Statistics Norway
Research Department

John K. Dagsvik

Probabilistic Models for Qualitative Choice Behavior: An Introduction

Statistics Norway
Statistisk sentralbyrå

Documents 98/15 • Statistics Norway, August 1998

John K. Dagsvik

Probabilistic Models for Qualitative Choice Behavior: An Introduction

Preface:
The econometric disipline has been criticized for being too similar to mathematical statistics and only to a limited degree linked to formalized theoretical models. This is particularly the case as regards formulation and specification of the stochastic elements in econometric models. Ragnar Frisch, who is known to be the originator of econometrics, expressed both in theory and practice an opposite ideal; namely econometrics as an almost symbiotic blend of statistical methodology and mathematically formulated theory, cf. Frisch (1926). See also Bjerkholt (1995).

Theory and econometric methodology for qualitative choice behavior is developed in a tradition which I believe is somewhat closer to the ideal of Frisch than much of the traditional textbook approach to econometrics. This stems from the fact that the theory of qualitative choice is rooted in a tradition where probabilistic concepts and formulations play a key role in contrast to the point of departure in traditional micro theory, which is deterministic. Since probabilistic concepts are integral parts of the theory of qualitative choice this means that the gap between theory and empirical model specification in applications often becomes less wide than is the case in the traditional micro-economic approach.

The present compendium is a revised version of an introductory course in the theory of qualitative choice behavior (often called the theory of discrete choice). Some of the material I present here draws on a Ph.D. course I gave at the Department of Economics, University of Wisconsin, during the Fall semester of 1990.

Acknowledgement: I acknowledge the helpful comments of Steinar Strøm and Anne Skoglund for word processing assistance.

Address: John K. Dagsvik, Statistics Norway, Research Department, P.O.Box 8131 Dep., N-0033 Oslo, Norway. E-mail: 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, may 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 startet 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.

In contrast to standard textbooks and surveys in econometric modelling 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.[1]

This survey is organized as follows: In chapter 2 I give a brief overview of reduced form type specifications and estimation of models with discrete or limited dependent response. In chapter 3 I discuss some important elements of probabilistic choice theory, and in chapter 4 the issue is functional forms and econometric specification of discrete choice models. In chapter 5 I discuss the modeling of a few selected applications of discrete choice analysis, and in chapter 6 the extension to discrete/continuous choice model is treated. In the final chapter I discuss applications on discrete/continuous modeling.

[1] An elementary survey in Norwegian is Dagsvik (1985).

2. Statistical analysis when the dependent variable is discrete

As mentioned in the introduction there are many interesting phenomena which naturally can be modelled with a dependent variable being qualitative (discrete) or where the dependent variables 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 and tools which are useful for analyzing such phenomena, 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 with discrete response

When analyzing "demand for housing", "tourist destinations", "type of accident", etc. the response—or dependent variable—is typically discrete and it often has the structure of a binomial, or more generally, a multinomial variable. Recall that in multinomial experiments with m possible categories only one out of m outcomes can occur in each experiment. In other words, the outcomes are mutually exclusive. For example, out of m possible housing alternatives the household will only select one. Similarly, a student who has the choice between m different schools will only select one. Statistically, a multinomial model is represented by probabilities, P_j, $j = 1,2,\ldots,m$, where P_j is the probability that outcome j shall occur.

Let Y_j denote the corresponding response variable, where $Y_j = 1$ if outcome j occurs and zero otherwise. (For simplicity, we suppress the indexation of the agent.) Then $E\,Y_j = P(Y_j = 1)\cdot 1 + P(Y_j = 0)\cdot 0 = P(Y_j = 1) \equiv P_j$. We can therefore write

$$Y_j = P_j + e_j \tag{2.1}$$

where $\{e_j\}$ are random terms with zero mean. Thus, once the systematic term P_j has been specified as a function of explanatory variables, one could estimate the unknown parameters by regression analysis. However, it is problematic to specify the probabilities $\{P_j\}$ as linear functions of the explanatory variables due to the fact that a linear specification does not necessarily satisfy the constraints that $0 \le P_j \le 1$, and $\sum_j P_j = 1$ (cf. Maddala, 1983, pp. 15-16, or Greene, 1990, pp. 636-441).

Example 2.1

Consider the modelling of labor force participation. In this case $m=2$, where alternative one represents participation, while alternative two represents nonparticipation. It is believed that a number of factors, such as age, marital status, number of small children, education, etc., explain the outcome. Let X be the vector of relevant (observable) variables that explain the outcome. Thus

$$P_2 = \psi(X\beta) \tag{2.2}$$

where $\psi(\cdot)$ is a suitable chosen functional form while β is a vector of unknown parameters. If one could estimate β it would for example be possible to assess the marginal effect of education on the labor force participation. We realize that $\psi(\cdot)$ must be positive and $0 \le \psi(\cdot) \le 1$.

2.1.1. The multinomial Logit model

One convenient and commonly used specification that fulfills the above restrictions is the multinomial logit model. One version of the multinomial logit model has the structure

$$P_j = H_j(X) \equiv \frac{\exp(X\beta_j)}{\sum_{k=1}^{m} \exp(X\beta_k)} \tag{2.3}$$

where X is, typically, a vector of agent-specific variables and β_j, $j=1,2,\ldots,m$, are vectors of unknown parameters. This specification is also convenient for estimation purposes as we shall see below.

From (2.3) it follows that

$$\log\left(\frac{H_j(X)}{H_1(X)}\right) = X(\beta_j - \beta_1). \tag{2.4}$$

Eq. (2.4) demonstrates that at most $\beta_j - \beta_1$ can be identified. To realize this, suppose β_j^*, $j=1,2,\ldots,m$, are parameter vectors such that $\beta_j^* \neq \beta_j$. If

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

for $j=2,\ldots,m$, then $\{\beta_j^*\}$ will satisfy (2.4), and consequently $\{\beta_j\}$ are not identified. We can therefore, without loss of generality, put $\beta_1 = 0$, and write

$$H_1(X) = \frac{1}{1+\sum_{k=2}^{m} \exp(X\beta_k)} \tag{2.5a}$$

and

$$H_j(X) = \frac{\exp(X\beta_j)}{1+\sum_{k=2}^{m} \exp(X\beta_k)} \tag{2.5b}$$

for $j = 2,3,\ldots,m$.

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 now be modelled as in (2.5).

2.1.2. The binary Probit model

The binary probit model is often motivated by a latent variable specification such as in (2.7), but with u normally distributed instead of logistically distributed. If $\Phi(\cdot)$ denotes the cumulative normal distribution, N(0,1), then the probit model follows by replacing L(y) by $\Phi(y)$ in which case we obtain the binary Probit model as

$$P(Y_2 = 1) = \Phi(X\beta) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X\beta} \exp\left(-\frac{t^2}{2}\right) dt. \tag{2.6}$$

The normal and the logistic distributions are rather close, and in most applications one has found that the binary logit and probit models are almost indistinguishable.

In case there are extreme values of the explanatory variables the predictions from the logit and probit model conditional on these extreme values may, however, differ since the logistic distribution has slightly heavier tails than the normal distribution.

2.1.3. Binary models derived from latent variable specifications

For the sake of motivation let us reconsider Example 2.1. Let now U_j be the individual's utility of alternative j, $j = 1,2$, and let

$$U_j = X\beta_j + u_j \tag{2.7}$$

where u_j is a random variable that is supposed to capture unobserved variables that affect the utility of alternative j. Let

$$Y^* \equiv U_2 - U_1 = X\beta - u \tag{2.8}$$

where $\beta = \beta_1 - \beta_2$ and $u = u_1 - u_2$. Let $\psi(y) \equiv P(u \leq y)$, be the cumulative distribution function of u, which we assume is independent of X. Consistent with the notation in Example 2.1, let the observable variable, Y_2, be given by

$$Y_2 = \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

and $Y_1 = 1 - Y_2$. From (2.8) it follows that the probability of participation equals

$$\begin{aligned} P_2 &= P(Y_2 = 1) = P(Y^* > 0) \\ &= P(X\beta - u > 0) = P(X\beta > u) = \psi(X\beta). \end{aligned}$$

If $\psi(y) = \Phi(y)$, where $\Phi(\cdot)$ is given by (2,6), the *Probit* model follows, whereas if

$$\psi(y) = \frac{1}{1 + \exp(-y)} \tag{2.9}$$

the binary *Logit* model follows. The distribution function (2.9) is known as the *Logistic* distribution.

For example, in the labor force participation example, Y^* may be interpreted as the difference between the agent's (expected) market wage and the reservation wage. This, and further examples will be discussed in chapter 5.

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 were 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, $i = 1,2,\ldots,m$. Thurstone assumed that

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

where ε_j^i, $j = 1,2$, $i = 1,2,\ldots,m$, are independent and normally distributed with zero mean and standard deviation equal to σ. 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

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

$j = 1,2$. 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 $\Phi(y)$ 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_j. 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 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).

3.1.3. General systems of choice probabilities

Formally, we shall decribe 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) A set of agent-specific characteristics.

(iii) A random utility function U_j, where U_j is the agent's utility of alternative j, $j \in S$, and a distribution function M which yields the joint distribution of the utilities in S, i.e.,

$$M(u_1, u_2, \ldots) = P(U_1 \leq u_1, U_2 \leq u_2, \ldots). \tag{3.2}$$

From the assumptions above it is possible in principle to derive the system of choice probabilities, $\{P_j(B)\}$, where $P_j(B)$ is defined by

$$P_j(B) = P\left(U_j = \max_{k \in B} U_k\right) \tag{3.3}$$

and $j \in B \subset S$. The interpretation of (3.3) is as the probability that the agent will choose alternative j when the set of feasible alternatives are equal to B. It is important to stress that a choice probability is a function of *two* arguments, namely j and B. For each given B, $P_j(B)$, $j \in B$, are multinomial probabilities. The relationship between $P_j(B)$ and $P_j(A)$ for two different choice sets A and B is governed by the joint distribution of the utilities. 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,

$$U_j = v_j + \varepsilon_j, \tag{3.4}$$

where v_j is a deterministic term and ε_j is a random variable with joint distribution of the terms $\{\varepsilon_j\}$ 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:* We may 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. Furthermore, 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. 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 are $\{P_j(B)\}$ and $\{P_j(A)\}$ 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 agents. 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

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

Definition 2; Independence from irrelevant alternatives (IIA)

A system of choice probabilities, $\{P_j(B)\}$, with $P_j(B) \neq 0,1$, satisfies IIA if and only if for all j, A, B such that $j \in A \subset B \subset S$,

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

Eq. (3.5) states that the probability of choosing alternative j from B equals the probability of selecting a subset A of the "best" alternatives 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 is 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".

It may be instructive for the sake of clarification of the IIA property to consider the relationship between $P_j(B)$ and the conditional choice probability given that the chosen alternative belongs to B. More specifically, suppose for example that the universal set S is feasible. Then the conditional choice probability that alternative j is chosen, given that the chosen alternative belongs to $B \subset S$, equals

$$\frac{P_j(S)}{P_B(S)},$$

which only coincides with $P_j(B)$ when IIA holds. While $P_j(B)$ expresses the probability that j is chosen when the choice set equals B, $P_j(S)/P_B(S)$ expresses the probability that j is chosen when the choice set is S, given that the chosen outcome belongs to B. The empirical counterpart to $P_j(S)/P_B(S)$ is the number of agents that face choice set S and have chosen j, to the number of agents that face choice set S and whose choice outcomes belong to B.

Definition 3; 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 $j, k \in B \subset S$,

(3.6) $$P_j(\{k, j\})/P_k(\{k, j\}) = P_j(B)/P_k(B)$$

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. Then the IIA assumption holds if and only if there exist positive scalars, a(j), $j \in S$, such that the choice probabilities equal

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

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

Proof: Assume first that (3.7) holds. Then it follows immediately that (3.5) holds. Assume next that (3.5) holds. Define $a(j)=c\,P_j(S)$, where c is an arbitrary positive constant. Then by (3.5) 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 $B\subset S$. This shows that $P_j(B)$ has the structure (3.7).

To show uniqueness (apart from multiplication by a constant), let $\tilde{a}(j)$ be positive scalars such that (3.7) 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 (3.7) 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 must 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 $a_1=1$). 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 holds it follows that for distinct i, j and $k \in S$

(3.8)
$$P_i(\{i,j\})P_j(\{j,k\})P_k(\{k,i\})=P_i(\{i,k\})P_k(\{k,j\})P_j(\{j,i\}).$$

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 (3.8) is that

$$P(i\succ j\succ k\succ i)=P(i\succ k\succ j\succ i)$$

where $\succ$ means "preferred to". In other words, the intransitive chains $i\succ j\succ k\succ i$ and $i\succ k\succ j\succ i$ 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 (3.8) holds then the binary choice probabilities must have the form

$$P_j(\{i,j\}) = \frac{a(j)}{a(i)+a(j)} \tag{3.9}$$

where $\{a(j), j \in S\}$ are unique up to multiplication by a constant, cf. Luce and Suppes (1965). However, (3.8) does not imply IIA. Equation (3.8) 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, $U_j = \log a(j) + \varepsilon_j$, *where* ε_j, $j \in S$, *are i.i.d. according to the standard type III extreme value distribution*[2]

$$P(\varepsilon_j \le x) = exp(-e^{-x}). \tag{3.10}$$

Then, for $j \in B \subset S$,

$$P_j(B) \equiv P\left(U_j = \max_{k \in B} U_k\right) = \frac{a(j)}{\sum_{k \in B} a(k)}. \tag{3.11}$$

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

$$P(V_j \le y) = \prod_{k \in B \setminus \{j\}} P(\varepsilon_k \le y - \log a(k)) = \prod_{k \in B \setminus \{j\}} \exp(-a(k)e^{-y}) = \exp(-e^{-y} D_j) \tag{3.12}$$

where

$$D_j = \sum_{k \in B \setminus \{j\}} a(k). \tag{3.13}$$

---

[2] In the following the distribution function (3.10) will be called the standard extreme value distribution.

Hence

$$(3.14)\quad P\left(U_j = \max_{k\in B} U_k\right) = P\left(U_j > V_j\right) = P\left(\varepsilon_j + \log a(j) > V_j\right) = \int_{-\infty}^{\infty} P\left(y > V_j\right) P\left(U_j \in (y, y+dy)\right).$$

Note next that since by (3.10)

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

it follows that

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

Hence

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

Since

$$D_j + a(j) = \sum_{k\in B} a(k)$$

the result of the Theorem follows from (3.14) and (3.15)

Q.E.D.

An interesting question is whether or not there exists other distribution functions than (3.10) 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 $U_j = \log a(j) + \varepsilon_j$, *where* ε_j, $j \in S$, *are i.i.d. with cumulative distribution function that is strictly increasing on the real line. Then (3.11) 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

(3.16) $$U_j = Z_j\beta + \varepsilon_j\sigma$$

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

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

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

(3.18) $$P_j = \frac{\exp\left(Z_j\tilde{\beta}\right)}{\sum_{k=1}^{m}\exp\left(Z_k\tilde{\beta}\right)}.$$

Clearly, $\tilde{\beta}$ is identified, since

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

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 (3.17) corresponds to a small $\tilde{\beta}$.

