equals NP_j where

$$P_j = Q_j(\mathbf{w}) \equiv \frac{\exp\left(\frac{a_j - w_j}{\sigma}\right)}{\sum_{k=1}^{m} \exp\left(\frac{a_k - w_k}{\sigma}\right)}. \tag{4.27}$$

Assume next that the firm knows the mean fractional demands $\{Q_j(\mathbf{w})\}$ as a function of prices, $\mathbf{w}$. Consequently, a firm that produces variant j can calculate expected profit, π_j, conditional on the prices;

$$\pi_j = (w_j - c_j) N Q_j(\mathbf{w}) - K_j. \tag{4.28}$$

Now firm j takes the prices set by other firms as given and chooses the price of variant j that maximizes (4.28). Anderson et al. (1992) demonstrate that there exists a unique Nash equilibrium set of prices, $\mathbf{w}^* = (w_1^*, w_2^*, \ldots, w_m^*)$ which are determined by

$$w_j^* = c_j + \frac{\sigma}{1 - Q_j(\mathbf{w}^*)}. \tag{4.29}$$

4.5. Social network

This example is borrowed from Dagsvik (1985). In the time-use survey conducted by Statistics Norway, 1980-1981, the survey respondents were asked who they would turn to if they needed help. The respondents were divided into two age groups, where group (i) and (ii) consist of individuals less than 45 years of age and more than 45 years of age, respectively. Here, we shall only analyze the subsample of individuals less than 45 years of age. The univers of alternatives S consisted of five alternatives, namely

$$S = \{\text{Mother}(1), \text{father}(2), \text{brother}(3), \text{sister}(4), \text{neighbor}(5)\}.$$

However, the set of feasible alternatives (choice set) were less for many of the respondents. Specifically, there turn out to be 11 different choice sets in the sample; $B_1, B_2, \ldots, B_{11}$. The data for each of the 11 groups are given in Table 4.5. Group (i) consists of 526 individuals.

The question is whether the above data can be rationalized by a choice model. To this end we first estimated a logit model

$$P_j(B_k) = \frac{e^{v_j}}{\sum_{r \in B_k} e^{v_r}}, \; j \in B_k, \tag{4.30}$$

where $k = 1,2,\ldots,11,$ and $v_5 = 0.$ Thus this model contains four parameters to be estimated. Let $\hat{P}_{jk}$ be the observed choice frequencies conditional on choice set B_k. Let ℓ^* denote the loglikelihood obtained when the respective choice probabilities are estimated by $\hat{P}_{jk}$, $j \in B_k$. From Table 4.5 it follows that $\ell^* = -405.8.$ In the logit model there are four free parameters, while there are 24 "free" probabilities in the 11 multinomial models in the a priori statistical model. Consequently, if ℓ_1 denotes the loglikelihood under the hypothesis of a logit model it follows that $-2(\ell_1 - \ell^*)$ is (asymptotically) Chi squared distributed with 20 degrees of freedom. Since the corresponding critical value at 5 per cent significance level equals 31.4 it follows from estimation results reported in Table 4.4 that the logit model is rejected against the non-structural multinomial model. One interesting hypothesis that might explain this rejection is that alternative five ("neighbor") differs from the "family" alternatives in the sense that the family alternatives depend on a latent variable which represents the "family aspect", that make the family alternatives more "close" than non-family alternatives. As a consequence, the family alternatives will have correlated utilities. To allow for this effect we postulate a nested logit structure with utilities that are correlated for the family alternatives. Specifically, we assume that

$$\text{corr}(U_i, U_j) = 1 - \theta^2, \tag{4.31}$$

for $i \neq j,$ $i, j \neq 5,$ and

$$\text{corr}(U_i, U_5) = 0, \tag{4.32}$$

for $i < 5,$ where $0 < \theta \leq 1.$ This yields

$$P_j(B) = \frac{e^{v_j/\theta}}{\sum_{r \in B} e^{v_r/\theta}} \tag{4.33}$$

when $B \ni 5$,

$$P_j(B) = \frac{e^{v_j/\theta}\left(\sum_{r\in B\setminus\{5\}} e^{v_r/\theta}\right)^{\theta-1}}{e^{v_5} + \left(\sum_{r\in B\setminus\{5\}} e^{v_r/\theta}\right)^{\theta}} \tag{4.34}$$

when $j \neq 5,\ 5 \in B,$ and

$$P_5(B) = \frac{e^{v_5}}{e^{v_5} + \left(\sum_{r\in B\setminus\{5\}} e^{v_r/\theta}\right)^{\theta}}. \tag{4.35}$$

As above we set $v_5 = 0$.

The parameter estimates in the nested logit case are also given in Table 4.4. We notice that while only v_1 and v_4 are precisely determined in the logit case all the parameters are rather precisely determined in the nested logit case. The estimate of θ implies that the correlation between the utilities of the family alternatives equals 0.79.

From Table 4.4 we find that twice the difference in loglikelihood between the two models equals 17.6. Since the critical value of the Chi squared distribution with one degree of freedom at 5 per cent level equals 3.8, it follows that the logit model is rejected against the nested logit alternative.

As above we can also compare the nested logit model to the non-structural multinomial model. Let ℓ_2 denote the loglikelihood of the nested logit model. Since the nested logit model has five parameters it follows that $-2\left(\ell_2 - \ell^*\right)$ is (asymptotically) Chi squared distributed with 19 degrees of freedom (under the hypothesis of the nested logit model). The corresponding critical value is 30.1 at 5 per cent significance level and therefore the estimate of $-2\left(\ell_2 - \ell^*\right)$ in Table 4.4 implies that the nested logit model is not rejected against the non-structural multinomial model. As measured by McFaddens ρ^2, the difference in goodness-of-fit is only one per cent.

Table 4.4. Parameter estimates

	Logit model		Nested logit model	
Parameters	Estimates	t-values	Estimates	t-values
v_1	2.119	18.9	1.932	31.8
v_2	-0.519	0.7	0.654	5.5
v_3	0.099	0.2	0.801	8.3
v_4	0.725	4.8	1.242	16.8
θ			0.455	15.0
loglikelihood ℓ_j	-424.9		-416.1	
McFadden's ρ^2	0.33		0.34	
$-2\left(\ell_j - \ell^*\right)$	38.2		20.6	

In Table 4.5 we report the data and the prediction performance of the two model versions. The table shows that the nested logit model predicts the fractions of observed choices rather well.

At this point it is perhaps of interest to recall the limitation of this type of statistical significance testing. Of course, when the sample size increases we will always get rejection of the null hypothesis of a "perfect model". Since we already know that our models are more or less crude approximations to the "true model", this is as it should be, but is hardly very interesting. What, however, is of interest is how the model performs in predictions, preferably out-of-sample predictions.

Since the logit and the nested-logit model predict almost equally well within sample, it is not possible to discriminate between the two models on the basis of (aggregate) predictions. One argument that supports the selection of the nested logit model is that even if this model contains an additional parameter, the precision of the estimates is considerably higher than in the case of the logit model. This suggests that the nested logit model captures more of the "true" underlying structure than the logit model.

Table 4.5. Prediction performance of the logit- and the nested logit model

Choice sets			Alternatives 1 Mother	2 Father	3 Brother	4 Sister	5 Neighbor	# observations
	Observed		30	NF	NF	NF	6	36
B_1	Predicted	Logit	32.1	NF	NF	NF	3.9	
	Predicted	Nested logit	31.4	NF	NF	NF	4.6	
	Observed		NF	NF	36	NF	20	56
B_2	Predicted	Logit	NF	NF	29.4	NF	26.6	
	Predicted	Nested logit	NF	NF	38.6	NF	17.3	
	Observed		21	NF	2	NF	1	24
B_3	Predicted	Logit	19.2	NF	2.5	NF	2.3	
	Predicted	Nested logit	19.4	NF	1.5	NF	2.9	
	Observed		NF	NF	9	21	2	32
B_4	Predicted	Logit	NF	NF	8.5	15.8	7.7	
	Predicted	Nested logit	NF	NF	7.0	18.6	6.4	
	Observed		NF	5	NF	NF	2	7
B_5	Predicted	Logit	NF	2.6	NF	NF	4.4	
	Predicted	Nested logit	NF	4.6	NF	NF	2.4	
	Observed		65	3	NF	NF	10	78
B_6	Predicted	Logit	65.4	4.7	NF	NF	7.9	
	Predicted	Nested logit	64.5	3.9	NF	NF	9.6	
	Observed		50	4	4	NF	6	64
B_7	Predicted	Logit	48.3	3.5	6.4	NF	5.8	
	Predicted	Nested logit	49.2	3.0	4.1	NF	7.7	
	Observed		23	NF	NF	7	8	38
B_8	Predicted	Logit	27.8	NF	NF	6.9	3.3	
	Predicted	Nested logit	27.5	NF	NF	6.0	4.4	
	Observed		45	2	NF	5	8	60
B_9	Predicted	Logit	41.7	3.0	NF	10.3	5	
	Predicted	Nested logit	41.5	2.5	NF	9.1	6.8	
	Observed		21	NF	2	6	8	37
B_{10}	Predicted	Logit	24.7	NF	3.3	6.1	3.0	
	Predicted	Nested logit	25.2	NF	2.1	5.5	4.2	
	Observed		64	4	5	15	6	94
B_{11}	Predicted	Logit	60.0	4.3	7.9	14.8	7.2	
	Predicted	Nested logit	61.3	3.7	5.1	13.4	10.5	

NF = Not feasible.

5. Maximum likelihood estimation of multinomial probability models

Suppose the multinomial probability model has been specified. Let $Y_{ij} = 1$, if agent i in a sample of randomly selected agents, falls into category j and zero otherwise, and let

$$P(Y_{ij} = 1 \mid \mathbf{Z}, X_i) = H_j(\mathbf{Z}, X_i; \beta)$$

$\{H_j(X_i;\boldsymbol{\beta})\}$ be the corresponding multinomial logit probabilities, where X_i is the vector of individual characteristics for agent i and $\mathbf{Z} = (Z_1, Z_2,\ldots,Z_m)$. The total likelihood of the observed outcome equals

$$\prod_{i=1}^{N} \prod_{j=1}^{m} H_j(Z, X_i; \beta)^{Y_{ij}}$$

where N is the sample size. The loglikelihood function can therefore be written as

(5.1) $$\ell = \sum_{i=1}^{N} \sum_{j=1}^{m} Y_{ij} \log H_j(Z, X_i; \beta).$$

By the maximum likelihood principle the unknown parameters are estimated by maximizing ℓ with respect to the unknown parameters.

The logit structure implies that the first order conditions of the loglikelihood function equals

(5.2) $$\frac{\partial \ell}{\partial \beta_{rk}} = \sum_{i=1}^{N} (Y_{ik} - H_r(Z, X_i; \beta) X_{ik}) = 0$$

for $r = 2,3,\ldots, m$, $k = 1,2,\ldots, K$, where X_{ik} is the k-th component component of X_i, with associated coefficient β_{rk}.

5.1. Estimation of the multinomial logit model

Suppose next that the logit model has the structure

$$H_j(\mathbf{Z}, X_i; \boldsymbol{\beta}) = \frac{\exp\left(h\left(Z_j, X_i\right)\beta\right)}{\sum_{k=1}^{m} \exp\left(h\left(Z_k, X_i\right)\beta\right)} \tag{5.3}$$

where

$$h\left(Z_j, X_i\right)\beta = \sum_{r=1}^{K} h_r\left(Z_j, X_i\right)\beta_r \,. \tag{5.4}$$

Examples of this structure were given in Section 3.5. Note that in this case the parameters are *not* alternative-specific.

When the logit model has the structure given by (5.3) and (5.4), then the first order conditions yield

$$\frac{\partial \ell}{\partial \beta_k} = \sum_{i=1}^{N} \sum_{j=1}^{m} \left(Y_{ij} - H_j\left(\mathbf{Z}, X_i; \boldsymbol{\beta}\right)\right) h_k\left(Z_j, X_i\right) = 0 \tag{5.5}$$

for $k = 1,2,\ldots,K$.

McFadden (1973) has proved that when the probabilities are given by (5.3) and (5.4), the loglikelihood function is globally strictly concave, and therefore a unique solution to (5.5) is guarantied.