When $\tilde{\beta}$ has been estimated one can compute the aggregate own- and cross-price elasticities according to the formulae

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

and

(3.20) $$\frac{\partial \log P_j}{\partial \log Z_k} = -\tilde{\beta}\,Z_k\,P_k$$

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

$$Z_{i1j} = \begin{cases} 1 & \text{if } j=1 \\ 0 & \text{otherwise,} \end{cases}$$

Z_{i2j} = In-vehicle time, alternative j,

Z_{i3j} = Out-of-vehicle time, alternative j,

Z_{i4j} = Transportation cost, alternative j,

$$Z_{i51} = \begin{cases} 1 & \text{if the household own a car} \\ 0 & \text{otherwise,} \end{cases}$$

and

$Z_{i52} = 0.$

The utility function is assumed to have the structure

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

where $Z_{ij} = (Z_{i1j}, Z_{i2j}, Z_{i3j}, Z_{i4j}, Z_{i5j})$, 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

(3.21)
$$P_{ij} = \frac{\exp(Z_{ij}\beta)}{\sum_{k=1}^{2} \exp(Z_{ik}\beta)}.$$

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 equals

(3.22)
$$P_j = \frac{\exp(Z_j\beta)}{\sum_{k=1}^{2} \exp(Z_k\beta)}$$

for $j=1,2$. From (3.22) it follows that

(3.23)
$$\frac{\partial \log P_j}{\partial \log Z_{jr}} = \beta_r Z_{jr}(1 - P_j)$$

and

$$\frac{\partial \log P_j}{\partial \log Z_{kr}} = -\beta_r Z_{kr} P_k \tag{3.24}$$

for $k \neq j$. Eq. (3.23) expresses the "own elasticities" while (3.24) expresses the "cross elasticities". Specifically, (3.23) 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.

3.4. The independent random utility model

If ε_j, $j \in S$, are independent then the choice probabilities can be expressed as

$$P_j(B) = \int \prod_{k \in B \setminus \{j\}} F_k(y - v_k) F_j'(y - v_j)\, dy \tag{3.25}$$

where $F_j(y) = P(\varepsilon_j \leq y)$, and $B \subset S$.

To realize that (4.9) hold note that since ε_j, $j \in S$, are independent we get

$$P\left(\max_{k \in B \setminus \{j\}} U_k \leq y\right) = P\left(\bigcap_{k \in B \setminus \{j\}} (\varepsilon_k \leq y - v_k)\right) = \prod_{k \in B \setminus \{j\}} P(\varepsilon_k \leq y - v_k) = \prod_{k \in B \setminus \{j\}} F_k(y - v_k).$$

Furthermore,

$$P(U_j \in (y, y + dy)) = F_j'(y)\, dy.$$

Hence,

$$P_j(B) = P\left(U_j > \max_{k \in B \setminus \{j\}} U_k\right) = \int_{-\infty}^{\infty} P\left(y > \max_{k \in B \setminus \{j\}} U_k\right) F_j'(y)\, dy = \int_{-\infty}^{\infty} \prod_{k \in B \setminus \{j\}} F_k(y - v_k) F_j'(y)\, dy.$$

Example 3.3. (Multinomial logit)

Assume that

$$F_j(y) = e^{-e^{-y}}. \tag{3.26}$$

Then (3.25) yields

$$P_j(B) = \frac{e^{v_j}}{\sum_{k \in B} e^{v_k}}. \tag{3.27}$$

Example 3.4. (Independent multinomial probit)

If

$$F_j'(y) = \Phi'(y) \equiv \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \tag{3.28}$$

then

$$P_j(B) = \int_{-\infty}^{\infty} \prod_{k \in B \setminus \{j\}} \Phi(y - v_k) \exp\left(-\frac{1}{2}(y - v_j)^2\right) dy. \tag{3.29}$$

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

$$P(U_1 > U_2) = \Phi(v_1 - v_2). \tag{3.30}$$

Example 3.6. (Binary Arcus-tangens)

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

$$F_j'(y) = \frac{2}{\pi(1 + 4y^2)}. \tag{3.31}$$

The density (3.31) is the density of the standard Cauchy distribution. Then

$$P(U_1 > U_2) = \frac{1}{2} + \frac{1}{\pi} \operatorname{Arctg}(v_1 - v_2). \tag{3.32}$$

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.

3.5. 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

$$v_j = Z_j \beta \equiv \sum_{k=1}^{K} Z_{jk} \beta_k. \tag{3.33}$$

A more general specification, which was already mentioned in chapter 2, is

$$v_j = \sum_{k=1}^{K} R_k(Z_j, X)\beta_k \tag{3.34}$$

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

Example 3.7

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

$$v_j = Z_{j1}\beta_1 + Z_{j2}\beta_2 + Z_{j1}X_1\beta_3 + Z_{j1}X_2\beta_4 + Z_{j2}X_1\beta_5 + Z_{j2}X_2\beta_6 . \tag{3.35}$$

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

$$v_j = \left(\frac{Z_{j1}^{\alpha_1}-1}{\alpha_1}\right)\beta_1 + \left(\frac{Z_{j2}^{\alpha_2}-1}{\alpha_2}\right)\beta_2 \tag{3.36}$$

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

$$\frac{y^{\alpha}-1}{\alpha}, \tag{3.37}$$

$y>0,$ 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 $\alpha<1$, $(y^{\alpha}-1)/\alpha$ is concave while it is convex when $\alpha>1$. For any α, $(y^{\alpha}-1)/\alpha$ is increasing in y.

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 again

Example 3.7, and suppose that Z_{j2} and X_2 are not observed. Suppose that we try to deal with the missing variable problem by applying $Z_{j1}\mu_1$ as a proxy for $Z_{j1}X_2\beta_4$, $X_1\mu_{j2}$ as a proxy for $Z_{j2}X_1\beta_5$ and μ_{j3} as a "proxy" for $Z_{j2}X_2\beta_5 + Z_{j2}\beta_2$, where μ_1, μ_{j2} and μ_{j3} are unknown parameters. This corresponds to a utility function with error term

$$\varepsilon_j^* = \varepsilon_j + Z_{j1}X_2\beta_4 - Z_{j1}\mu_1 + Z_{j2}X_2\beta_5 + Z_{j2}\beta_2 - \mu_{j2}. \tag{3.38}$$

Now if X_1 and X_2 are correlated we realize that ε_j^* will be correlated with the deterministic term

$$v_j^* = Z_{j1}(\beta_1 + \mu_1) + Z_{j1}X_1\beta_3 + X_1\mu_{j2} + \mu_{j3}. \tag{3.39}$$

This simple example shows that simultaneous equation bias may be a serious problem in many cases where data contains limited information about population heterogeneity. 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 detterministic term of the utility function. This is easily demonstrated by a similar argument as the one above.

3.6. Aggregation of latent alternatives

In this section we shall obtain a characterization of the choice model that may be justified in applications that conform to the following general description: For the sake of expository convenience we proceed by means of a concrete example.

Consider migration choice: The agent faces a set B of feasible regions. Within region j there is a set B_j of feasible schooling and/or employment opportunities. The agent's problem is to choose his favorite opportunity. The researcher only observes the choice of region but not the choice within the chosen region. The agent is assumed to have the utility function with structure

$$U_{jr} = v_j + \varepsilon_{jr} \tag{3.40}$$

where $j = 1,2,\ldots,m$, indexes the regions and $r \in B_j$ indexes the opportunities within B_j. The term v_j is deterministic and represents the systematic mean utility across all opportunities within B_j, while ε_{jr}, $r \in B$, $j = 1,2,\ldots,m$, are i.i.d. with cumulative distribution function F. Let n_j be the number of opportunities in B_j. Evidently the (indirect) utility of choosing region j equals

$$U_j \equiv \max_{r \in B_j} U_{jr} = v_j + \tilde{\varepsilon}_j$$

where

$$\tilde{\varepsilon}_j \equiv \max_{r \in B_j} \varepsilon_j = \max_{r \le n_j} \varepsilon_{jr}.$$

Suppose next that F satisfies Condition (A.6) in Appendix A. Then Theorem A3 implies, provided n_j is large, that

$$P\left(\max_{r\le n_j}\varepsilon_{jr}-\log(c\,n_j)\le x\right)\equiv\exp\left(-e^{-x}\right)$$

which means that

$$v_j+\tilde{\varepsilon}_j\equiv v_j+\log n_j+\log c+\varepsilon_j \tag{3.41}$$

where ε_j, $j=1,2,\ldots,m$, are standard type III extreme value distributed. Thus we obtain fromTheorem 3 that the probability of moving to region j equals

$$P_j\equiv P\left(U_j=\max_{k\in B}U_k\right)=\frac{\exp\left(v_j+\log c+\log n_j\right)}{\sum_{k\in B}\exp\left(v_k+\log c+\log n_k\right)}$$

$$=\frac{c\,n_j e^{v_j}}{c\sum_{k\in B}n_k e^{v_k}}=\frac{n_j e^{v_j}}{\sum_{k\in B}n_k e^{v_k}}.$$

If variables that characterize the regions are available these can be utilized to model $\{n_j\}$ and $\{v_j\}$.

The crucial point in the development above is that even if we are only interested in the analysis of the choice of region, we can exploit the (theoretical) structure of the choice problem to obtain a characterization of the choice model. Specifically, we have demonstrated that aggregating of a large number of latent alternative in fact yields IIA. Moreover, the set of latent alternatives $\{B_j\}$ are represented in the model by the respective sizes $\{n_j\}$.

3.7. 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 been applied to analyze the potential demand for products that may be introduced in the market.

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 $\rho_B = (\rho_1, \rho_2, \ldots, \rho_m)$ be the rank ordering of the alternatives in B, where m is the number of alternatives in B, and $B \subset S$. This means that ρ_i denotes the element in B that has the i'th rank. Moreover, let $P(\rho_B)$ denote the probability that the agent shall prefer rank ordering ρ_B of B, and, consistent with the notation above, let $P_{\rho_i}(B)$ be the probability that the agent shall rank alternative i on top when B is the set of feasible alternatives. Recall that the empirical counterpart of these probabilities are 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.

Definition 4

The ranking probabilities constitute a random utility model if and only if

$$P(\rho_B) = P\left(U(\rho_1) > U(\rho_2) > \ldots > U(\rho_m)\right)$$

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

Definition 5: Generalized IIA

The ranking probabilities satisfy the Independence from Irrelevant Alternatives (IIA) property if and only if for any $B \subset S$

(3.42)
$$P(\rho_B) = P_{\rho_1}(B)\, P_{\rho_2}\left(B \setminus \{\rho_1\}\right) \ldots P_{\rho_{m-1}}\left(\{\rho_{m-1}, \rho_m\}\right).$$

Definition 5 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 2 is a special case of Definition 5.

Theorem 5

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

(3.43) $$P(\rho_B)=\prod_{i\in B}\frac{a(\rho_i)}{\sum_{k\in B\ \{\rho_0,\ldots,\rho_{i-1}\}}a(\rho_k)},$$

for $B\subset S$, where $\rho_0\equiv\{\varnothing\}$. The scalars, $\{a(j)\}$, are uniquely determined up to multiplication by a positive constant.

Block and Marschak (1960, p. 109) have proved Theorem 5, the first part of which is a generalization of a result in Luce (1959, p. 72), cf. Luce and Suppes (1965). As an example consider the case when $B=\{1,2,3\}$ and $\rho_B=(2,3,1)$. Then (3.17) reduces to

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

The next result shows that (3.17) is consistent with a simple random utility representation.

Theorem 6

Assume a random utility model with $U(j)=\log a(j)+\varepsilon_j$, where ε_j, $j\in S$, are i.i.d. with standard extreme value distribution function. Then

(3.44) $$P(\rho_B)\equiv P\left(U(\rho_1)>U(\rho_2)>\ldots>U(\rho_m)\right)=\prod_{i\in B}\frac{a(\rho_i)}{\sum_{k\in B\ \{\rho_0,\rho_1,\ldots,\rho_{i-1}\}}a(\rho_k)}.$$

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

3.8. 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 multinominal 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. However, the assumption of independent random terms is rather restrictive in some cases, which the following example will demonstrate.

Example 3.9. (A version of the red-bus/blue-bus problem, Debreu, 1960)

Consider a commuter choice problem in which there are two transportation alternatives, namely "car", (1), "bus", (2). The fraction of commuters that go by car and bus is 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 bus service is introduced (alternative 3) that is equal in all attributes to the existing bus service except that its buses have a different color from the original buses. Thus, there are now red and blue buses which constitute two bus transportation alternatives. Since the new bus alternative is essential equivalent to the existing bus service 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 (3.7) 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 commuters will tend to treat the two bus alternatives as a single alternative so that

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

and

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

This example demonstrates that if alternatives are "similar" in some sense, then the Luce model is not likely to be valid.

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

$$U_j = Z_{j1}\beta_1 + X_1 Z_{j1}\beta_2 + X_2 Z_{j2}\beta_3 + \varepsilon_j \tag{3.45}$$

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

(3.46) $$U_j^* \equiv Z_{j1}\beta_1 + X_1 Z_{j1}\beta_2 + \mu_j + \varepsilon_j^*$$

where μ_j has the interpretation

(3.47) $$\mu_j = \beta_3 Z_{j2} EX_2,$$

(3.48) $$\varepsilon_j^* = \varepsilon_j + X_2 Z_{j2}\beta_3 - \beta_3 Z_{j2} EX_2,$$

and where we now treat the unobservable components X_2 and Z_{j2} as random variable. (In (3.47) and (3.48) the mean is taken across the population.) Suppose that ε_j, $j = 1,2,\ldots$, are independent. By (3.48) we get

(3.49) $$\operatorname{cov}\left(\varepsilon_i^*, \varepsilon_j^*\right) = \beta_3^2 Z_{i2} Z_{j2} \operatorname{Var} X_2.$$

Thus, we realize in this case that the error terms $\{\varepsilon_j^*\}$ are correlated.

If X_2 is observable but $\{Z_{j2}\}$ is not, we may in empirical estimation resort to the specification

(3.50) $$U_j^{**} = Z_{j1}\beta_1 + X_1 Z_{j1}\beta_2 + X_1\tilde{\mu}_j + \varepsilon_j$$

where

$$\tilde{\mu}_j = \beta_3 Z_{j2}.$$

In this case we therefore still have independent error terms provided we introduce alternative-specific dummies in the deterministic terms of the utilities.

Finally suppose that $\{Z_{j2}\}$ are observable while X_2 is not. Then a natural specification would be

(3.51) $$\tilde{U}_j = Z_{j1}\beta_1 + X_1 Z_{j2}\beta_2 + Z_{j3}\tilde{\beta}_3 + \tilde{\varepsilon}_j$$

where

(3.52) $$\tilde{\varepsilon}_j = \varepsilon_j + X_2 Z_{j2}\beta_3 - Z_{j2}\tilde{\beta}_3$$

and

(3.53) $$\tilde{\beta}_3 = \beta_3 \, E X_2 .$$

Hence we get

(3.54) $$\operatorname{cov}\left(\tilde{\varepsilon}_i , \tilde{\varepsilon}_j\right) = \beta_3^2 \, Z_i \, Z_j \, \operatorname{Var} X_2$$

which demonstrate that we may get interdependent random terms solely from unobserved population heterogeneity.

3.9. 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 and costly 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 n-dimensional multinormal zero mean vector-variable with covariance matrix Ω. We have

(3.55) $$h(x;\Omega) = (2\pi)^{-n/2} \, |\Omega|^{-1/2} \exp\left(-\frac{1}{2} x' \Omega^{-1} x\right)$$

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

(3.56) $$P\left(v_j + \varepsilon_j = \max_{k \le n}\left(v_k + \varepsilon_k\right)\right) = \int_{-\infty}^{v_j - v_1} \cdots \int_{-\infty}^{v_j - v_j} \cdots \int_{-\infty}^{v_j - v_n} h\left(x_1, \ldots, x_j, \ldots, x_n ; \Omega\right) dx_1 \ldots dx_j \ldots dx_n .$$

From (3.56) we see that an n-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.