5.2. Berkson's method (Minimum logit chi-square method)

If we have a case with several observations for each value of the explanatory variable it is possible to carry out estimation by Berkson's method (Berkson, 1953). Model (3.17) in Example 3.1 is an example of a case where this method is applicable, since this model does not depend on individual characteristics. Let

$$\hat{H}_j = \frac{1}{N} \sum_{i=1}^{N} Y_{ij}$$

and replace H_j by $\hat{H}_j$ in (3.17). We then obtain

(5.6) $$\log\left(\frac{\hat{H}_j}{\hat{H}_1}\right) = \left(Z_j - Z_1\right)\beta + \eta_j ,$$

where η_j is a random error term. By the strong law of large numbers $\hat{H}_j \to H_j$ with probability one as the sample size increases, the error term η_j will be small when N is "large". Also by first order Taylor approximation we get

$$\log\left(\frac{\hat{H}_j}{\hat{H}_1}\right) = \log\hat{H}_j - \log\hat{H}_1 \approx \log\left(\frac{H_j}{H_1}\right) + \frac{\left(\hat{H}_j - H_j\right)}{H_j} - \frac{\left(\hat{H}_1 - H_1\right)}{H_1}$$

which shows that

(5.7)
$$\begin{aligned} E\eta_j &= E\log\left(\frac{\hat{H}_j}{\hat{H}_1}\right) - \left(Z_j - Z_1\right)\beta \\ &\approx \log\left(\frac{H_j}{H_1}\right) + \frac{E\hat{H}_j - H_j}{H_j} - \frac{\left(E\hat{H}_1 - H_1\right)}{H_1} - \left(Z_j - Z_1\right)\beta \\ &= \log\left(\frac{H_j}{H_1}\right) - \left(Z_j - Z_1\right)\beta = 0. \end{aligned}$$

Thus, even in samples of limited size the mean of the error terms $\{\eta_j\}$ is approximately equal to zero. Define the dependent variable $\tilde{Y}_j$ by

$$\tilde{Y}_j = \log\left(\frac{\hat{H}_j}{\hat{H}_1}\right).$$

We now realize that due to (5.6) we can estimate β by regression analysis with $\{\tilde{Y}_j\}$ as dependent variables and $\{Z_j - Z_1\}$ as independent variables. However, the error terms in (5.6) are correlated with covariance matrix that depends on the probabilities. Therefore one needs to apply GLS methods to obtain efficient estimation. See Maddala (1983, p. 30) for a more detailed treatment of Berkson's method.

6. The nonstructural Tobit model

In this section we shall describe a type of statistical model, usually called the Tobit model. The Tobit model (Tobin, 1958) is specified as follows: The dependent variable Y is defined by

$$Y = \begin{cases} X\beta + u\sigma \text{ if } X\beta + u\sigma > 0 \\ 0 \quad \text{otherwise}, \end{cases} \tag{6.1}$$

where $\sigma > 0$ is a scale parameter, and u is a zero mean random variable with cumulative distribution function F(·). Another way of expressing (6.1) is as

$$Y = \max(0, X\beta + u\sigma). \tag{6.2}$$

Tobin (1958) assumed that u is normally distributed N(0,1), but it is also convenient to work with the logistic distribution.

An example of a Tobit formulation is the standard labor supply model. Here we may interpret $X\beta c + u\sigma c$ as an index that measures the desire to work of an agent with characteristics X. Specifically, one may interpret $X\beta c + u\sigma c$ as the difference between the utility of working and the utility of not working. When this index is positive, the desired hours of work is typically assumed proportional to $X\beta c + u\sigma c$ where 1/c is the proportionality factor. The variable vector X may contain education, work experience, and the unobservable term u may capture the effect of unobservable variables such as specific skills and training. When the index $X\beta c + u\sigma c$ is negative and large, say, it means that the agent has strong tendence to choose leisure. Since the actual hours of work always will be non-negative we therefore get the structure (6.1).

As regards structural models, see for example Hanemann (1984) and Dubin and McFadden (1984) and McFadden 1981) who discuss multivariate structural discrete/continuous choice models of the Tobit type.

6.1. Maximum likelihood estimation of the Tobit model

Notice first that due to the form of (6.2) ordinary regression analysis will not do because of the nonlinear operation on the right hand side of (6.2).

From (6.2) it follows that

(6.3) $$P(Y=0)=P(u\le -X\beta/\sigma)=F(-X\beta/\sigma)$$

where F(y) denotes the cumulative distribution of u, and

(6.4) $$P(Y\in(y,y+dy))=P(u\sigma\in(y-X\beta,y+dy-X\beta))=\frac{1}{\sigma}F'\left(\frac{y-X\beta}{\sigma}\right)dy,$$

for $y>0$. Consider now the estimation of the unknown parameters based on observations from a random sample of individuals, and as above, let $i=1,2,\ldots$ be an indexation of the individuals in the sample. Let S_1 be the set of N_1 individuals for which $Y_i>0$ and S_0 the remaining set of individuals for whom $Y_i=0$. We shall distinguish between two cases, namely the cases where we observe X_i and Y_i for all the individuals (Case I), and the case where we do not observe X_i when $i\in S_0$ (Case II).

<u>Case I</u>: X_i is observed for all $i\in S_0\cup S_1$ (Censored case)

From (6.4) it follows that the density of Y_i when $Y_i>0$ equals

$$F'\left(\frac{y-X_i\beta}{\sigma}\right)\frac{1}{\sigma}$$

while, by (6.3), the probability that $i\in S_0$ equals

$$F\left(\frac{-X_i\beta}{\sigma}\right).$$

Therefore the total loglikelihood equals

(6.5) $$\ell=\sum_{i\in S_1}\left(\log F'\left(\frac{Y_i-X_i\beta}{\sigma}\right)-\log\sigma\right)+\sum_{i\in S_0}\log F\left(\frac{-X_i\beta}{\sigma}\right).$$

Example 6.1

Suppose F(y) is a standard normal distribution function, Φ(y). Then, since

$$\Phi'(u)=\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$$

it follows that the loglikelihood in this case reduces to

(6.6) $$\ell = -\sum_{i\in S_1} \frac{(Y_i - X_i\beta)^2}{2\sigma^2} - N_1 \log\sigma + \sum_{i\in S_0} \log\Phi\left(\frac{-X_i\beta}{\sigma}\right) - \frac{N_1}{2}\log(2\pi).$$

We realize that applying OLS to the equation $Y = X\beta + u\sigma$ corresponds to neglecting the last term in (6.6) and will therefore produce biased estimates.