3.10. 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 7

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

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

(ii) $\lim_{x_i \to \infty} G(x_1, \ldots, x_i, \ldots, x_n) = \infty, \quad i = 1,2,\ldots,n,$

(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, the joint distribution function

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

is a well defined multivariate (type III) extreme value distribution function. Moreover, if the random terms of the utility function has joint distribution function given by (3.57), then it follows that

(3.58) $$P\left(v_j + \varepsilon_j = \max_{k \le n}\left(v_k + \varepsilon_k\right)\right) = \frac{e^{v_j}\,\partial_j G\left(e^{v_1}, e^{v_2}, \ldots, e^{v_n}\right)}{G\left(e^{v_1}, e^{v_2}, \ldots, e^{v_n}\right)},$$

where ∂_j denotes the partial derivative with respect to component j.

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 (3.58).

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

(3.59) $$G(x_1, \ldots, x_n) = \sum_{k=1}^{n} x_k$$

from which it follows from (3.32) that

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

Example 3.10

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

(3.60) $$G(x_1, x_2, x_3) = x_1 + \left(x_2^{1/\theta} + x_3^{1/\theta}\right)^{\theta}$$

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

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

and

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

From Theorem 7 we obtain that

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

and

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

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

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

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

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

and

(3.66) $$P_j(S)=\frac{\exp\left(\max(v_2,v_3)\right)I\left(v_j=\max(v_2,v_3)\right)}{e^{v_1}+\exp\left(\max(v_2,v_3)\right)},$$

for $j=2,3$, where I(A) is the indicator function that is equal to one if A is true and zero otherwise, provided $v_2\neq v_3$. For $v_2=v_3$ we obtain

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

and

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

for $j = 2,3$.

Consider the red-bus/blue-bus problem on page 31, where $v_2 = v_3$, which by (3.38), (3.43) and (3.68) yield

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

and

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

Thus the model generated from (3.34) with θ close to zero is able to capture the underlying structure of the red-bus/blue-bus problem.

3.10.1. 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 subsets that 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 red-bus/blue-bus problem has such a tree structure: Here the first stage concern the choice between car and bus while the second stage alternatives are "red-bus" and "blue-bus" in case the first stage choice was bus.

Example 3.11

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 locations within city j. Let $U_{jk} = v_{jk} + \varepsilon_{jk}$, where

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

and

$$G(y_{11}, y_{12}, \ldots, y_{21}, \ldots) = \sum_{j=1}^{2} \left(\sum_{k \in L_j} y_{jk}^{1/\theta_j} \right)^{\theta_j}. \tag{3.70}$$

The structure (3.70) implies that

$$\operatorname{corr}(\varepsilon_{jk}, \varepsilon_{jr}) = 1 - \theta_j^2, \text{ for } r \neq k, \tag{3.71}$$

and

$$\operatorname{corr}(\varepsilon_{js}, \varepsilon_{ir}) = 0 \text{ for } j \neq i. \tag{3.72}$$

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 7 we get that

$$P_{jr} \equiv P\left(U_{jr} = \max_{i=1,2} \left(\max_{r \in L_i} U_{ir} \right) \right) = \frac{e^{v_{jk}} \partial_{jk} G\left(e^{v_{11}}, e^{v_{12}}, \ldots\right)}{G\left(e^{v_{11}}, e^{v_{12}}\right)}$$

$$= \frac{\left(\sum_{k \in L_j} e^{v_{jk}/\theta_j} \right)^{\theta_j - 1} e^{v_{jr}/\theta_j}}{\sum_{i=1}^{2} \left(\sum_{k \in L_i} e^{v_{jk}/\theta_i} \right)^{\theta_i}} \tag{3.73}$$

where $\partial_{jk}G$ is the partial derivative of G with respect to component (j,k). Note that we can rewrite (3.73) as

$$P_{jr} = \frac{\left(\sum_{k \in L_j} e^{v_{jk}/\theta_j} \right)^{\theta_j}}{\sum_{i=1}^{2} \left(\sum_{k \in L_i} e^{v_{ik}/\theta_i} \right)^{\theta_i}} \cdot \frac{e^{v_{jr}/\theta_j}}{\sum_{k \in L_j} e^{v_{jk}/\theta_j}} = P_j \cdot \frac{e^{v_{jr}/\theta_j}}{\sum_{k \in L_j} e^{v_{jk}/\theta_j}}, \tag{3.74}$$

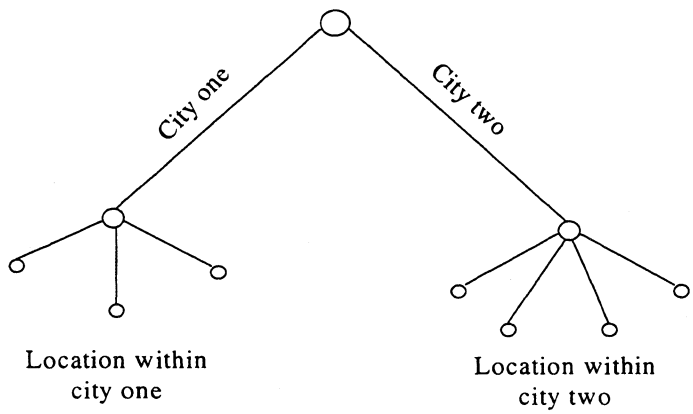
where

$$P_j = \sum_{k \in L_j} P_{jk}. \tag{3.75}$$

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

$$\frac{P_{jr}}{P_j} = \frac{e^{v_{jr}/\theta_j}}{\sum_{k \in L_j} e^{v_{jk}/\theta_j}} \tag{3.76}$$

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 a structure as if L_j were the choice set. We realize that it is therefore consistent with the Luce model. However, only when $\theta = 1$ are the probabilities P_1 and P_2 consistent with the Luce model. Graphically, the above tree structure looks as follows:



So far no theoretical motivation for the GEV model has been given, apart from the property that it contains the Luce model as a special case. I shall therefore conclude this section by reviewing two invariance properties that characterize the GEV class, and discuss their implications.

Definition 6; The DIM property[3]

The utilities $\{U_j\}$ *satisfy DIM if and only if the distribution of* $\max_j U_j$ *is independent of which variable attains the maximum.*

[3] DIM is an acronym for; Distribution in Invariant of which variable attains the Maximum.

Definition 7; The MSD property[4]

The utilities $\{U_j\}$ *satisfy MSD if and only if the distribution of* $\max_j U_j$ *is the same (apart from a location shift) as the distribution of* U_l.

If the utilities satisfy DIM it means that the indirect utility is not correlated with the utility of the *chosen* alternative.

This property corresponds to the notion that the indirect utility in the deterministic micro theory has prices and income as arguments, but the chosen quantities do not enter as arguments, nor do their corresponding direct utility.

The MSD property is natural, since it implies that the stochastic properties of the utilities are invariant under aggregation of alternatives. To realize this suppose that the univers of alternatives is divided into subsets of alternatives called "aggregate alternatives". Thus each aggregate alternative consists of one or several "basic" alternatives. It is understood that the consumer's choice of an aggregate alternative means that he chooses a basic alternative that belongs to the aggregate one. Consequently, the utility of the aggregate alternative must be the maximum of the utilities of the basic alternatives within the aggregate one. Under MSD, the utility of the aggregate alternative will therefore have the same distribution (apart from a location shift) as the basic utilities.

Theorem 8

Assume that $U_j = v_j + \varepsilon_j$, *where the c.d.f. F of* $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ *does not depend on* $\{v_j\}$.

(i) Then F satisfies DIM if and only if

(3.77) $$F(y_1, y_2, \ldots, y_n) = \psi\left(G\left(e^{-y_1}, e^{-y_2}, \ldots, e^{-y_n}\right)\right)$$

where G is a homogeneous function and ψ *is a positive function (subject to F being a proper distribution function).*

(ii) If $\varepsilon_1, \varepsilon_2, \ldots$ *have a common cumulative distribution function then F satisfies MSD if and only if (3.77) holds.*

A proof of Theorem 8 is given by Robertson and Strauss (1981).

From (3.77) and Theorem 7 we realize that when $\psi(x) = \exp(-x)$ we obtain the GEV class.

Strauss (1979) has proved the following result which follows readily from Theorem 8, and extends the result of Theorem 7. This result shows that the choice probabilities do not depend on ψ.

---

[4] MSD is an acronym for; The Maximum utility has the Same Distribution as the distribution of U_1 + b.

Corollary 2

If (3.77) holds then the choice probabilities are given by

$$P\left(v_j+\varepsilon_j=\max_{k\le n}\left(v_k+\varepsilon_k\right)\right)=\frac{e^{v_j}\,\partial_j G\left(e^{v_1},e^{v_2},\ldots,e^{v_n}\right)}{G\left(e^{v_1},e^{v_2},\ldots,e^{v_n}\right)}.$$

Thus, from Theorem 7 we realize that the class of models determined by (3.53) is equivalent to the GEV class.

Until resently it has not been clear which restrictions on the choice probabilities are implied by the GEV class. Dagsvik (1995) proved that the GEV class is very large; in fact the GEV class yields no other restrictions on the choice probabilities beyond those following from the random utility assumption.

Theorem 9

Assume that $U_j=v_j+\varepsilon_j$, where the cumulative distribution function F of ε does not depend on $\{v_j\}$. If (3.75) holds then IIA holds if and only if

(3.78) $$F\left(y_1,y_2,\ldots,y_n\right)=\psi\left(\sum_{k=1}^{n}e^{-\alpha y_k}\right)$$

where $\alpha>0$ is an arbitrary constant.

A proof of Theorem 9 is given by Strauss (1979).

From (3.78) we realize that when $\psi(x)=\exp(-x)$ we obtain the independent extreme value model.

Example 3.12

Another example is obtained when

(3.79) $$\psi(x)=\frac{1}{1+x},$$

in which case (3.78) yields

(3.80) $$F\left(y_1,y_2,\ldots,y_n\right)=\frac{1}{1+\sum_{k=1}^{n}e^{-\alpha y_k}}.$$

Example 3.13

Assume that

(3.81) $$\psi(x) = \exp\left(-x^{1/\alpha}\right)$$

with $\alpha > 1$. Then (3.78) implies that

(3.82) $$F(y_1, y_2, \ldots, y_n) = \exp\left(-\left(\sum_{k=1}^{n} e^{-\alpha y_k}\right)^{1/\alpha}\right).$$

In this model it can be demonstrated that

(3.83) $$\operatorname{corr}(\varepsilon_i, \varepsilon_j) = 1 - \frac{1}{\alpha^2}$$

which shows that the Luce model is consistent with a random utility model with any correlation (different from zero and one) between the utilities as long as the correlation structure is symmetric.

4. More advanced examples of discrete choice analysis

4.1. Labor supply (I)

Consider the binary decision problem of wanting to work or not. 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

$$C = hW + I \tag{4.1}$$

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

$$h + L \leq M\,(=8760). \tag{4.2}$$

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

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

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

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

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

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

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

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

(5.6)

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, we cannot base the estimation

procedure immediately on (4.6) 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. We assume that the parameter β_2/β_1 depend on age and number of small children, X_2, such that

$$\log\left(\frac{\beta_2}{\beta_1}\right) = X_2 b + \varepsilon_2 \tag{4.7}$$

where ε_2 is a random term which accounts for unobserved variables that affect the preferences and b is a parameter vector. 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

$$\log W = X_1 a + \varepsilon_1 \tag{4.8}$$

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. For simplicity we assume that α_1 is common to all agents. If ε_1 and ε_2 are independent and normally distributed with $E\varepsilon_j = 0$, $\mathrm{Var}\,\varepsilon_j = \sigma_j^2$, 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

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

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 rest of the 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

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

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

Suppose that the random terms ε_1 and ε_2 are i.i. standard extreme value distributed. Then it follows that P_2 becomes a binary logit model given by

(4.12) $$P_2 = \frac{\exp\left(E\log W\right)}{\exp\left(E\log W\right)+\exp\left(E\log W^*\right)} = \frac{1}{1+\exp\left(-X\tilde{s}+\tilde{c}\log I\right)}$$

where $\tilde{s} = s^*\pi/\sqrt{3}$ and $\tilde{c} = c\pi/\sqrt{3}$. From (4.12) we now obtain the elasticity with respect to I as

(4.13) $$\frac{\partial \log P_2}{\partial \log I} = -c\left(1-P_2\right) = -\frac{c}{1+\exp\left(X\tilde{s}-\tilde{c}\log I\right)}.$$

4.2. Labor supply (II)

Consider the choice of whether or not to work. The agent is assumed to face a set B of feasible jobs where job j has wage rate W_j. The set B is unobservable to the econometrician. The econometrician only observe if the agent works or not and the corresponding wage rate if the agent works. Let

(4.14) $$U_j = \theta\log W_j + \varepsilon_j^*,\ j\in B$$

be the utility of job j, where ε_j^* is supposed to account for non-pecuniary aspects with job j, and $\theta > 0$ is a parameter. The utility of not working equals

(4.15) $$U_0 = v_0 + \varepsilon_0$$

where v_0 is a structural term and ε_0 is a random variable. In (4.14), W_j is possibly correlated with ε_j^* and we therefore introduce an instrument variable equation

(4.16) $$\log W_j = X\beta + \eta_j$$

where X is a vector that consists of individual characteristics such as length of education and experience, and η_j is a zero mean random term that may be correlated with ε_j^*. However, we assume that η_j and ε_k^* are independent when $k \neq j$. When (4.16) is inserted into (4.14) we get

(4.17) $$U_j = \theta X\beta + \varepsilon_j$$

where $\varepsilon_j = \varepsilon_j^* + \theta\eta_j$. Let n be the number of jobs in B. If we assume that ε_j, $j = 0,1,2,\ldots,n$, are i.i. standard extreme value distributed then the probability of choosing job j equals

$$P\left(U_j = \max\left(U_0, \max_{k\in B} U_k\right)\right) = \frac{e^{\theta X\beta}}{e^{v_0} + \sum_{k\in B} e^{\theta X\beta}} = \frac{e^{\theta X\beta}}{e^{v_0} + n e^{\theta X\beta}}. \tag{4.18}$$

Hence the probability of working (which is the probability of choosing one of the jobs in B) equals

$$P_2 = \frac{n e^{\theta X\beta}}{e^{v_0} + n e^{\theta X\beta}}. \tag{4.19}$$

Suppose n depends on regional and/or group-specific unemployment rate, Z, in the following manner

$$\log n = \rho Z + \delta \tag{4.20}$$

where ρ and δ are unknown parameters. Then P_2 takes the form

$$P_2 = \frac{1}{1 + \exp(v_0 - \delta - \rho Z - X\beta\theta)}. \tag{4.21}$$

Consider next the estimation of (4.16) from the subsample of working individuals. Since $\varepsilon_j = \varepsilon_j^* + \theta\eta_j$ it follows that the mean of η_j is not necessarily equal to zero, given that j is the chosen alternative, i.e.,

$$E\left(\eta_j \,\middle|\, U_j = \max\left(U_0, \max_{k\in B} U_k\right)\right) \neq 0.$$

Define $\tilde{\eta}_j$ by

$$\eta_j = \alpha\varepsilon_j + \tilde{\eta}_j \tag{4.22}$$

where α is an unknown parameter that is equal to

$$\alpha = \operatorname{cov}(\eta_j, \varepsilon_j)/\operatorname{Var}\varepsilon_j. \tag{4.23}$$

This implies that ε_j and $\tilde{\eta}_j$ are uncorrelated. Moreover, we have, by Lemma 1 in Appendix A that

$$\begin{aligned} E\left(\varepsilon_j \,\middle|\, U_j = \max\left(U_0, \max_{k\in B} U_k\right)\right) &= E\left(U_j \,\middle|\, U_j = \max\left(U_0, \max_{k\in B} U_k\right)\right) - X\beta\theta \\ &= E\max\left(U_0, \max_{k\in B} U_k\right) - X\beta\theta. \end{aligned} \tag{4.24}$$

Under the assumption of extreme value distributed utility terms we get

$$(4.25)\qquad E\max\left(U_0, \max_{k\in B} U_k\right) = \log\left(\sum_{k\in B} e^{X\beta\theta} + e^{v_0}\right) + 0.5772 = \log\left(n\, e^{X\beta\theta} + e^{v_0}\right) + 0.5772.$$

Hence, by combining (4.25) and (4.22) we get

$$\begin{aligned}(4.26)\qquad & E\left(\eta_j \,\middle|\, U_j = \max\left(U_0, \max_{k\in B} U_k\right)\right) = \alpha E\left(\varepsilon_j \,\middle|\, U_j = \max\left(U_0, \max_{k\in B} U_k\right)\right) \\ & = \alpha \log\left(n\, e^{X\beta\theta} + e^{v_0}\right) - \alpha X\beta\theta + 0.5772\alpha \\ & = -\alpha \log P_2 + \alpha \log n + 0.5772\cdot\alpha = -\alpha \log P_2 + \alpha\rho Z + \alpha\delta + 0.5772\cdot\alpha.\end{aligned}$$

Consequently, we can write the wage equation as

$$(4.27)\qquad \log W_j = X\beta - \alpha \log P_2 + \alpha\rho Z + \delta^* + \eta_j^*$$

where $\delta^* = \alpha\delta + 0.5772\cdot\alpha$ and η_j^* is a random term with the property that

$$(4.28)\qquad E\left(\eta_j^* \,\middle|\, U_j = \max\left(U_0, \max_{k\in B} U_k\right)\right) = 0.$$

Thus we can estimate (4.27) consistently from the subsample of working individuals.