Example 6.2

Suppose that F(y) is a standard logistic distribution, L(y), given by (2.12). Since $1 - L(-y) = L(y)$ and

(6.7) $$L'(y) = L(y)(1 - L(y))$$

it follows from (6.5) that the loglikelihood function in this case is

(6.8) $$\ell = \sum_{i\in S_1} \left(\log L\left(\frac{Y_i - X_i\beta}{\sigma}\right) + \log\left(1 - L\left(\frac{Y_i - X_i\beta}{\sigma}\right)\right)\right) - N_1 \log\sigma + \sum_{i\in S_0} \log L\left(\frac{-X_i\beta}{\sigma}\right).$$

<u>Case II</u>: X_i is not observed for $i \in S_0$ (Truncated case)

In this case we must evaluate the conditional likelihood function given that the individuals belong to S_1. The conditional probability of $Y_i \in (y, y + dy), y > 0,$ given that $Y_i > 0$ equals

$$P(Y_i \in (y, y+dy) \mid Y_i > 0) = \frac{P(Y_i \in (y, y+dy), Y_i > 0)}{P(Y_i > 0)} = \frac{P(Y_i \in (y, y+dy))}{P(Y_i > 0)} = \frac{F'\left(\frac{y - X_i\beta}{\sigma}\right)\frac{1}{\sigma}dy}{1 - F\left(\frac{-X_i\beta}{\sigma}\right)}.$$

Therefore, the conditional loglikelihood given that $Y_i > 0$ for all i, equals

(6.9) $$\ell = \sum_{i\in S_1} \left(\log F'\left(\frac{Y_i - X_i\beta}{\sigma}\right) - \log\left(1 - F\left(\frac{-X_i\beta}{\sigma}\right)\right)\right) - N_1 \log\sigma.$$

6.2. Estimation of the Tobit model by Heckman's two stage method

Heckman (1979) suggested a two stage method for estimating the tobit model. We shall briefly review his method for the case where F(y) is either the normal distribution or the logistic distribution.

6.2.1. Heckman's method with normally distributed random terms

As above $\Phi(\cdot)$ denotes the cumulative normal distribution function. From (6.2) we get

$$\mathrm{E}\left(\mathrm{Y}\,|\,\mathrm{Y}>0\right)=\mathrm{X}\beta+\sigma\,\mathrm{E}\left(\mathrm{u}\,|\,\mathrm{Y}>0\right). \tag{6.10}$$

Since $\mathrm{E}\left(\mathrm{u}\,|\,\mathrm{Y}>0\right)$ in general is different from zero we cannot, as mentioned above, do linear regression analysis based on the subsample of individuals in S_1. Now note that

$$\begin{aligned}\mathrm{P}\left(\mathrm{u}\in(\mathrm{y},\mathrm{y}+\mathrm{dy})\,|\,\mathrm{Y}>0\right)&=\mathrm{P}\left(\mathrm{u}\in(\mathrm{y},\mathrm{y}+\mathrm{dy})\,\middle|\,\mathrm{u}>-\frac{\mathrm{X}\beta}{\sigma}\right)\\ &=\frac{\mathrm{P}\left(\mathrm{u}\in(\mathrm{y},\mathrm{y}+\mathrm{dy}),\mathrm{u}>-\frac{\mathrm{X}\beta}{\sigma}\right)}{\mathrm{P}\left(\mathrm{u}>-\frac{\mathrm{X}\beta}{\sigma}\right)}=\frac{\mathrm{P}\left(\mathrm{u}\in(\mathrm{y},\mathrm{y}+\mathrm{dy})\right)}{\mathrm{P}\left(-\mathrm{u}<\frac{\mathrm{X}\beta}{\sigma}\right)}=\frac{\Phi'(\mathrm{y})\,\mathrm{dy}}{\Phi\left(\frac{\mathrm{X}\beta}{\sigma}\right)}\end{aligned} \tag{6.11}$$

since -u has the same distribution as u due to symmetry. We therefore get

$$\mathrm{E}\left(\mathrm{u}\,|\,\mathrm{Y}>0\right)=\frac{1}{\Phi\left(\frac{\mathrm{X}\beta}{\sigma}\right)}\int_{-\frac{\mathrm{X}\beta}{\sigma}}^{\infty}\mathrm{u}\,\Phi'(\mathrm{u})\,\mathrm{du}. \tag{6.12}$$

But

$$\int_{-\frac{\mathrm{X}\beta}{\sigma}}^{\infty}\mathrm{u}\,\Phi'(\mathrm{u})\,\mathrm{du}=\int_{-\frac{\mathrm{X}\beta}{\sigma}}^{\infty}\frac{\mathrm{u}\,\mathrm{e}^{-\frac{\mathrm{u}^2}{2}}}{\sqrt{2\pi}}\,\mathrm{du}=-\int_{-\frac{\mathrm{X}\beta}{\sigma}}^{\infty}\frac{\mathrm{e}^{-\frac{\mathrm{u}^2}{2}}}{\sqrt{2\pi}}=\frac{1}{\sqrt{2\pi}}\cdot\exp\left(-\left(\frac{\mathrm{X}\beta}{\sigma}\right)^2\Big/2\right)=\Phi'\left(\frac{\mathrm{X}\beta}{\sigma}\right) \tag{6.13}$$

which together with (6.11) yields

$$\mathrm{E}\left(\mathrm{u}\,|\,\mathrm{Y}>0\right)=\frac{\Phi'\left(\frac{\mathrm{X}\beta}{\sigma}\right)}{\Phi\left(\frac{\mathrm{X}\beta}{\sigma}\right)}\equiv\lambda\left(\frac{\mathrm{X}\beta}{\sigma}\right) \tag{6.14}$$

where the last notation (λ) is introduced for convenience.

Heckman suggested the following approach: First estimate β/σ by probit analysis, i.e., by maximizing the likelihood with the dependent variable equal to one if $\mathrm{i}\in S_1$ and zero otherwise. The corresponding loglikelihood equals

(6.15) $$\ell = \sum_{i \in S_1} \log \Phi\left(\frac{X_i\beta}{\sigma}\right) + \sum_{i \in S_0} \log\left(1 - \Phi\left(\frac{X_i\beta}{\sigma}\right)\right).$$

From the estimates β^* of β/σ, compute

$$\hat{\lambda}_i = \frac{\Phi'\left(X_i\beta^*\right)}{\Phi\left(X_i\beta^*\right)}$$

and estimate β and σ by regression analysis on the basis of

(6.16) $$Y_i = X_i\beta + \sigma\hat{\lambda}_i + \eta_i$$

by applying the observations from S_1. This gives unbiased estimates because it follows from (6.10) and (6.14) that

$$\begin{aligned} &E\left(\eta_i \mid Y_i > 0\right) = E\left(Y_i - X_i\beta - \sigma\hat{\lambda}_i \mid Y_i > 0\right) \\ &= E\left(\sigma u_i - \sigma\hat{\lambda}_i \mid Y_i > 0\right) = \sigma\, E\left(u_i \mid Y_i > 0\right) - \sigma\hat{\lambda}_i \\ &= \sigma\lambda\left(\frac{X_i\beta}{\sigma}\right) - \sigma\hat{\lambda}_i \approx 0. \end{aligned}$$

Heckman (1979) has also obtained the asymptotic covariance matrix of the parameter estimates that take into account that one of the regressors, λ_i, is represented by the estimate, $\hat{\lambda}_i$.

Note that this procedure leads to two separate estimates of σ, namely the one obtained as a regression coefficient in (7.21) and the one that follows by dividing the mean component value of the estimated β by the corresponding mean based on β^*.

6.2.2 Heckman's method with logistically distributed random term

Assume now that u is distributed according to the logistic distribution L(y). Then by Lemma A3 in Appendix A it is proved that

(6.17) $$E\left(u \mid Y > 0\right) = \left(1 + \exp(-X\beta / \sigma)\right)\log\left(1 + \exp(X\beta / \sigma)\right) - X\beta / \sigma.$$

In this case the regression model that corresponds to (6.21) equals

(6.18) $$Y_i = X_i\beta + \sigma\hat{\theta}_i + \tilde{\eta}_i$$

where

(6.19) $$\hat{\theta}_i = \left(1+\exp\left(-X_i\beta^*\right)\right)\log\left(1+\exp\left(X_i\beta^*\right)\right) - X_i\beta^*$$

and β^* is the first stage maximum likelihood estimate of β/σ based on the binary logit model with loglikelihood equal to (6.15) with $\Phi(y)$ replaced by $L(y)$.

A modified version of Heckman's method

Since

$$P(Y>0) = \frac{1}{1+\exp\left(-X\beta/\sigma\right)}$$

it follows from (6.17) that

(6.20)
$$\begin{aligned} EY &= P(Y>0)\left(E\left(u \mid Y>0\right)\sigma + X\beta\right) \\ &= \sigma\log\left(1+\exp\left(X\beta/\sigma\right)\right) \\ &= \sigma\log\left(1+\exp\left(-X\beta/\sigma\right)\right) + X\beta = X\beta - \sigma\log P(Y>0). \end{aligned}$$

Eq. (6.20) implies that we may alternatively apply regression analysis on the whole sample based on the regression equation

(6.21) $$Y_i = X_i\beta + \sigma\hat{\mu}_i + \delta_i$$

where

(6.22) $$\hat{\mu}_i = \log\left(1+\exp\left(-X_i\beta^*\right)\right)$$

and δ_i is an error term with zero mean. This is so because (6.20) implies that

$$E\,\delta_i = E\left(Y_i - X_i\beta + \sigma\log P\left(Y_i>0\right)\right) = 0.$$

With the present state of computer software, where maximum likelihood procedures are readily available and easy to apply, Heckman's two stage approach may thus be of less interest.

6.3. The likelihood ratio test

The likelihood ratio test is a very general method which can be applied in wide variety of cases. A typical null hypothesis (H) is that there are specific constraints on the parameter values. For example, several parameters may be equal to zero, or two or more parameters may be equal to each other. Let $\hat{\beta}^H$ denote the constrained maximum likelihood estimate obtained when the likelihood is maximized

subject to the restrictions on the parameters under H. Similarly, let $\hat{\beta}$ denote the parameter estimate obtained from unconstrained maximization of the likelihood. Let $\ell\left(\hat{\beta}^{H}\right)$ and $\ell\left(\hat{\beta}\right)$ denote the loglikelihood values evaluated at $\hat{\beta}^{H}$ and $\hat{\beta}$, respectively. Let r be the number of independent restrictions implied by the null hypothesis. By "independent restrictions" it is meant that no restriction should be a function of the other restrictions. It can be demonstrated that under the null hypothesis

$$-2\left(\ell\left(\hat{\beta}^{H}\right)-\ell\left(\hat{\beta}\right)\right)$$

is asymptotically chi squared distributed with r degrees of freedom. Thus, if $-2\left(\ell\left(\hat{\beta}^{H}\right)-\ell\left(\hat{\beta}\right)\right)$ is "large" (i.e. exceeds the critical value of the chi squared with r degrees of freedom), then the null hypothesis is rejected.

In the literature, other types of tests, particularly designed for testing the "Independence from Irrelevant Alternatives" hypothesis have been developed. I refer to Ben-Akiva and Lerman (1985, p. 183), for a review of these tests.

6.4. McFadden's goodness-of-fit measure

As a goodness-of-fit measure McFadden has proposed a measure given by

(6.23) $$\rho^{2}=1-\frac{\ell\left(\hat{\beta}\right)}{\ell(0)}$$

where, as before, $\ell\left(\hat{\beta}\right)$ is the unrestricted loglikelihood evaluated at $\hat{\beta}$ and $\ell(0)$ is the loglikelihood evaluated by setting all parameters equal to zero. A motivation for (6.23) is as follows: If the estimated parameters do no better than the model with zero parameters then $\ell\left(\hat{\beta}\right)=\ell(0)$, and thus $\rho^{2}=0$. This is the lowest value that ρ^{2} can take (since if $\ell\left(\hat{\beta}\right)$ is less than $\ell(0)$, then $\hat{\beta}$ would not be the maximum likelihood estimate). Suppose instead that the model was so good that each outcome in the sample could be predicted perfectly. Then the corresponding likelihood would be one which means that the loglikelihood $\ell\left(\hat{\beta}\right)$ is equal to zero. Thus in this case $\rho^{2}=1,$ which is the highest value ρ^{2} can take. This goodness-of-fit measure is similar to the familiar R^{2} measure used in regression analysis in that it ranges between zero and one. However, there are no general guidelines for when a ρ^{2} value is sufficiently high.