Consider finally the conditional variance

$$\mathrm{Var}\left(\eta_j \,\middle|\, U_j = \max\left(U_0, \max_{k\in B} U_k\right)\right).$$

From Lemma 1 in Appendix A we get

$$\begin{aligned}(4.29)\qquad & \mathrm{Var}\left(\varepsilon_j \,\middle|\, U_j = \max\left(U_0, \max_{k\in B} U_k\right)\right) \\ & = \mathrm{Var}\left(U_j \,\middle|\, U_j = \max\left(U_0, \max_{k\in B} U_k\right)\right) \\ & = \mathrm{Var}\left(\max\left(U_0, \max_{k\in B} U_k\right)\right) = \mathrm{Var}\,\varepsilon_j.\end{aligned}$$

The last equality in (4.29) follows from the fact that

$$\max\left(U_0, \max_{k\in B} U_k\right)$$

has the same distribution as ε_j, apart from an additive deterministic term. Consequently, since ε_j and $\tilde{\eta}_j$ are independent,

(4.30) $$\operatorname{Var}\left(\eta_j \,\middle|\, U_j = \max\left(U_0, \max_{k\in B} U_k\right)\right) = \operatorname{Var}\tilde{\eta}_j + \alpha^2 \operatorname{Var}\varepsilon_j = \operatorname{Var}\eta_j .$$

The last result shows that in contrast to the case with normally distributed disturbances, (cf. Heckman, 1979) the conditional variance of η_j given that j is the chosen alternative equals the corresponding unconditional variance.

4.3. Labor supply (III)

Consider an alternative modelling framework to the one discussed in section 4.2. We assume that the agent faces a set B (unobservable) of feasible job opportunities. Let

(4.31) $$U_j = v(W_j) + \varepsilon_j$$

$j = 1,3,\ldots,n,$ be the utility of job j with wage rate W_j, where $v(W_j)$ is the structural part of the utility function that is common to all agents, while ε_j is an agent-specific random term that accounts for non-pecuniary aspect associated with job j. Similarly, let

(4.32) $$U_0 = v_0 + \varepsilon_0$$

be the utility of not working. Suppose furthermore that ε_j, $j = 0,1,\ldots,$ are i.i. standard extreme value distributed.

Let B(w) be the subset of B that consists of all feasible jobs with wage rate w, and let n(w) be the number of jobs in B(w), and let D be the set of all possible wages. The probability of choosing job j in B equals

(4.33) $$P_j \equiv P\left(U_j = \max\left(U_0, \max_{k\in B} U_k\right)\right) = \frac{e^{v(W_j)}}{e^{v_0} + \sum_{k\in B} e^{v(W_k)}}$$
$$= \frac{e^{v(W_j)}}{e^{v_0} + \sum_{y\in D}\sum_{k\in B(y)} e^{v(W_k)}} = \frac{e^{v(W_j)}}{e^{v_0} + \sum_{y\in D} n(y) e^{v(y)}}.$$

Hence the probability of choosing a job with wage rate w equals

(4.34) $$P(w) \equiv \sum_{j\in B(w)} P_j = \frac{\sum_{j\in B(w)} e^{v(W_j)}}{e^{v_0} + \sum_{y\in D} n(y) e^{v(y)}}$$
$$= \frac{n(w) e^{v(w)}}{e^{v_0} + \sum_{y\in D} n(y) e^{v(y)}} = \frac{e^{\tilde{v}(w)}}{e^{v_0} + \sum_{y\in D} e^{\tilde{v}(y)}}$$

where

$$\tilde{v}(y) = \log n(y) + v(y). \tag{4.35}$$

From (4.35) we realize that we cannot without further assumptions separate n(w) from v(w). To this end suppose that the agent also receives nonlabor income. For example, a married woman or man may receive income from the spouse. In this case

$$v(w) = v^*(w + I) \tag{4.36}$$

where I denoted nonlabor income, and $v^*(\cdot)$ is a concave parametric function.

4.4. 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

$$U_j(x) = U_{1j} + U_2(x). \tag{4.37}$$

The budget constraint is given by

$$p'x = y - q_j, x \geq 0, \tag{4.38}$$

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 (5.39) we obtain the conditional indirect utility, given j, as

$$V_j(p, y - q_j) = U_{1j} + V_2(p, y - q_j) \tag{4.39}$$

where

$$V_2(p, y) = \max_{p'x = y} U_2(x). \tag{4.40}$$

Assume that

$$U_{1j} = \beta \log T_j + \varepsilon_j \tag{4.41}$$

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

(4.42) $$V_2(p, y - q_j) = V_3(p) + \theta \log(y - q_j)$$

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

(4.43) $$V_j = \beta \log T_j + \theta \log(y - q_j) + \varepsilon_j$$

which implies that

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

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

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

4.5. Firms' location of plants (I)

In this example we outline a framework for analyzing firms' location of plants. Specifically, we assume that the firms face the choice of establishing a plant in one of m differents sites (counties). Suppose furthermore that firms profit functions (or expected profit functions) depend on observable characteristics that are common for all sites within particular regions. Let C_r denote the set of counties within region r, $r = 1,2,\ldots,m$, and let n_r be the number of counties in C_r. The regional attributes of interest may be population density and macro indicators that describe the industry structure. Finally, certain tax rates may differ across regions (tax shelters). Consider an arbitrarily selected firm. Let $U_{rj} = v_r + \varepsilon_{rj}$ denote the firms utility of establishing a plant in county $j \in C_r$, where $\{\varepsilon_{rj}\}$ are i.i. standard extreme value distributed terms that account for unobserved region and county-specific attributes and $\{v_r\}$ are structural terms that depend on the attributes specific to region r. Let P_{rj} be the probability of a location in county j in region r. We get

(4.46) $$P_{rj} \equiv P\left(U_{rj} = \max_i \left(\max_{k \in C_i} U_{ik}\right)\right) = \frac{e^{v_r}}{\sum_{i=1}^{m} \sum_{k \in C_i} e^{v_r}} = \frac{e^{v_r}}{\sum_{i=1}^{m} n_i e^{v_i}}.$$

Hence, we get that the probability of a location within region r equals

(4.47) $$P_r = \sum_{j \in C_r} P_{rj} = \frac{n_r e^{v_r}}{\sum_{i=1}^{m} n_i e^{v_i}} = \frac{e^{\tilde{v}_r}}{\sum_{i=1}^{m} e^{\tilde{v}_i}},$$

where

(4.48) $$\tilde{v}_r = v_r + \log n_r .$$

If we assume that $v_r = Z_r\beta$, where Z_r is the vector of observable attributes associated with region r, we get

(4.49) $$\tilde{v}_r = Z_r\beta + \log n_r .$$

4.6. Firms' location of plants (II)

We now consider an extension of the setting in Section 4.5. Suppose now that the error terms for counties within a common region are correlated. This may be a plausible assumption since it is often the case that counties within regions are more homogeneous than counties across regions. We shall now apply the nested logit framework to model this case. Let

(4.50) $$G(\mathbf{y}) = \sum_{r=1}^{m} \left(\sum_{j=1}^{n_r} y_{rj}^{1/\theta} \right)^{\theta}$$

and let

$$F(\mathbf{x}) = \exp\left(-G\left(e^{-x_{11}}, e^{-x_{12}}, \ldots\right)\right)$$

be the joint distribution function of $\left(\varepsilon_{11}, \ldots, \varepsilon_{1n_1}, \ldots, \varepsilon_{m1}, \ldots, \varepsilon_{mn_m}\right)$. Then it follows that

(4.51) $$\operatorname{corr}\left(\varepsilon_{ri}, \varepsilon_{rj}\right) = 1 - \theta^2$$

for $i \neq j$, $i, j \in C_r$, and

(4.52) $$\operatorname{corr}\left(\varepsilon_{ri}, \varepsilon_{sj}\right) = 0$$

for $i \in C_r$, $j \in C_s$, $r \neq s$, where $0 < \theta \leq 1$. From Theorem 7 we get

(4.53)
$$P_{rj} = \frac{\left(\sum_{j\in C_r} e^{v_i/\theta}\right)^{\theta-1} e^{v_r/\theta}}{\sum_{i=1}^{m}\left(\sum_{j\in C_i} e^{v_i/\theta}\right)^{\theta}} = \frac{e^{v_r} n_r^{1/\theta}}{\sum_{i=1}^{m} e^{v_i} n_i^{1/\theta}} \cdot \frac{1}{n_r}$$

which yields

(4.54)
$$P_r = \sum_{j\in C_r} P_{rj} = \frac{e^{v_r} n_r^{1/\theta}}{\sum_{i=1}^{m} e^{v_i} n_i^{1/\theta}} = \frac{e^{v_r^*}}{\sum_{i=1}^{m} e^{v_i^*}}$$

where

(4.55)
$$v_r^* = v_r + \frac{1}{\theta}\log n_r = Z_r\beta + \frac{1}{\theta}\log n_r .$$

Provided $n_1, n_2, \ldots,$ are known we can estimate β and $1/\theta$ from observations on plant locations with $\{Z_r, \log n_r\}$ as explanatory variables.

From (4.54) we get

(4.56)
$$\frac{\partial \log P_r}{\partial \log n_r} = \frac{1}{\theta}(1 - P_r)$$

and

(4.57)
$$\frac{\partial \log P_k}{\partial \log n_r} = -\frac{1}{\theta} P_r$$

for $k \neq r$. The interpretation of (4.56) and (4.57) is as the effect from increasing the size of C_r. For example, one may wish to assess the effect of changing the number of counties that belong to a region with "tax shelters".

4.7. Firms' location of plants (III)

The setting here is the same as the one in Section 4.6. Suppose now that $\{n_r\}$ are unobservable, but that we observe the number of locations in at least one county in each region, say in county number one. Let M_{r1} be the observed number of locations in county one in C_r, and let M_r be the total number of observed locations within region r. Finally, let $M = \sum_{r=1}^{m} M_r$. Then M_{r1}/M_r is an estimate of P_{r1} and M_r/M is an estimate of P_r. Since by (4.53)

$$P_{r1} = P_r \cdot \frac{1}{n_r}$$

it follows that consistent estimates for n_r is given by

(4.58) $$\hat{n}_r = \frac{M_r^2}{M_{r1}\, M}, \quad r = 1,2,\ldots,m.$$

4.8. Potential demand for alternative fuel vehicles

This example is taken from Dagsvik et al. (1996). 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

(4.59) $$U_j(t) = Z_j(t)\beta + \mu_j + \varepsilon_j(t)$$

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 section 3.4 it follows that if $\{\varepsilon_j(t)\}$ are assumed to be i.i. standard extreme value distributed then

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

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

$$(4.61)\qquad \ell = \sum_{t=1}^{15} \left(\sum_{h\in A} \sum_{i} \sum_{j} Y_{ij}^{h}(t) \log P_{ijt}(C_A) + \sum_{h\in B} \sum_{i} \sum_{j} Y_{ij}^{h}(t) \log P_{ijt}(C_B) \right).$$

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	-2.176	-1.549	-2.159	-1.550	-1.394
	(-17.7)	(-15.2)	(-15.0)	(-20.6)	(-11.9)	(-11.8)
Top speed (100 km/h)	-0.274	0.488	-0.820	-0.571	-0.320	-0.339
	(-0.9)	(1.5)	(-3.3)	(-2.4)	(-1.1)	(-1.2)
Driving range (1 000 km)	1.861	2.130	1.018	1.465	0.140	1.000
	(3.1)	(3.3)	(2.0)	(3.2)	(0.2)	(1.8)
Fuel consumption (liter per 10 km)	-0.902	-1.692	-0.624	-1.509	-0.446	-1.030
	(-3.0)	(-5.1)	(-2.5)	(6.7)	(-1.5)	(-3.7)
Dummy, electric	0.890	-0.448	0.627	-0.180	0.765	-0.195
	(4.2)	(-2.0)	(3.6)	(-1.1)	(3.6)	(-1.0)
Dummy, hybrid	1.185	0.461	1.380	0.649	1.216	0.666
	(7.6)	(2.8)	(10.6)	(5.6)	(7.7)	(4.6)
Dummy, lpg	1.010	0.236	0.945	0.778	0.698	0.676
	(8.2)	(1.9)	(9.2)	(8.5)	(5.7)	(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.

4.9. Oligopoly with product differentiation

This example is taken from Anderson et al. (1994). Consider n 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

$$U_{ij} = y_i + a_j - w_j + \sigma\varepsilon_{ij}. \tag{4.62}$$

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}, $j = 1,2,\ldots,n$, $i = 1,2,\ldots,N$, are i.i. standard extreme value distributed we get that the aggregate demand for variant j equals NP_j where

$$P_j = Q_j(\mathbf{w}) \equiv \frac{\exp\left(\frac{a_j - w_j}{\sigma}\right)}{\sum_{k=1}^{n} \exp\left(\frac{a_k - w_k}{\sigma}\right)}. \tag{4.63}$$

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;

$$\pi_j = (w_j - c_j) N Q_j(\mathbf{w}) - K_j. \tag{4.64}$$

Now firm j takes the prices set by other firms as given and chooses the price of variant j that maximizes (4.64). Anderson et al. (1994) demonstrate that there exists a unique Nash equilibrium set of prices, $\mathbf{w}^* = (w_1^*, w_2^*, \ldots, w_n^*)$ which are determined by

$$w_j^* = c_j + \frac{\sigma}{1 - Q_j(\mathbf{w}^*)}. \tag{4.65}$$

Thus, when estimating the model (4.63) one should take into account the additional restrictions determined by (4.65).