Appendix A

Some properties of the extreme value and the logistic distributions

In this appendix we collect some classical results about the logistic and the extreme value distributions.

Let $X_1, X_2, \ldots,$ be independent random variables with a common distribution function F(x). Let

(A.1) $$M_n = \max(X_1, X_2, \ldots, X_n).$$

Theorem A1

Suppose that, for some $\alpha > 0$,

(A.2) $$\lim_{x \to \infty} x^{\alpha} (1 - F(x)) = c,$$

where $c > 0$. *Then*

(A.3) $$\lim_{n \to \infty} P\left(\frac{M_n}{(cn)^{1/\alpha}} \le x \right) = \begin{cases} exp\left(-x^{-\alpha}\right) & for \quad x > 0, \\ 0 & for \quad x \le 0. \end{cases}$$

Theorem A2

Suppose that for some x_0,, $F(x_0) = 1$, *and that for some* $\alpha > 0$,

(A.4) $$\lim_{x \to x_0} (x_0 - x)^{-\alpha} (1 - F(x)) = c,$$

where $c > 0$. *Then*

(A.5) $$\lim_{n \to \infty} P\left(\frac{M_n - x_0}{(cn)^{1/\alpha}} \le x \right) = \begin{cases} exp\left(-|x|^{\alpha}\right) & for \quad x < 0 \\ 1 & for \quad x \ge 0. \end{cases}$$

Theorem A3

Suppose that

(A.6) $$\lim_{x\to\infty} e^{x}\left(1-F(x)\right)=c\,,$$

where $c>0$ *. Then*

(A.7) $$\lim_{n\to\infty} P\left(M_n-\log(cn)\le x\right)=\exp\left(-e^{-x}\right)$$

for all x.

Proofs of Theorems A1 to A3 are found in Lamperti (1996), for example. Moreover, it can be proved that the distributions (A.3), (A.5) and (A.7) are the only ones possible.

The three classes of limiting distributions for maxima were discovered during the 1920s by M. Fréchet, R.A. Fisher and L.H.C. Tippett. In 1943 B. Gnedenko gave a systematic exposition of limiting distributions of the maximum of a random sample.

Note that there is some similarity between the Central Limit Theorem and the results above in that the limiting distributions are, apart from rather general conditions, independent of the original distribution. While the Central Limit Theorem yields only one limiting distribution, the limiting distributions of maxima are of three types, depending on the tail behavior of the distribution. The three types of distributions (A.3), (A.5) and (A.7) are called standard type I, II and III *extreme value* distributions, cf. Resnick (1987).

The extreme value distributions have the following property: if X_1 and X_2 are type III independent extreme value distributed with different location parameters, i.e.,

$$\mathrm{P}\left(\mathrm{X_j}\le \mathrm{x_j}\right)=\exp\left(-\mathrm{e}^{\mathrm{b_j}-\mathrm{x_j}}\right)$$

where b_1 and b_2 are constants, then $\mathrm{X}\equiv\max\left(\mathrm{X_1},\mathrm{X_2}\right)$ is also type III extreme value distributed. This is seen as follows: We have

$$\begin{aligned}\mathrm{P}\left(\mathrm{X}\le\mathrm{x}\right)&=\mathrm{P}\left(\left(\mathrm{X_1}\le\mathrm{x}\right)\cap\left(\mathrm{X_2}\le\mathrm{x}\right)\right)\\&=\mathrm{P}\left(\mathrm{X_1}\le\mathrm{x}\right)\mathrm{P}\left(\mathrm{X_2}\le\mathrm{x}\right)=\exp\left(-\mathrm{e}^{\mathrm{b_1}-\mathrm{x}}\right)\cdot\exp\left(-\mathrm{e}^{\mathrm{b_2}-\mathrm{x}}\right)\\&=\exp\left(-\mathrm{e}^{-\mathrm{x}}\left(\mathrm{e}^{\mathrm{b_1}}+\mathrm{e}^{\mathrm{b_2}}\right)\right)=\exp\left(-\mathrm{e}^{\mathrm{b}-\mathrm{x}}\right)\end{aligned}$$

where

$$\mathrm{b}=\log\left(\mathrm{e}^{\mathrm{b_1}}+\mathrm{e}^{\mathrm{b_2}}\right).$$

Similar results hold for the other two types of extreme value distributions.

In the multivariate case where the random variables are vectors, there exists similar asymptotic results for maxima as in the univariate case, where maximum of a vector is defined as maximum taken componentwise. The resulting limiting distributions are called multivariate extreme value distributions, and they are of three types as in the univariate case. A characterization of type III is given in Theorem 8 in Section 3.10. More details about the multivariate extreme value distributions can be found in Resnick (1987).

A general type III extreme value distribution has the form

$$\exp\left(-e^{-(x-b)/a}\right)$$

and it has the mean $b + 0.5772\ldots.$, and variance equal to $a^2\pi^2/6$, cf. Lemma A1 below.

Lemma A1

Let ε be standard type III extreme value distributed and let $s < 1$. Then

$$E\, e^{s\varepsilon} = \Gamma(1-s)$$

where $\Gamma(\cdot)$ denotes the Gamma function. In particular

$$E\,\varepsilon = -\Gamma'(1) = 0.5772\ldots$$

and

$$Var\,\varepsilon = \Gamma''(1) - \Gamma'(1)^2 = \frac{\pi^2}{6}.$$

Proof:

We have

$$E\, e^{s\varepsilon} = \int_{-\infty}^{\infty} e^{sx} \exp\left(-e^{-x}\right) e^{-x}\, dx.$$

By change of variable $t = e^{-x}$ this expression reduces to

$$E\, e^{s\varepsilon} = \int_{-\infty}^{\infty} t^{-s} e^{-t}\, dt = \Gamma(1-s).$$

Moreover, the formulaes $E\varepsilon = -\Gamma'(1)$ and $E\varepsilon^2 = \Gamma''(1)$ follows immediately. The values of $\Gamma'(1)$ and $\Gamma''(1)$ can be found in any standard tables on the Gamma function.

Q.E.D.