4.10. 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

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

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

$$P_j(B_k) = \frac{e^{v_j}}{\sum_{k \in B_k} e^{v_k}}, \quad j \in B_k, \tag{4.66}$$

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 $\hat{P}_{kj}, j \in B_k$. 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 based on the 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

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

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

$$\operatorname{corr}(U_i, U_5) = 0,$$

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

$$P_j(B) = \frac{e^{v_j/\theta}}{\sum_{k \in B} e^{v_k/\theta}} \tag{4.67}$$

when $5 \notin B$,

(4.68)
$$P_j(B) = \frac{e^{v_j/\theta}\left(\sum_{k\in B\setminus\{5\}} e^{v_k/\theta}\right)^{\theta-1}}{e^{v_5} + \left(\sum_{k\in B\setminus\{5\}} e^{v_k/\theta}\right)^{\theta}}$$

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

(4.69)
$$P_5(B) = \frac{e^{v_5}}{e^{v_5} + \left(\sum_{k\in B\setminus\{5\}} e^{v_k/\theta}\right)^{\theta}}.$$

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 $-2(\ell_2 - \ell^*)$ is (asymptotically) Chi squared distributed with 19 degrees of freedom. The corresponding critical value is 30.1 at 5 per cent significance level and therefore the estimate of $-2(\ell_2 - \ell^*)$ in Table 4.4 implies that the nested logit model is not rejected against the non-structural multinomial model. Thus, in terms of goodness-of-fit there seems to be an essential difference between the logit and the nested logit formulation. However, as measured by McFaddens ρ^2, the difference in goodness-of-fit is only one per cent! This shows that one should be very cautious when interpreting ρ^2.

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	
$-2(\ell_j - \ell^*)$	38.2		21.6	

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.

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	
	Oserved		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. Discrete/continuous choice

5.1. 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 motivated from the latent variable specification (2.7), in section 2.1.1, but in contrast to the case described there we now also observe the left hand side variable when it is positive. Thus we observe Y defined by

$$Y = \begin{cases} X\beta + u\sigma & \text{if } X\beta + u\sigma > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5.1}$$

where $\sigma > 0$ is a scale parameter, and u is a zero mean random variable with cumulative distribution function $F(\cdot)$. Another way of expressing (5.1) is as

$$Y = \max(0, X\beta + u\sigma). \tag{5.2}$$

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. 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 preference for leisure. Since the actual hours og work always will be non-negative we therefore get the structure (5.1).

5.2. The general structural setting

Models such as the Tobit one account for some of the statistical nature of the data, but is not structural in a "deep" sense. We shall now discuss structural specifications derived from choice theory. In many situations a decision-maker makes interrelated choices where one choice is discrete and the other is continuous. For example, a worker may face the decision problem of which job to choose and how many hours to work, (conditional on the choice of job). Another example is a consumer that-considers purchasing electric versus gas appliances, as well as how much electricity or gas to consume. A third example is a household that chooses which type of car to own and the intensity of car use.

Such choice situations are called discrete/continuous, reflecting the fact that the choice set along one dimension is discrete while it is continuous along another dimension. Theories and methods for specifying and estimating structural models for discrete/continuous choice have been developed among others by Heckman (1974, 1979), Dubin and McFadden (1984), Lee and Trost (1978), King (1980) and Dagsvik (1994).

We now consider an agent that faces two choices; first which alternative to choose, from a finite and exhaustive set of mutually exclusive alternatives, and second; how much of a particular good to consume. Since it is often the case that these choices depend on the same underlying factors this should be taken into account in the formulation of the model and in the corresponding econometric specification. Suppose for expository simplicity that there are only two continuous goods. Let $U_j(x_1, x_2)$ be the utility of alternative (j, x_1, x_2), where $j = 1,2,\ldots,$ indexes the discrete alternatives and (x_1, x_2) the continuous ones. Thus the agent's optimization problem is to maximize $U_j(x_1, x_2)$ with respect to (j, x_1, x_2) subject to the budget constraints $j \in B$ and

$$x_1 p_1 + x_2 p_2 + \sum_k \delta_k c_k = y, \; x_1 \geq 0, \; x_2 \geq 0, \tag{5.3}$$

where B is the choice set of feasible (discrete) alternatives, p_1, p_2 are prices, y is the agent's income (exogenous), c_j is the cost (or annual user cost) of the discrete alternative j and $\delta_k = 1$ if alternative $k \in B$ is chosen and zero otherwise. Consider now the continuous choice given the discrete alternative j. Let

$$V_j(p, y - c_j) = \max_{\substack{x_1 p_1 + x_2 p_2 = y - c_j \\ x_1 \geq 0, x_2 \geq 0}} U_j(x_1, x_2) \tag{5.4}$$

which means that $V_j(p, y - c_j)$ is the conditional indirect utility, given that the discrete alternative j is chosen. Since $V_j(p, y - c_j)$ expresses the highest possible utility conditional on alternative j, it must be the case that alternative j is chosen if

$$V_j(p, y - c_j) = \max_{k \in B} V_k(p, y - c_k). \tag{5.5}$$

Second, it follows from Roy's identity that under standard regularity conditions we obtain the corresponding continuous demands by

$$\tilde{x}_{rj} = -\frac{\partial V_j(p, y - c_j)/\partial p_r}{\partial V_j(p, y - c_j)/\partial y} \tag{5.6}$$

for $r = 1,2$, given that j is the preferred discrete alternative, i.e., given that (5.5) holds. Thus the discrete as well as the continuous choices are here derived from a common representation of the preferences.

It is known from duality theory that under standard regularity conditions the specification of the indirect utility is equivalent to the specification of the corresponding direct utility. Therefore, in econometric model building, it is convenient to start with a parametric functional form of the indirect utility function, including alternative-specific random terms.

5.3. The Gorman Polar functional form

When the conditional indirect utility function belongs to the class of functional forms called "Gorman Polar forms", (Gorman, 1953), then the structure of the demand equations and choice probabilities become particularly convenient. The Gorman Polar functional form is given by

$$V_j \equiv V_j(p, y - c_j) = \frac{y - c_j + a(p)(\varepsilon_j + m_j)}{b(p)} \tag{5.7}$$

where $a(\cdot)$ and $b(\cdot)$ are functions that are homogeneous of degree one, concave and non-decreasing in p and $\{m_j\}$ are alternative-specific terms which are independent of prices and income. It then follows that V_j is non-increasing and convex in prices. Here $\{\varepsilon_j\}$ are random terms that are supposed to account for unobservables that affect preferences and m_j is (possibly) a function of observable attributes associated with alternative j.

From (5.7) it follows that the choice probabilities are given by

$$P_j(B) = P\left(\varepsilon_j + m_j - \frac{c_j}{a(p)} = \max_{k \in B}\left(\varepsilon_k + m_k - \frac{c_k}{a(p)}\right)\right). \tag{5.8}$$

In case $\{\varepsilon_j\}$ are i.i. extreme value distributed we obtain

$$P_j(B) = \frac{\exp\left(m_j - c_j/a(p)\right)}{\sum_{k \in B} \exp\left(m_k - c_k/a(p)\right)}. \tag{5.9}$$

By Roy's identity we obtain the demands as

$$\tilde{x}_{rj} = \left(\frac{a(p)\, b_r(p)}{b(p)} - a_r(p)\right) m_j - \left(y - c_j\right)\frac{b_r(p)}{b(p)} + \left(\frac{a(p)\, b_r(p)}{b(p)} - a_r(p)\right)\varepsilon_j \tag{5.10}$$

where $a_r(p)$ and $b_r(p)$ denote the respective partial derivatives with respect to component r.

Recall, however, that due to the selectivity problem we cannot automatically apply standard methods to estimate (5.10), as we shall discuss in further detail below.

Example 5.1

Assume that the conditional indirect utility function has the form

$$V_j(p, y - c_j) = \log\left[\left(Z_j\alpha + \beta_{j1}p_1 + \beta_{j2}p_2 + \theta(y - c_j) + \varepsilon_j\right)e^{-\mu\theta p_1}\right] - \beta_3 \log p_2 \tag{5.11}$$

where $\{\varepsilon_j\}$ are i.i. standard extreme value distributed random terms which have mean 0.5772 and α, β_{jr}, β_3, $r = 1,2$, θ and μ are unknown parameters.[5] The specification (5.11) has been applied by Dubin and McFadden (1984). However, (5.11) is not a Gorman Polar functional form. First, we obtain

$$\frac{\partial V_j(p, y - c_j)}{\partial p_1} = \beta_{j1} e^{-\mu p_1} - \mu V_j(p, y - c_j) \tag{5.12}$$

and

$$\frac{\partial V_j(p, y - c_j)}{\partial y} = \theta e^{\mu p_1}. \tag{5.13}$$

Consequently, by (5.6)

$$\tilde{x}_{1j} = \left(Z_j\alpha + \beta_{j1}p_1 + \beta_{j2}p_2 + \theta(y - c_j)\right)\mu - \frac{\beta_{j1}}{\theta} + \varepsilon_j\mu. \tag{5.14}$$

Second, note that maximization of $V_j(p, y - c_j)$ in (5.11) with respect to j is equivalent to maximizing

$$Z_j\alpha + \beta_{j1}p_1 + \beta_{j2}p_2 + \theta(y - c_j) + \varepsilon_j,$$

since $\exp(-\mu\theta p_1)$ does not depend on j. Therefore, the probability of choosing alternative j equals

---

[5] Note that (5.11) is not homogeneous of degree zero in prices and income. We may, however, interpret (5.11) as an indirect utility function in *normalized* prices and income. This is possible because a function v(p,y) of normalized prices and income is the indirect utility function of some locally nonsatiated utility function if and only if it is lower semicontinuous, quasi-convex, increasing in y, nonincreasing in p, and has $v(\lambda p, \lambda y)$ nondecreasing in λ.

$$P_j = P\left(Z_j\alpha+\beta_{j1}p_1+\beta_{j2}p_2+\theta(y-c_j)+\varepsilon_j = \max_k\left(Z_k\alpha+\beta_{k1}p_1+\beta_{k2}p_2+\theta(y-c_k)+\varepsilon_k\right)\right)$$

(5.15)
$$= \frac{\exp\left(Z_j\alpha+\beta_{j1}p_1+\beta_{j2}p_2-\theta c_j\right)}{\sum_k \exp\left(Z_k\alpha+\beta_{k1}p_1+\beta_{k2}p_2-\theta c_k\right)}.$$

Recall that while the unconditional mean of ε_j is 0.5772 the conditional mean of ε_j given that j is the *chosen* alternative is *not* equal to 0.5772. By Lemma A2 in Appendix A we have that when $\{\varepsilon_j\}$ are extreme value distributed then

(5.16)
$$E\left(\left(v_j+\varepsilon_j\right)\mu \,\middle|\, v_j+\varepsilon_j = \max_k\left(v_k+\varepsilon_k\right)\right) = \mu\, E\max_k\left(v_k+\varepsilon_k\right).$$

Since by Lemma 1 in Appendix A

$$\mu\max_k\left(v_k+\varepsilon_k\right) = \mu\log\left(\sum_k e^{v_k}\right)+\varepsilon\mu,$$

where ε has the same distribution as ε_k, it follows that

$$\begin{aligned}
&\mu E\left(\varepsilon_j \,\middle|\, v_j+\varepsilon_j = \max_k\left(v_k+\varepsilon_k\right)\right) = \mu E\left(v_j+\varepsilon_j \,\middle|\, v_j+\varepsilon_j = \max_k\left(v_k+\varepsilon_k\right)\right) - v_j\mu \\
&= \mu E\left(\max_k\left(v_k+\varepsilon_k\right)\right) - v_j\mu = \mu\log\left(\sum_k e^{v_k}\right) - v_j\mu + E\varepsilon\mu \\
&= \mu\log\left(\sum_k e^{v_k}\right) - v_j\mu + 0.5772\mu.
\end{aligned} \tag{5.17}$$

From this result it follows that

$$\begin{aligned}
&E\left(\tilde{x}_{1j} \,\middle|\, V_j(p, y-c_j) = \max_k V_k(p, y-c_k)\right) \\
&= \left(Z_j\alpha+\beta_{j1}p_1+\beta_{j2}p_2+\theta(y-c_j)\right)\mu - \frac{\beta_{j1}}{\theta} + 0.5772\mu \\
&- \left(Z_j\alpha+\beta_{j1}p_1+\beta_{j2}p_2-\theta c_j\right)\mu + \mu\log\left(\sum_k \exp\left(Z_k\alpha+\beta_{k1}p_1+\beta_{k2}p_2-\theta c_k\right)\right) \\
&= 0.5772\mu - \frac{\beta_{j1}}{\theta} + \theta\mu y + \mu\log\left(\sum_k \exp\left(Z_k\alpha+\beta_{k1}p_1+\beta_{2k}p_2-\theta c_k\right)\right).
\end{aligned} \tag{5.18}$$

The interpretation of (5.18) is as the mean demand of good one given that j is the preferred discrete alternative.

The result in (5.18) implies that if one runs regression analysis based directly on (5.14) this will produce biased estimates. Instead one should apply the specification

(5.19)
$$\tilde{x}_{1j} = \beta^*_{j1} + \theta\mu y + \mu\log\left(\sum_k \exp\left(Z_k\alpha+\beta_{k1}\,p_1+\beta_{2k}\,p_2-\theta c_k\right)\right)+\eta_j$$

where $\beta_{j1}^{*} = 0.5771\mu - \beta_{j1}/\theta$ and η_j is a random error term with the property that the mean of η_j given that j is the chosen alternative equals zero. The estimation can be carried out in two steps: First estimate α, β_{k1}, β_{k2} and θ by the maximum likelihood procedure. Second apply these estimates to compute

$$\log\left(\sum_k \exp\left(Z_k\alpha + \beta_{k1}\, p_1 + \beta_{2k}\, p_2 - \theta c_k\right)\right)$$

which, analogous to Heckman's two stage procedure, is used as a known regressor in (5.19), and the remaining parameters θ, β_{j1}^{*} and μ can be estimated by OLS in a second stage.

Example 5.2

Assume that the conditional indirect utility has the Gorman Polar form with

(5.20) $$a(p) = a_0 \prod_k p_k^{\alpha_k}$$

and

(5.21) $$b(p) = b_0 \prod_k p_k^{\beta_k}$$

where a_0, b_0, α_k, β_k are positive and

$$\sum_k \alpha_k = \sum_k \alpha_k = 1.$$

From (5.10), (5.20) and (5.21) it follows that

(5.22) $$\tilde{x}_{rj}\, p_r = a(p)\left(\beta_r - \alpha_r\right)m_j - \left(y - c_j\right)\beta_r + a(p)\left(\beta_r - \alpha_r\right)\varepsilon_j .$$

If $\{\varepsilon_j\}$ are standard extreme value distributed the discrete choice probabilities are as in (5.9) with (5.20) inserted. If for example $m_j = Z_j\gamma + \delta$, where Z_j is an observable attribute vector and γ and δ are parameters, then if $\{Z_j\}, \{c_j\}$ and $\{p_j\}$ vary sufficiently across a randomly selected sample of agents then it is possible to estimate γ, $\{\alpha_k\}$ and a_0 from observations on the agents' discrete choices. The remaining paramaters to be estimated are $\{\beta_r\}$ and δ. These paramters can be estimated in a second stage by applying (5.22) and controlling for the selectivity bias as explained in Example 5.1.

5.4. Perfect substitute models

We now consider choice problems in which there are $m+1$ goods of which m brands are perfect substitutes, cf. Hanemann (1984). The utility function has the structure

(5.23) $$\tilde{U}(x,\psi,z)=U\left(\sum_{k=1}^{m}\psi_k x_k, z\right)$$

and the budget constraint is

(5.24) $$\sum_{k=1}^{m} p_k x_k + z = y.$$

Here, $\{\psi_k\}$ are unknown parameters and U is a conventional utility function. Letting $\psi_k x_k = z_k$, the corresponding utility maximization problem can be written as

(5.25) $$\max U\left(\sum_{k=1}^{m} z_k, z\right)$$

subject to

(5.26) $$\sum_{k=1}^{m} \frac{p_k}{\psi_k} z_k + z = y,\ x_k \geq 0.$$

Clearly, this maximization problem implies a "corner" solution where the consumer selects the brand with the lowest "price", $p_k^* \equiv p_k/\psi_k$. Thus, brand j is chosen if

(5.27) $$\frac{p_j}{\psi_j} = \min_k\left(\frac{p_k}{\psi_k}\right)$$

while $x_k = 0$, for $k \neq j$. The corresponding indirect utility equals

(5.28) $$V_j \equiv \max_{z+z_j p_j/\psi_j = y} U(z_j, z) = V\left(\frac{p_j}{\psi_j}, y\right)$$

where V(q,y) is the indirect utility that corresponds to the direct utility $U(z_j, z)$, i.e.,

(5.29) $$V(q,y) = \max_{z+qz_j=y} U(z_j, z)$$

Now assume that

(5.30) $$\log \psi_j = Z_j \beta/\mu + \varepsilon_j/\mu$$

where Z_j is a vector of non-pecuniary attributes associated with brand j while β and $\mu > 0$ are unknown parameters and ε_j are i.i. standard extreme value distributed. Now from (5.23) and (5.25) we obtain that brand j is chosen if $U_j = \max_k U_k$, where

$$U_j = Z_j\beta - \mu \log p_j + \varepsilon_j, \tag{5.31}$$

and therefore the choice probabilities are given by

$$P_j = \frac{\exp\left(Z_j\beta - \mu \log p_j\right)}{\sum_k \exp\left(Z_k\beta - \mu \log p_k\right)}. \tag{5.32}$$

Note that in this case there are no fixed costs associated with the discrete choice. As above the continuous demands follow by applying Roy's identity.

Example 5.3 (Hanemann, 1984, p. 550)

Let

$$V(q,y) = \frac{\theta q^{1-\rho}}{\rho - 1} - \frac{e^{-\eta y}}{\eta}, \quad \theta > 0, \quad \eta \neq 0,$$

which yields

$$-\frac{\partial_1 V(q,y)}{\partial_2 V(q,y)} = \theta q^{-\rho} e^{\eta y}, \tag{5.33}$$

where ∂_1 and ∂_2 denote the respective partial derivatives, and therefore it follows from (5.27) that the continuous demand for brand j is given by

$$\tilde{x}_j = -\frac{\partial_1 V\left(\frac{p_j}{\psi_j}, y\right)}{\psi_j \partial_2 V\left(\frac{p_j}{\psi_j}, y\right)} = \theta p_j^{-\rho} \psi_j^{\rho-1} e^{\eta y}. \tag{5.34}$$

From (5.34) and (5.30) we get

$$\begin{aligned} \log\left(\tilde{x}_j p_j\right) &= \log\theta + (\rho-1)\log\psi_j + (1-\rho)\log p_j + \eta y \\ &= \log\theta + \frac{(\rho-1)}{\mu} Z_j\beta + (1-\rho)\log p_j + \eta y + \frac{(\rho-1)}{\mu}\varepsilon_j. \end{aligned} \tag{5.35}$$

Hence, it follows that

$$E\left(\log\left(\tilde{x}\,p_j\right)\middle|\, U_j = \max_k U_k\right) = \log\theta + \eta\, y + \frac{(\rho-1)}{\mu} E\left(U_j \middle|\, U_j = \max_k U_k\right).$$

From Lemma A2 in Appendix we have that

$$\begin{aligned} E\left(U_j \middle|\, U_j = \max_k U_k\right) &= E \max U_k \\ &= 0.5772 + \log\left(\sum_k \exp\left(Z_k\beta - \mu \log p_k\right)\right) \end{aligned} \tag{5.36}$$

which implies that

$$\begin{aligned} &E\left(\log\left(\tilde{x}_j\, p_j\right)\middle|\, U_j = \max_k U_k\right) \\ &= \log\theta + 0.5772\frac{(\rho-1)}{\mu} + \eta\, y + \frac{(\rho-1)}{\mu}\log\left(\sum_k \exp\left(Z_k\beta - \mu \log p_k\right)\right). \end{aligned} \tag{5.37}$$

Similarly, Lemma A2 implies that

$$\mathrm{Var}\left(U_j \middle|\, U_j = \max_k U_k\right) = \mathrm{Var}\left(\max_k U_k\right). \tag{5.38}$$

Note that in the conditional expectations and variance above it is implicitly understood that y and $\{Z_k\}$ are *given*. Apart from an additive deterministic term $\max_k U_k$ has the same distribution as ε_j. Consequently, (5.35) implies that

$$\mathrm{Var}\left(\log\left(\tilde{x}_j\, p_j\right)\middle|\, U_j = \max_k U_k\right) = \mathrm{Var}\left(\frac{\rho-1}{\mu}\varepsilon_j\right) = \frac{(\rho-1)^2\,\pi^2}{6\mu^2}. \tag{5.39}$$

Suppose now that our sample only consists of a simple cross-section. Then, since $\{Z_k\}$ do not vary across individuals we may write

$$\log\left(\tilde{x}_j\, p_j\right) = a + \eta\, y + \delta_j \tag{5.40}$$

where

$$a = \log\theta + 0.5772\frac{(\rho-1)}{\mu} + \frac{(\rho-1)}{\mu}\log\left(\sum_k \exp\left(Z_k\beta - \mu \log p_k\right)\right) \tag{5.41}$$

and δ_j is a random term which due to (5.41) has the property that

$$E\left(\delta_j \middle|\, U_j = \max_k U_k\right) = 0$$

$$\mathrm{Var}\left(\delta_j \,\middle|\, U_j = \max_k U_k\right) = \frac{(\rho-1)^2\,\pi^2}{6\mu^2}.$$

Thus, the model parameters can be estimated in two stages.

Stage 1: Estimate β and μ from data on the discrete choices by means of model (5.32).

Stage 2: Estimate a, η and $\rho - 1/\mu$ on the basis on (5.40). By inserting the estimates of a, μ, $\rho - 1$ and β in (5.41) an estimate of a can be obtained.

Similarly to (5.37) it is easy to prove that

$$\begin{aligned}&\log E\left(\tilde{x}_j\, p_j \,\middle|\, U_j = \max_k U_k\right)\\ &= \log\theta + \log\Gamma\left(1+\frac{1-\rho}{\mu}\right) + \eta\, y + \frac{(\rho-1)}{\mu}\log\left(\sum_k \exp\left(Z_k\beta - \mu\log p_k\right)\right)\end{aligned} \tag{5.42}$$

where $\Gamma(\cdot)$ is the Gamma function. Suppose now that microdata are not available but that one has macro time series data for $\{Z_k\}$, $\tilde{x}_j p_j$, prices, the mean income, y, and the aggregate shares, $\{P_j\}$. Then it is possible to use (5.32) and (5.42) to estimate the unknown parameters in a two-stage procedure.

6. Estimation

We shall briefly review maximum likelihood estimation, Berkson's method and finally Heckman's two stage method.

6.1. Maximum likelihood

Suppose the multinomial probability model has been specified, for example as (2.2), (2.4), (2.6), or as a binary Probit model. Let $Y_{ij} = 1$, if agent i in a sample of randomly selected agents, falls into category j and zero otherwise, and let $\{H_j(X_i)\}$ be the corresponding multinomial logit probabilities given by (2.2) where X_i is the vector of explanatory variables for agent i. The total likelihood of the observed outcome equals

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

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

$$\ell = \sum_{i=1}^{N} \sum_{j=1}^{m} Y_{ij} \log H_j(X_i). \tag{6.1}$$

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

$$\frac{\partial \ell}{\partial \beta_{rk}} = \sum_{i=1}^{N} \left(X_{ik} - H_r(X_i) X_{ik}\right) = 0 \tag{6.2}$$

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

When the logit model has the structure (2.6) then the first order conditions yield

$$\frac{\partial \ell}{\partial \beta_k} = \sum_{i=1}^{N} \sum_{j=1}^{m} \left(Y_{ij} - H_j(\mathbf{Z}, X_i)\right) R_k(Z_j, X_i) = 0 \tag{6.3}$$

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

McFadden (1973) has proved that when the probabilities are given by (2.6), the loglikelihood function is globally strictly concave, and therefore a unique solution to (2.15) is guarantied.

6.2. Berkson's 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 (2.4) 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 (2.5). We then obtain

$$\log\left(\frac{\hat{H}_j}{\hat{H}_1}\right) = (Z_j - Z_1)\beta + \eta_j, \tag{6.4}$$

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

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

which shows that

$$\begin{aligned} E\eta_j &= E\log\left(\frac{\hat{H}_j}{\hat{H}_1}\right) - (Z_j - Z_1)\beta \\ &\approx \log\left(\frac{H_j}{H_1}\right) + \frac{E\hat{H}_j - H_j}{H_j} - \frac{(E\hat{H}_1 - H_1)}{H_1} - (Z_1 - Z_1)\beta \\ &= \log\left(\frac{H_j}{H_1}\right) - (Z_j - Z_1)\beta = 0. \end{aligned} \tag{6.5}$$

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 (2.16) we can estimate β by regression analysis with $\{\tilde{Y}_j\}$ as dependent variables and $\{Z_j - Z_1\}$ as independent variables.

6.3. Maximum likelihood estimation of the Tobit model

Notice first that due to the form of (5.2) ordinary regression analysis will not do because of the nonlinear operation on the right hand side of (5.2).

From (5.2) it follows that

$$P(Y=0)=P(u\le -X\beta/\sigma)=F(-X\beta/\sigma) \tag{6.6}$$

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

$$P(Y\in(y,y+dy))=P(u\sigma\in(y-X\beta,y+dy-X\beta))=\frac{1}{\sigma}F'\left(\frac{y-X\beta}{\sigma}\right)dy, \tag{6.7}$$

for $y>0$. Consider now the estimation of the unknown parameters based on observations from a random sample of N 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 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).

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

From (6.7) 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 (6.6), the probability that $i\in S_0$ equals

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

Therefore the total loglikelihood equals

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

Example 6.1

Suppose F(y) is a standard normal distribution function, $\Phi(y)$. Then since

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

it follows that the loglikelihood in this case reduces to

$$\ell = -\sum_{i\in S_1} \frac{(Y_i - X_i\beta)^2}{2\sigma^2} - N\log\sigma + \sum_{i\in S_0} \log\Phi\left(\frac{-X_i\beta}{\sigma}\right). \tag{6.9}$$

We realize that applying OLS to the equation $Y = X\beta + u\sigma$ correspond to neglecting the last term in (2.20) and will therefore produce biased estimates.

Example 6.2

Suppose that F(y) is a standard logistic distribution, L(y), given by (2.9). Since $1 - L(-y) = L(y)$ and

$$L'(y) = L(y)(1 - L(y)) \tag{6.10}$$

the loglikelihood function in this case is

$$\ell = \sum_{i\in S_1} \left(\log L\left(\frac{Y_i - X_i\beta}{\sigma}\right) + \log\left(1 - L\left(\frac{Y_i - X_i\beta}{\sigma}\right)\right)\right) - N\log\sigma + \sum_{i\in S_0} \log L\left(\frac{-X_i\beta}{\sigma}\right). \tag{6.11}$$

Case II: 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

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

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

$$\ell = \sum_{i\in S_1} \left(\log F'\left(\frac{Y_i - X_i\beta}{\sigma}\right) - \log\left(1 - F\left(\frac{-X_i\beta}{\sigma}\right)\right)\right) - N\log\sigma. \tag{6.12}$$

6.4. 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.

6.4.1. Heckman's method with normally distributed random terms

As above $\Phi(\cdot)$ denotes the cumulative normal distribution function. From (5.2) we get

$$E(Y|Y>0)=X\beta+\sigma E(u|Y>0). \tag{6.13}$$

Since $E(u|Y>0)$ 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

$$\begin{aligned} P(u\in(y,y+dy)|Y>0) &= P\left(u\in(y,y+dy)\middle| u>-\frac{X\beta}{\sigma}\right) \\ &= \frac{P\left(u\in(y,y+dy),u>-\frac{X\beta}{\sigma}\right)}{P\left(u>-\frac{X\beta}{\sigma}\right)} = \frac{P(u\in(y,y+dy))}{P\left(-u<\frac{X\beta}{\sigma}\right)} = \frac{\Phi'(y)dy}{\Phi\left(\frac{X\beta}{\sigma}\right)} \end{aligned} \tag{6.14}$$

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

$$E(u|Y>0)=\frac{1}{\Phi\left(\frac{X\beta}{\sigma}\right)}\int_{-\frac{X\beta}{\sigma}}^{\infty} u\Phi'(u)du. \tag{6.15}$$

But

$$\int_{-\frac{X\beta}{\sigma}}^{\infty} u\Phi'(u)du=\int_{-\frac{X\beta}{\sigma}}^{\infty}\frac{ue^{-\frac{u^2}{2}}}{\sqrt{2\pi}}du=-\int_{-\frac{X\beta}{\sigma}}^{\infty}\frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}}=\frac{1}{\sqrt{2\pi}}\cdot\exp\left(-\left(\frac{X\beta}{\sigma}\right)^2\Big/2\right)=\Phi'\left(\frac{X\beta}{\sigma}\right) \tag{6.16}$$

which together with (6.14) yields

$$E(u|Y>0)=\frac{\Phi'\left(\frac{X\beta}{\sigma}\right)}{\Phi\left(\frac{X\beta}{\sigma}\right)}\equiv\lambda\left(\frac{X\beta}{\sigma}\right) \tag{6.17}$$

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 $i\in S_1$ and zero otherwise. The corresponding loglikelihood equals

$$\ell=\sum_{i\in S_1}\log\Phi\left(\frac{X_i\beta}{\sigma}\right)+\sum_{i\in S_0}\log\left(1-\Phi\left(\frac{X_i\beta}{\sigma}\right)\right). \tag{6.18}$$

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

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

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

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

by applying the observations from S_1. This gives unbiased estimates because it follows from (6.13) and (6.17) that

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

Heckman (1979) has 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 (6.19) and the one that follows by dividing the mean component value of the estimated β by the corresponding mean based on β^*.

6.4.2. Heckman's method with logistically distributed random term

Assume now that u is distributed according to the logistic distribution L(y). Then by Lemma 2 in Appendix A it is proved that

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

In this case the regression model that corresponds to (6.19) equals

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

where

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

and β^* is the first stage maximum likelihood estimate of β/σ based on the binary logit model with loglikelihood equal to (6.18) with $\Phi(y)$ replaced by L(y).

A modified version of Heckman's method

Since

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

it follows from (6.20) that

(6.23)

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

Eq. (6.23) implies that we may alternatively apply regression analysis on the whole sample based on the model

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

where

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

and δ_i is an error term with zero mean. This is so because (6.23) implies that

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

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

6.5. 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. It can be demonstrated that

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

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.

6.6. McFadden's goodness-of-fit measure

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

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

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 (2.38) 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, cf. sections 4.8 and 4.10.

7. Advanced examples of discrete/continuous choice analysis

7.1. Behavior of the firm when technology is a discrete choice variable

Suppose the firm faces the choice of choosing one out of m possible technologies. Let

$$\pi_j = f(p_j, q_j)\exp(\varepsilon_j/\alpha), \tag{7.1}$$

$j = 1,2,\ldots,m$, be the firm's profit conditional on technology j, where p_j is the output price, q_j is a vector of input prices, ε_j is a random term that accounts for unobservable variables that affect production with technology j. We assume that $\{\varepsilon_j\}$ are i.i. standard extreme value distributed and $\alpha > 0$ is a constant. We realize that when α decreases then the effect of unobservable heterogeneity will increase.