Lemma A2

Suppose $U_j = v_j + \varepsilon_j$, *where* $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m)$ *is multivariate extreme value distributed. Then*

$$P\left(max_k\, U_k \le y \,\middle|\, U_j = max_k\, U_k\right) = P\left(U_j \le y \,\middle|\, U_j = max_k\, U_k\right) = P\left(max_k\, U_k \le y\right).$$

Proof: According to the definition of the multivariate extreme value distribution

(A.8) $$P(U_1 \le y_1, U_2 \le y_2, \ldots, U_m \le y_m) \equiv F(y_1, y_2, \ldots) = \exp\left(-G\left(e^{v_1 - y_1}, e^{v_2 - y_2}, \ldots, e^{v_m - y_m}\right)\right)$$

where $G(\cdot)$ is homogeneous of degree one. For notational simplicity let $j = 1$, since the general case is completely analogous. Let ∂_j denote the partial derivative with respect to component j. We have

(A.9)
$$P\left(\max_k U_k \in (z, z + dz), U_1 = \max_k U_k\right) = P\left(U_1 \in (z, z + dz), U_2 \le z, \ldots, U_m \le z\right) = \partial_1 F(z, z, \ldots, z)\, dz.$$

Since by assumption

(A.10) $$G\left(e^{v_1 - y_1}, e^{v_2 - y_2}, \ldots, e^{v_m - y_m}\right) = e^{-y}\, G\left(e^{v_1 - y_1 + y}, e^{v_2 - y_2 + y}, \ldots, e^{v_m - y_m + y}\right)$$

we get

(A.11) $$\partial_1 F(z, z, \ldots) = \exp\left(-e^{-z}\, G\left(e^{v_1}, e^{v_2}, \ldots, e^{v_m}\right)\right) \partial_1 G\left(e^{v_1}, e^{v_2}, \ldots, e^{v_m}\right) e^{v_1 - z}.$$

Hence

$$
\begin{aligned}
&P\left(\max_k U_k \le y, U_1 = \max_k U_k\right) = \int_{-\infty}^{y} \partial_1 F(z,z,\ldots,z)\,dz \\
&= e^{v_1}\,\partial_1 G\left(e^{v_1}, e^{v_2}, \ldots, e^{v_m}\right) \int_{-\infty}^{y} \exp\left(-e^{-z}\, G\left(e^{v_1}, e^{v_2}, \ldots, e^{v_m}\right)\right) e^{-z}\,dz \\
&= \frac{e^{v_1}\,\partial_1 G\left(e^{v_1}, e^{v_2}, \ldots, e^{v_m}\right)}{G\left(e^{v_1}, e^{v_2}, \ldots, e^{v_m}\right)} \cdot \exp\left(-e^{-y}\, G\left(e^{v_1}, e^{v_2}, \ldots, e^{v_m}\right)\right).
\end{aligned}
\tag{A.12}
$$

With $y = \infty$ in (A.12) we realize that the first factor on the right hand side equals the choice probability, $P\left(U_1 = \max_k U_k\right)$. Hence we have proved Theorem 8 as well. This implies also that the second factor on the right hand side equals $P\left(\max_k U_k \le y\right)$. Moreover, it follows that the events $\left\{U_1 = \max_k U_k\right\}$ and $\left\{\max_k U_k \le y\right\}$ are stochastically independent.

Q.E.D.

Lemma A3

Assume that $Y = \mu + \sigma u$, *where*

$$P(u \le y) = \frac{1}{1 + exp(-y)}.$$

Then

$$P\left(u > y \,|\, Y > 0\right) = \frac{1 + exp\left(-\frac{\mu}{\sigma}\right)}{1 + exp\,(y)} \tag{A.13}$$

for $y > -\frac{\mu}{\sigma}$, *and equal to one for* $y \le \frac{\mu}{\sigma}$. *Furthermore,*

$$E\left(u \,|\, Y > 0\right) = \left(1 + exp\left(-\frac{\mu}{\sigma}\right)\right) log\left(1 + exp\left(\frac{\mu}{\sigma}\right)\right) - \frac{\mu}{\sigma} = -\frac{log\, P\left(Y < 0\right)}{P\left(Y > 0\right)} - \frac{\mu}{\sigma}. \tag{A.14}$$

Proof:

For $y > -\frac{\mu}{\sigma}$ we have

$$P(u > y \mid Y > 0) = \frac{P\left(u > y, u > -\frac{\mu}{\sigma}\right)}{P\left(u > -\frac{\mu}{\sigma}\right)}$$

(A.15)

$$= \frac{P(-u < -y)}{P\left(-u < \frac{\mu}{\sigma}\right)} = \frac{P(u < -y)}{P\left(u < \frac{\mu}{\sigma}\right)} = \frac{1 + \exp\left(-\frac{\mu}{\sigma}\right)}{1 + \exp(y)}$$

which proves (A.13).

Consider next (A.14). Let $\tilde{Y} = Y/\sigma$. Then for $y \geq 0$

(A.16) $$P\left(\tilde{Y} > y \,\middle|\, \tilde{Y} > 0\right) = \frac{P\left(\tilde{Y} > y, \tilde{Y} > 0\right)}{P\left(\tilde{Y} > 0\right)} = \frac{P\left(\tilde{Y} > y\right)}{P\left(\tilde{Y} > 0\right)} = \frac{1 + \exp\left(-\frac{\mu}{\sigma}\right)}{1 + \exp\left(y - \frac{\mu}{\sigma}\right)}.$$

Hence

$$E\left(\tilde{Y} \,\middle|\, \tilde{Y} > 0\right) = \int_0^\infty P\left(\tilde{Y} > y \,\middle|\, \tilde{Y} > 0\right) dy = \left(1 + \exp\left(-\frac{\mu}{\sigma}\right)\right) \int_0^\infty \frac{dy}{1 + \exp\left(y - \frac{\mu}{\sigma}\right)}$$

(A.17)
$$= \left(1 + \exp\left(-\frac{\mu}{\sigma}\right)\right) \int_0^\infty \frac{\exp\left(\frac{\mu}{\sigma} - y\right) dy}{1 + \exp\left(\frac{\mu}{\sigma} - y\right)} = \left(1 + \exp\left(-\frac{\mu}{\sigma}\right)\right) \Big|_0^\infty \left(-\log\left(1 + \exp\left(\frac{\mu}{\sigma} - y\right)\right)\right)$$

$$= \left(1 + \exp\left(-\frac{\mu}{\sigma}\right)\right) \log\left(1 + \exp\left(\frac{\mu}{\sigma}\right)\right).$$

This implies that

$$E(u \mid Y > 0) = E\left(\tilde{Y} \,\middle|\, \tilde{Y} > 0\right) - \frac{\mu}{\sigma} = \left(1 + \exp\left(-\frac{\mu}{\sigma}\right)\right) \log\left(1 + \exp\left(\frac{\mu}{\sigma}\right)\right) - \frac{\mu}{\sigma}$$

and (A.14) has thus been proved.