By Hotelling's Lemma we obtain that output, Y_j, conditional on technology j, is given by

$$\tilde{y}_j = \frac{\partial f(p_j, q_j)}{\partial p_j}\exp(\varepsilon_j/\alpha) \tag{7.2}$$

and similarly input of type r, conditional on technology j is equal to

$$\tilde{x}_{rj} = -\frac{\partial f(p_j, q_j)}{\partial q_{rj}}\exp(\varepsilon_j/\alpha). \tag{7.3}$$

Let

$$V_j = \alpha \log f(p_j, q_j) + \varepsilon_j . \tag{7.4}$$

It follows from (6.1) and (6.4) that the probability that the firm shall choose technology j equals

$$P_j \equiv P(\pi_j = \max_k \pi_k) = P(V_j = \max_k V_k) = \frac{\exp(\alpha \log f(p_j, q_j))}{\sum_k \exp(\alpha \log f(p_k, q_k))}. \tag{7.5}$$

Recall that by Lemma A2 in Appendix A

$$P(\max_k V_k \le y \mid V_j = \max_k V_k) = P(\max_k V_k \le y). \tag{7.6}$$

Therefore we obtain that

$$E\exp\left(\frac{1}{\alpha}V_j \mid V_j = \max_k V_k\right) = E\exp\left(\frac{1}{\alpha}\max_k V_k\right). \tag{7.7}$$

Moreover,

$$P(\max_k V_k \le y) = \prod_k P(V_k \le y) = \exp(-e^{-y} A) \tag{7.8}$$

where

$$A = \sum_k \exp(\alpha \log f(p_k, q_k)). \tag{7.9}$$

Hence

$$E\exp\left(\frac{1}{\alpha}\max_k V_k\right) = \int_{-\infty}^{\infty} e^{y/\alpha} \cdot \exp(-e^{-y} A) A e^{-y} dy \tag{7.10}$$

which by change of variable, $A e^{-y} = x$, reduces to

$$E\exp\left(\frac{1}{\alpha}\max_k V_k\right) = A^{1/\alpha} \int_0^{\infty} x^{-1/\alpha} e^{-x} dx = A^{1/\alpha} \Gamma\left(1-\frac{1}{\alpha}\right) \tag{7.11}$$

provided $\alpha > 1$. When $\alpha \le 1$ this mean is infinite. From (7.2), (7.7) and (7.11) we get

$$\begin{aligned} E(\tilde{y}_j \mid \pi_j = \max_k \pi_k) &= \frac{\partial f(p_j, q_j)}{f(p_j, q_j)\partial p_j} E\left(\exp\left(\frac{1}{\alpha} V_j\right) \mid V_j = \max_k V_k\right) \\ &= \frac{\partial \log f(p_j, q_j)}{\partial p_j} E\exp\left(\frac{1}{\alpha}\max_k V_k\right) \\ &= \frac{\partial \log f(p_j, q_j)}{\partial p_j} \left[\sum_k \exp(\alpha \log f(p_k, q_k))\right]^{1/\alpha} \Gamma\left(1-\frac{1}{\alpha}\right). \end{aligned} \tag{7.12}$$

Similarly, it follows that

$$E(\tilde{x}_{rj} \mid \pi_j = \max_k \pi_k) = -\frac{\partial \log f(p_j, q_j)}{\partial q_{rj}} \left(\sum_k \exp(\alpha \log f(p_k, q_k))\right)^{1/\alpha} \Gamma\left(1-\frac{1}{\alpha}\right), \tag{7.13}$$

$$E(\pi_j \mid \pi_j = \max_k \pi_k) = E(\max_k \pi_k) = \left[\sum_k \exp(\alpha \log f(p_k, q_k))\right]^{1/\alpha} \Gamma\left(1-\frac{1}{\alpha}\right) \tag{7.14}$$

and

$$E(\log \pi_j \mid \pi_j = \max_k \pi_k) = E(\max_k \log \pi_k) = \frac{1}{\alpha} \log\left[\sum_k \exp(\alpha \log f(p_k, q_k))\right] + \frac{0.5772}{\alpha}. \tag{7.15}$$

From the results above we can deduce an interesting aggregation property. We get from (7.14) that

$$\frac{\partial E(\max_k \pi_k)}{\partial p_j} = \Gamma\left(1-\frac{1}{\alpha}\right)\left[\sum_k \exp\left(\alpha \log f(p_k, q_k)\right)\right]^{1/\alpha - 1} \exp\left(\alpha \log f(p_j, q_j)\right)\frac{\partial \log f(p_j, q_j)}{\partial p_j}$$

(7.16)

$$= \Gamma\left(1-\frac{1}{\alpha}\right)\left[\sum_k \exp\left(\alpha \log f(p_k, q_k)\right)\right]^{1/\alpha} \frac{\partial \log f(p_j, q_j)}{\partial p_j} P_j .$$

But by comparing (7.12) and (7.16) we realize that

$$\frac{\partial E(\max_k \pi_k)}{\partial p_j} = P_j \, E\left(\tilde{y}_j \,\middle|\, \pi_j = \max_k \pi_k\right) = E\,\tilde{y}_j . \tag{7.17}$$

Similarly, it follows readily that

$$\frac{\partial E(\max_k \pi_k)}{\partial q_{rj}} = P_j \, E\left(\tilde{x}_{rj} \,\middle|\, \pi_j = \max_k \pi_k\right) = E\,\tilde{x}_{rj} . \tag{7.18}$$

Finally, it can easily be demonstrated that

$$P_j = \frac{\partial \log E(\max_k \pi_k)}{\partial \log \pi_j} . \tag{7.19}$$

The results above demonstrate that assumptions (7.1) and (7.2) imply that it is possible to define a representative agent with profit function $E(\max_k \pi_k)$, from which one can derive fractional technology choice rates, P_j, and aggregate demands. These are equivalent to the choice probabilities and aggregate demands and production derived from profitmaximizing micro agents.

7.2. Labor supply with taxes (I)

This example is an extension of the example in section 4.1. Consider the choice of "working" versus "not working", and annual hours of work when working. We assume that there is no rationing in the market so that of the agent wishes to work he will be able to get work. Let the agent's utility function in consumption and (normalized) leisure, $L = 1 - h / M$, be given by

$$V(C,L) = \frac{\left(C^{\alpha_1} - 1\right)\beta_1}{\alpha_1} + \frac{\left(\left(1-\frac{h}{M}\right)^{\alpha_2} - 1\right)\beta_2 M^{\alpha_2}}{\alpha_2} \tag{7.20}$$

where $M = 8760$, is total number of hours a year, h is hours of work and $\alpha_1 < 1, \alpha_2 < 1$, $\beta_1 > 0, \beta_2 > 0$. The budget constraint is given by

$$C = hW + I - S(hW, I) \tag{7.21}$$

where W is the wage rate, I is nonlabor income and $S(\cdot)$ is the tax function. There is no fixed cost of working.

The marginal rate of substitution equals

$$\frac{\partial_2 V(C,L)}{\partial_1 V(C,L)} = \frac{\left(1-\frac{h}{M}\right)^{\alpha_2-1}\beta_2}{\beta_1 C^{\alpha_1-1}}. \tag{7.22}$$

Let

$$g(x,y) = x + y - S(x,y). \tag{7.23}$$

Then it follows that the agent wishes to work if

$$W\,\partial_1 g(0,I) \geq \frac{\partial_2 V(g(0,I),1)}{\partial_1 V(g(0,I),1)} = \frac{\beta_2\, g(0,I)^{1-\alpha_1}}{\beta_1}, \tag{7.24}$$

and hours of work, $\tilde{h}$, is determined from

$$W\partial_1 g\left(\tilde{h}W,I\right) = \frac{\partial_2 V\left(g\left(\tilde{h}W,I\right),1-\tilde{h}/M\right)}{\partial_1 V\left(g\left(\tilde{h}W,I\right),1-\tilde{h}/M\right)} = \beta_2\left(1-\frac{\tilde{h}}{M}\right)^{\alpha_2-1} g\left(\tilde{h}W,I\right)^{1-\alpha_1}/\beta_1 \tag{7.25}$$

provided (7.24) holds. The left hand side of (7.24) is called the marginal wage rate at zero hours of work, and the right hand side of (7.24) is called the reservation wage. Assume that β_2/β_1 and W are specified as in (4.7) and (4.8).

Estimation by Heckman's two stage method

From (7.25) we have that hours of work is determined by

$$(\alpha_2-1)\log\left(1-\frac{\tilde{h}}{M}\right) = \log W + \log\partial_1 g\left(\tilde{h}W,I\right) + (\alpha_1-1)\log g\left(\tilde{h}W,I\right) - \log\left(\frac{\beta_2}{\beta_1}\right) \tag{7.26}$$

provided (7.24) holds. Therefore, we face the usual "Tobit problem" that the random term, $\varepsilon_1 - \varepsilon_2$, does not have zero expectation and consequently we cannot apply standard regression analysis. Both $\tilde{h}$ and W are endogenous variables. $\tilde{h}$ is endogenous because it is the hours of work function. Although W is exogenous theoretically it may be endogenous statistically due to unobservables that affect preferences through the hours of work function. If $\log(\beta_2/\beta_1)$ are replaced by (4.7) and we divide both sides of (7.26) by $\alpha_2 - 1$ we obtain

$$(7.27)\quad -\log\left(1-\frac{\tilde{h}}{M}\right)=\max\left(0,-X_2 b r_1+r_1 E\log W+r_1\log\partial_1 g\left(\tilde{h}W,I\right)+r_2\log g\left(\tilde{h}W,I\right)+r_1(\varepsilon_1-\varepsilon_2)\right)$$

where $r_1=1/(1-\alpha_2)$ and $r_2=(\alpha_1-1)/(1-\alpha_2)$, and where $E\log W$ is given by (4.8). Now the labor supply eq. (7.27) is well defined for both working and non-working individuals. However, it is nonlinear in parameters, and there still remains the endogenous variable $\tilde{h}W$ on the right hand side. On the subsample of those who work it is, however, linear, but we cannot apply standard regression analysis because, in addition to the endogeneity problem, the conditional expectation of the error terms given the subsample of workers is not equal to zero. To account for these problems we shall apply Heckman's two stage method. Let

$$(7.28)\quad \begin{aligned}\lambda &\equiv \frac{1}{\tau}E\left(\varepsilon_1-\varepsilon_2\middle|\tilde{h}>0\right)\\ &=\frac{1}{\tau}E\left(\varepsilon_1-\varepsilon_2\middle|-X_2 b r_1+r_1\log W+r_1\log\partial_1 g\left(\tilde{h}W,I\right)+r_2\log g(0,I)+r_1(\varepsilon_1-\varepsilon_2)>0\right)\end{aligned}$$

where

$$\tau^2=r_1^2\,\mathrm{Var}(\varepsilon_2-\varepsilon_1).$$

By applying the result obtained in section 6.4.1, it follows that

$$(7.29)\quad \lambda=\frac{\Phi'\left(\dfrac{Xsr_1+r_1\log\partial_1 g(0,I)+r_2\log g(0,I)}{\tau}\right)}{P_2}$$

where P_2 is the probability of working, and can be written as

$$(7.30)\quad P_2=\Phi\left(\frac{Xsr_1+r_1\log\partial_1 g(0,I)+r_2\log g(0,I)}{\tau}\right),$$

and where $Xs=X_1 a-X_2 b$. Hence, it follows that

$$(7.31)\quad E\left(-\log\left(1-\frac{\tilde{h}}{M}\right)\middle|\,\tilde{h}>0\right)=Xsr_1+r_1\log\partial_1 g\left(W\tilde{h},I\right)+r_2\log g\left(W\tilde{h},I\right)+\tau\lambda$$

which means that we can write

$$(7.32)\quad -\log\left(1-\frac{\tilde{h}}{M}\right)=Xsr_1+r_1\log\partial_1 g\left(W\tilde{h},I\right)+r_2\log g\left(W\tilde{h},I\right)+\tau\lambda+\eta_2$$

where η_2 is a random term with the property that

$$E\left(\eta_2 \middle| \tilde{h} > 0\right) = 0.$$

Similarly, it follows that

(7.33) $$E\left(\log W \middle| \tilde{h} > 0\right) = X_1 a + \rho\tau\lambda$$

where

$$\rho = \text{corr}\left(\varepsilon_1, \varepsilon_1 - \varepsilon_2\right).$$

The relation (7.33) is useful because it enables us to estimate the wage equation from a sample of working individuals, as we shall see in a moment. The term $\rho\tau\lambda$ in (7.33) may be called the "selectivity bias". It is different from zero when $\rho \neq 0$ due to the fact that in this case there is correlation between the random term in the wage equation and the sample selection criteria (namely, $\tilde{h} > 0$). Due to (7.33) we can write

(7.34) $$\log W = X_1 a + \rho\tau\lambda + \eta_1$$

where

$$E\left(\eta_1 \middle| \tilde{h} > 0\right) = 0.$$

If λ were known it would be possible to estimate (7.32) and (7.34) as a simultaneous equation system. Unfortunately, λ is unknown and this is therefore not possible. We can, however, apply the estimates from the probability of working to obtain an estimate of λ.

Step 1

Estimate the parameters of the probit model (7.30) on the basis of discrete observations on whether the agents are working or not working.

Step 2

Estimate the wage equation (7.34) by using $\hat{\lambda}$ as a regressor, where $\hat{\lambda}$ is an estimate of λ obtained from step one.

Step 3

Replace $\log \partial_1 g\left(W\tilde{h}, I\right)$ and $\log g\left(W\tilde{h}, I\right)$ by instrument relations

(7.35) $$\log \partial_1 g\left(W\tilde{h}, I\right) = Z\theta_1 + u_1$$

and

(7.36) $$\log g\left(W\tilde{h}, I\right) = Z\theta_2 + u_2$$

where Z is a set of instrument variables; $Z=(X,I)$, and u_1 and u_2 are zero mean random terms. Estimate (7.35) and (7.36).

Step 4

Insert $\hat{\lambda}$ and the estimated wage equation (without the selectivity term) and the estimated instrument relations (7.35) and (7.65) into (7.32) from which the structural parameters can be estimated.

Estimation by maximum likelihood

Since ε_1 and ε_2 are normally distributed we can write

(7.37) $$\varepsilon_2 - \varepsilon_1 = \theta\varepsilon_1 + \varepsilon_3$$

where ε_3 is a zero mean normal variable that is independent of ε_1 and θ is some constant. Let S_2 be the subsample of individuals that work and S_1 the subsample of individuals that do not work. Let i index individual i. From (7.26), (4.7), (4.8) and (7.37) we have that when $\tilde{h}_i > 0$

(7.38) $$\varepsilon_{3i} = -\theta\varepsilon_{1i} + (1-\alpha_2)\log\left(1-\frac{\tilde{h}_i}{M}\right) + X_{1i}a + \log\partial_1 g\left(\tilde{h}_i W_i, I_i\right) + (\alpha_1 - 1)\log g\left(\tilde{h}_i W_i, I_i\right) - X_{2i}b.$$

Note that we can express ε_{1i} as

(7.39) $$\varepsilon_{1i} = \log W_i - X_{1i}a.$$

Let l_2 be the (conditional) loglikelihood for the subsample of individuals that work. From (7.38) we have

(7.40) $$\frac{\partial\varepsilon_{3i}}{\partial\tilde{h}_i} = \frac{\alpha_2 - 1}{M - \tilde{h}_i} + \frac{W_i\,\partial_1^2 g\left(W_i\tilde{h}_i, I_i\right)}{\partial_1 g\left(W_i\tilde{h}_i, I_i\right)} + \frac{(\alpha_1 - 1)W_i\,\partial_1 g\left(W_i\tilde{h}_i, I_i\right)}{g\left(W_i\tilde{h}_i, I_i\right)}.$$

The loglikelihood for the subsample of those who work becomes

(7.41)

$$\log \ell_2$$

$$= \prod_{i \in S_2} \frac{1}{\sigma_3} \Phi' \left(\frac{\log \partial_1 g\left(\tilde{h}_i W_i, I_i\right) + (\alpha_1 - 1) \log g\left(\tilde{h}_i W_i, I_i\right) - \theta \log W_i + X_{1i} a(\theta+1) - X_{2i} b + (1-\alpha_2) \log\left(1 - \frac{\tilde{h}_i}{M}\right)}{\sigma_3} \right) \left| \frac{\partial \varepsilon_{i3}}{\partial \tilde{h}_i} \right|$$

$$\cdot \frac{1}{W_i} \Phi' \left(\frac{\log W_i - X_{1i} a}{\sigma_1} \right) \frac{1}{\sigma_1}$$

where $\Phi'(\cdot)$ is the standard normal density, $\sigma_1^2 = \operatorname{Var} \varepsilon_{1i}$ and $\sigma_3^2 = \operatorname{Var} \varepsilon_{3i}$.

The likelihood for non-working individuals equals

(7.42) $$\exp \ell_1 = \prod_{i \in S_1} \Phi \left(\frac{\log \partial_1 g(0, I_i) + (\alpha_1 - 1) \log g(0, I_i) + X_i d}{\sigma} \right)$$

where $\sigma^2 = \operatorname{Var}(\varepsilon_2 - \varepsilon_1)$. The total loglikelihood, ℓ, is therefore equal to

$$\ell = \ell_1 + \ell_2.$$

Results from empirical analysis of a sample of married women in Norway, 1979/1980

Dagsvik et al. (1986) analyze female labor supply in Norway based on a sample of married women from the level of living survey/tax return files, 1979/1980, by applying the model discussed above. The variables that affect the women's preferences are specified to be "Age", "Age squared", "Number of children below six years of age", "Number of children above six years", a disability dummy and an index of job opportunities for women.

The variables that affect the wage quation are assumed to be "Age", "Age squared" and "Years of education".

The estimates obtained by the four step procedure are displayed in Tables 7.1 and 7.2 below.

Table 7.1. Estimates of the parameter in the utility function

Independent variables	Estimate	Standard deviation
Intercept	-5.35	0.80
age	0.158	0.03
$10^{-2} \times$ age squared	-0.205	0.03
Number of children less than six years	-0.289	0.07
Number of children above six years	-0.079	0.04
Disability index	-0.398	0.09
Index of job-opportunities	0.727	0.59
α_1 (Consumption)	1.0	
α_2 (Leisure)	-4.28	0.11
Marginal wage ($1/\sigma$)	0.965	0.13

Table 7.2. Estimates of the wage equation

Independent variables	Estimate	Standard deviation
Intercept	2.161	0.28
Years of education	0.065	0.01
Age	0.030	0.01
$10^{-2}. \times$ age squared	-0.032	0.01
Selectivity, $\hat{\lambda}$	-0.105	0.06
R^2	0.16	

7.3. Labor supply with taxes (II)

We will now consider the case where ε_1 and ε_2 are jointly extreme value distributed. Dagsvik et al. (1988) have analyzed female labor supply in France based on the model formulation above, but where $(\varepsilon_1, \varepsilon_2)$ are bivariate extreme value distributed instead of bivariate normal. Thus,

$$P(\varepsilon_1 \le y_1, \varepsilon_2 \le y_2) = \exp\left(-\left(e^{-y_1/\rho\sigma} + e^{-y_2/\rho\sigma}\right)^{\rho}\right) \tag{7.43}$$

where $\rho, 0 < \rho \le 1$, is related to the correlation coefficient by

$$\operatorname{corr}(\varepsilon_1, \varepsilon_2) = 1 - \rho^2 \tag{7.44}$$

and

$$\frac{\pi^2\sigma^2}{6} = \operatorname{Var}\varepsilon_1 = \operatorname{Var}\varepsilon_2 . \tag{7.45}$$

Moreover, it follows that

$$\tau^2 \equiv \mathrm{Var}(\varepsilon_1 - \varepsilon_2) = \frac{\pi^2}{6}\sigma^2\rho^2. \tag{7.46}$$

Since ε_1 and ε_2 are jointly extreme value distributed we get by Theorem 7 that

$$\begin{aligned} P(\varepsilon_1 < \varepsilon_2 + y) &= P\left(\frac{\varepsilon_1}{\sigma} < \frac{\varepsilon_2}{\sigma} + \frac{y}{\sigma}\right) \\ &= \frac{\exp(y/\rho\sigma)}{1+\exp(y/\rho\sigma)} = \frac{1}{1+\exp(-y/\sigma\rho)}, \end{aligned} \tag{7.47}$$

which means that $\varepsilon_1 - \varepsilon_2$ has a logistic distribution. From (7.47) and (7.27) we get

$$P(\tilde{h} > 0) = \frac{1}{1+\exp\left(-(X s r_1 + r_1 \log \partial_1 g(0,I) + r_2 \log g(0,I))/\rho\sigma\right)}. \tag{7.48}$$

From Lemma A2 in Appendix A we get

$$\tilde{\lambda} \equiv \frac{1}{\tau}E(\varepsilon_1 - \varepsilon_2 \mid \tilde{h} > 0) = -\frac{\log(1 - P(\tilde{h} > 0))}{P(\tilde{h} > 0)} - \frac{1}{\tau}\left(X s r_1 + r_1 \log \partial_1 g(W\tilde{h}, I) + r_2 \log g(W\tilde{h}, I)\right). \tag{7.49}$$

From (7.32), (7.48) and (7.49) we thus obtain

$$-\log\left(1 - \frac{\tilde{h}}{M}\right) = X s r_1 + r_1 \log \partial_1 g(W\tilde{h}, I) + r_2 \log g(W\tilde{h}, I) + \tau\tilde{\lambda} + \tilde{\eta}_2 \tag{7.50}$$

where $\tilde{\eta}_2$ is a random term such that $E(\tilde{\eta}_2 \mid \tilde{h} > 0) = 0$. Similarly, it can be proved that

$$\log W = X_1 a - \rho\sigma \log P(\tilde{h} > 0) + \tilde{\eta}_1 \tag{7.51}$$

where $\tilde{\eta}_1$ is a random term such that $E(\tilde{\eta}_1 \mid \tilde{h} > 0) = 0$.

It is now clear that the model specified above can be estimated in the same way as the model specification in Section 7.2.

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 $X_1, X_2, \ldots,$ are 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) $$\lim_{n \to \infty} P\left(\frac{M_n}{(cn)^{1/\alpha}} \le x \right) = \begin{cases} exp(-x^{-\alpha}) & for \quad x > 0, \\ 0 & for \quad x \le 0. \end{cases}$$

Theorem A2

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

Theorem A3

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

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

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. Tippet. 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.

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.,

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

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

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

where

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

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 7 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 $b + 0.5772....$, and variance equal to $a^2\pi^2/6$.

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.

Proof:

We have

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

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

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

Q.E.D.

Lemma A2

Suppose $U_j = v_j + \varepsilon_j$, where $(\varepsilon_1, \varepsilon_2, ..., \varepsilon_m)$ 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

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

where G(·) is homogeneous of degree one. For notational simplicity let $j=1$, since the general case is completely analogous. 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

(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

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

But the last factor in (A.12) equals $P\left(\max_k U_k \le y\right)$, as is easily seen from (A.9) and (A.10). Moreover, by Theorem 7 the first factor on the right hand side of (A.12) equals $P\left(U_1 = \max_k U_k\right)$. Thus 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

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

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

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

Proof:

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

(A.15)
$$P(u>y|Y>0) = \frac{P\left(u>y, u>-\frac{\mu}{\sigma}\right)}{P\left(u>-\frac{\mu}{\sigma}\right)}$$
$$= \frac{P(-u<-y)}{P\left(-u<\frac{\mu}{\sigma}\right)} = \frac{P(u<-y)}{P\left(u<\frac{\mu}{\sigma}\right)} = \frac{1+\exp\left(-\frac{\mu}{\sigma}\right)}{1+\exp(y)}$$

which proves (A.13).

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

(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

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

and (A.14) has thus been proved.

Q.E.D.

The Tax function applied in Dagsvik et al. (1986)

Let

$$\psi(x)=\begin{cases}0.053x, & x\in[0,3000]\\ 3.38\cdot 10^{-4}\,(x-3000), & x\in[3000,49826]\\ 3.38\cdot 10^{-4}\,(0.81x+6467)^{1.61}+0.053x, & x\in[49826,23700]\\ -27472+0.651x, & x\in[237000,\infty).\end{cases}$$

Then the tax function is given by

$$T(hw,I)=\psi(hw+I),$$

when hw or I are less than NOK 22 000, and

$$T(hw,I)=\psi(hw)+\psi(I)$$

otherwise.

References

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

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.

Bjerkholt, O. (1995): Introduction: Ragnar Frisch, the Originator of Econometrics. In O. Bjerkholt (ed.): *Foundations of Modern Econometrics. The Selected Essays of Ragnar Frisch*, Vol. I. E. Elgar, Aldershot, UK.

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.

Dagsvik, J.K. (1985): Kvalitativ valghandlingsteori, en oversikt over feltet. *Sosialøkonomen*, no. 2, 1985, 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., F. Laisney, S. Strøm and J. Østervold (1988): Female Labour Supply and the Tax Benefit System in France. *Annales d'Économie et de Statistique*, **11**, 5-40.

Dagsvik, J.K., O. Ljones, S. Strøm and Rolf Aaberge (1986): Gifte kvinners arbeidstilbud, skatter og fordelingsvirkninger. Rapporter 86/14, Statistics Norway.

Dagsvik, J.K., D.G. Wetterwald and R. Aaberge (1996): Potential Demand for Alternative Fuel Vehicles. *Discussion Papers*, no. 165, Statistics Norway.

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.

Frisch, R. (1926): Sur un problème d'économie pure. English translation in O. Bjerkholt (ed.): *Foundation of Modern Econometrics. The Selected Essays of Ragnar Frisch*, 1995, Vol. I. E. Elgar, Aldershot, UK.

Gorman, W.M. (1953): Community Preference Fields. *Econometrica*, **21**, 63-80.

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.

Heckman, J.J. (1974): Shadow Prices, Market Wages, and Labor Supply. *Econometrica*, **42**, 679-694.

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

King, M. (1980): An Econometric Model of Tenure Choice and Demand for Housing as a Joint Decision. *Journal of Public Economics*, **14**, 137-159.

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

Lee, L.F., and R.P. Trost (1978): Estimation of Some Limited Dependent Variable Models with Application to Housing Demand. *Journal of Econometrics*, **8**, 357-382.

Lindquist, K.-G. (1992): Økonometriske modeller for kvalitative valg og trunkerte endogene variable. Arbeidsnotat no. 128/1992, Center for Research in Economics and Business Administration.

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.

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.

Robertson, C.A. and D. Strauss (1981): A Characterization Theorem for Random Utility Variables. *Journal of Mathematical Psychology*, **23**, 184-189.

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.

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.

Recent publications in the series Documents

96/25 T. Bye and S. Kverndokk: Nordic Negotiations on CO_2 Emissions Reduction. The Norwegian Negotiation Team's Considerations

96/26 L. Rogstad and M. Dysterud: Land Use Statistics for Urban Agglomerations. Development of a Method Based on the Use of Geographical Information Systems (GIS) and Administrative Records

96/27 K. Rypdal: NOSE – Nomenclature for Sources of Emissions

97/1 T.C. Mykkelbost and K. Rypdal: Material Flow Analysis of Cadmium and Di-2-etylhexylphthalate (DEHP) in Norway

97/2 S. Grepperud: The Impact of Policy on Farm Conservation Incentives in Developing countries: What can be Learned from Theory

97/3 M. Rolland: Military Expenditure in Norway's Main Partner Countries for Development Assistance. Revised and Expanded Version

97/4 N. Keilman: The Accuracy of the United Nation's World Population Projections

97/5 H.V. Sæbø: Managerial Issues of Information Technology in Statistics Norway

97/6 E.J. Fløttum, F. Foyn, T.J. Klette, P.Ø. Kolbjørnsen, S. Longva and J.E. Lystad: What Do the Statisticians Know about the Information Society and the Emerging User Needs for New Statistics?

97/7 A. Bråten: Technical Assistance on the Jordanian Consumer Price Index

97/8 H. Brunborg and E. Aurbakken: Evaluation of Systems for Registration and Identification of Persons in Mozambique

97/9 H. Berby and Y. Bergstrøm: Development of a Demonstration Data Base for Business Register Management. An Example of a Statistical Business Register According to the Regulation and Recommendations of the European Union

97/10 E. Holmøy: Is there Something Rotten in this State of Benchmark? A Note on the Ability of Numerical Models to Capture Welfare Effects due to Existing Tax Wedges

97/11 S. Blom: Residential Consentration among Immigrants in Oslo

97/12 Ø. Hagen and H.K. Østereng: Inter-Baltic Working Group Meeting in Bodø 3-6 August 1997 Foreign Trade Statistics

97/13 B. Bye and E. Holmøy: Household Behaviour in the MSG-6 Model

97/14 E. Berg, E. Canon and Y. Smeers: Modelling Strategic Investment in the European Natural Gas Market

97/15 A. Bråten: Data Editing with Artificial Neural Networks

98/1 A. Laihonen, I. Thomsen, E. Vassenden and B. Laberg: Final Report from the Development Project in the EEA: Reducing Costs of Censuses through use of Administrative Registers

98/2 F. Brunvoll: A Review of the Report "Environment Statistics in China"

98/3: S. Holtskog: Residential Consumption of Bioenergy in China. A Literature Study

98/4 B.K. Wold: Supply Response in a Gender-Perspective, The Case of Structural Adjustments in Zambia. Tecnical Appendices

98/5 J. Epland: Towards a register-based income statistics. The construction of the Norwegian Income Register

98/6 R. Chodhury: The Selection Model of Saudi Arabia. Revised Version 1998

98/7 A.B. Dahle, J. Thomasen and H.K. Østereng (eds.): The Mirror Statistics Exercise between the Nordic Countries 1995

98/8 H. Berby: A Demonstration Data Base for Business Register Management. A data base covering Statistical Units according to the Regulation of the European Union and Units of Administrative Registers

98/9 R. Kjeldstad: Single Parents in the Norwegian Labour Market. A changing Scene?

98/10 H. Brüngger and S. Longva: International Principles Governing Official Statistics at the National Level: are they Relevant for the Statistical Work of International Organisations as well?

98/11 H.V. Sæbø and S. Longva: Guidelines for Statistical Metadata on the Internet

98/12 M. Rønsen: Fertility and Public Policies - Evidence from Norway and Finland

98/13 A. Bråten and T. L. Andersen: The Consumer Price Index of Mozambique. An analysis of current methodology – proposals for a new one. A short-term mission 16 April - 7 May 1998

98/14 Sigurd Holtskog: Energy Use and Emmissions to Air in China: A Comparative Literature Study

98/15 J.K. Dagsvik: Probabilistic Models for Qualitative Choice Behavior: An introduction

Documents

PORTO BETALT
VED
INNLEVERINGA
P.P.
NORGE/NOREG

Tillatelse nr.
159 000/502

B

Returadresse:
Statistisk sentralbyrå
Postboks 8131 Dep.
N-0033 Oslo

Statistics Norway
P.O.B. 8131 Dep.
N-0033 Oslo

Tel: +47-22 86 45 00
Fax: +47-22 86 49 73

ISSN 0805-9411

Statistisk sentralbyrå
Statistics Norway