Q.E.D.

References and selected readings

Amemiya, T. (1981): Qualitative Response Models: A Survey. *Journal of Economic Literature*, **19**, 1483-1536.

Amemiya, T. (1985): *Advanced Econometrics.* Basil Blackwell Ltd. Oxford, UK.

Anderson, S.P., A. de Palma and J.-F. Thisse (1992): *Discrete Choice Theory of Product Differentiation*. MIT Press, Cambridge, Massachusetts.

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

Berkson, J. (1953): A Statistically Precise and Relatively Simple Method of Estimating the Bio-Assay with Quantal Response, Based on the Logistic Function. *Journal of the American Statistical Association*, **48**, 529-549.

Block, H.D., and J. Marschak (1960): Random Orderings and Stochastic Theories of Response. In I. Olkin (ed.): *Contributions to Probability and Statistics.* Stanford University Press, Stanford.

Cameron, A. C., and P. K. Trivedi (2005): *Microeconometrics: Methods and Applications.* Cambridge University Press, New York.

Dagsvik, J.K. (1985): Kvalitativ valghandlingsteori, en oversikt over feltet.(Qualitative Choice Theory, a survey.) *Sosialøkonomen*, no. 2, 32-38.

Dagsvik, J.K. (1994): Discrete and Continuous Choice, Max-Stable Processes and Independence from Irrelevant Attributes. *Econometrica*, **62**, 1179-1205.

Dagsvik, J.K. (1995): How Large is the Class of Generalized Extreme Value Random Utility Models? *Journal of Mathematical Psychology*, **39**, 90-98.

Dagsvik, J. K. (2001): James Heckman og Daniel McFadden: To pionerer i utviklingen av mikroøkonometri. (James Heckman and Daniel McFadden: Two pioneers in the development of micro-econometrics.) *Økonomisk forum*, **55**, 31-38.

Dagsvik, J.K. (2004): Hvordan skal arbeidstilbudseffekter tallfestes? En oversikt over den mikrobaserte arbeidstilbudsforskningen i Statistisk sentralbyrå. *Norsk Økonomisk Tidskrift*, **118**, 22-53.

Dagsvik, J.K., D.G. Wetterwald and R. Aaberge (2002): Potential Demand for Alternative Fuel Vehicles. *Transportation Research Part B*, **36**, 361-384.

Debreu, G. (1960): Review of R.D. Luce, Individual Choice Behavior: A Theoretical Analysis. *American Economic Review*, **50**, 186-188.

Dubin, J., and D. McFadden (1984): An Econometric Analysis of Residential Electric Appliance Holdings and Consumption. *Econometrica*, **52**, 345-362.

Georgescu-Roegen, N. (1958): Threshold in Choice and the Theory Demand. *Econometrica*, **26**, 157-168.

Greene, W.H. (1993): *Econometric Analysis*. Prentice Hall, Englewood Cliffs, New Jersey.

Hanemann, W.M. (1984): Discrete/Continuous Choice of Consumer Demand. *Econometrica*, **52**, 541-561.

Hausman, J., and D.A. Wise (1978): A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences. *Econometrica*, **46**, 403-426.

Heckman, J.J. (1979): Sample Selection Bias as a Specification Error. *Econometrica*, **47**, 153-161.

Lamperti, J.W. (1996): *Probability*. J. Wiley & Sons, Inc., New York.

Lattin, J., J. D. Carroll and P. E. Green ((2003): Analyzing Multivariate Data. Brooks & Cole
Luce, R.D. (1959): *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York.

Luce, R.D., and P. Suppes (1965): Preference, Utility and Subjective Probability. In R.D. Luce, R.R. Bush, and E. Galanter (eds.): *Handbook of Mathematical Psychology*, III. Wiley, New York.

Maddala, G.S. (1983): *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge University Press, New York.

Manski, C.F. (1977): The Structure of Random Utility Models. *Theory and Decision*, **8**, 229-254.

McFadden, D. (1973): Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press, New York.

McFadden, D. (1978): Modelling the Choice of Residential Location. In A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull (eds.): *Spatial Interaction Theory and Planning Models*. North Holland, Amsterdam.

McFadden, D. (1981): Econometric Models of Probabilistic Choice. In C.F. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*. MIT Press, Cambridge, Massachusetts.

McFadden, D. (1984): Econometric Analysis of Qualitative Response Models. In Z. Griliches and M.D. Intriligator (eds.): *Handbook of Econometrics*, Vol. II, Elsevier Science Publishers BV, New York.

McFadden, D. (1989): A Method of Simulated Moments of Discrete Response Models without Numerical Integration. *Econometrica*, **57**, 995-1026.

McFadden, D., and K. Train (2000): Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics*, **15**, 447-470.

Quandt, R.E. (1956): A Probabilistic Theory of Consumer Behavior. *Quarterly Journal of Economics*, **70**, 507-536.

Resnick, S.I. (1987): *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag, New York.

Strauss, D. (1979): Some Results on Random Utility Models. *Journal of Mathematical Psychology*, **20**, 35-52.

Thurstone, L.L. (1927): A Law of Comparative Judgment. *Psychological Review,* **34**, 273-286.

Tobin, J. (1958): Estimation of Relationships for Limited Dependent Variables. *Econometrica*, **26**, 24-36.

Train, K. (1986): *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand*. MIT Press, Cambridge, Massachusetts.

Train, K. (2003): *Discrete Choice Methods with Simulations*. Cambridge University Press, New York.

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

Yellott, J.I. (1977): The Relationship between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgment, and the Double Exponential Distribution. *Journal of Mathematical Psychology*, **15**, 109-144.