CONSTRUCTION OF MULTIVARIATE BINOMIAL CHOICE MODEL

A Thesis submitted in Partial Fulfillment of the Requirements for the Degree of M. Phil in Statistics

By
MD. SHAFIQUL ISLAM
REGISTRATION NO :-147
SESSION : 1998-1999

DEPARTMENT OF STATISTICS
UNIVERSITY OF DHAKA
DHAKA-1000, BANGLADESH
DECEMBER- 2007

Supervisor's Declaration

This is to certify that the disertation titled "Construction of Multivariate Binomial Choice Model" submitted by Md. Shafiqul Islam in partial fulfillment for the degree of M.Phil in the Department of Statistics, University of Dhaka is an original study carried out by him under my supervision and guidance.

Supervision
Professor Dr. Matiur Rahman
Department of Statistics
University of Dhaka
Dhaka-1000

Prof. Dr. Matiur Rahman
Department of Statistics
University of Dhaka
Dhaka, Bangladesh.

DEDICATED TO

MY PARENTS

ACKNOWLEDGEMENT

With deep sense of gratitude I acknowledge the invaluable guidance of my thesis supervisor Professor. Dr. M. Matiur Rahman, Department of Statistics, University of Dhaka, Dhaka-1000. In spite of his extremely busy schedule, he could find time to provide precious guidance and very generous support for preparing every material needed in completing the thesis work.

I express my sincere gratitude to All of the teachers, Dept. of Statistics, University of Dhaka, Dhaka-1000. They gave me valuable suggestion and inspired me for my thesis.

I acknowledge the support received from the BANSDOC Library, BCSIR, BDIIRM, Mohakhali, Dhaka to collect the important information necessary for preparing this thesis.

I also acknowledge the support received from British Council Library, Dhaka University, Dhaka University Library, ISRT Library, BANSDOC, ICDDRB Library, Mohakhali, Dhaka, Bangladesh Rice Research Institute, Gazipur, Bangladesh Bureau of Statistics, Dhaka.

I am very much indebted to all who have helped me in various ways to complete my thesis.

TABLE OF CONTENTS

ABBREVIATION

BCSIR	Bangladesh Council of Scientific and Industrial Research
ISRT	Institute of statistical reachrs and technology
ICDDRB	International Centre for Diarrhoea Disease Research of Bangladesh
WLS	Weighted Least Square
GEV	General form of extreme value distribution for e's.
NMNL	Nested Multinomial Logit Model
MNL	Multinomial Logit Model
IIA	Independence of Irrelevant Alternative
MNP	Multinomial Probit
MSM	Method of Simulated Moment
EBA	Elimination by Aspects
EBT	Elimination Tree
HEBA	Hierarchical Elimination by Aspects
MLE	Maximum Likelihood Estimation
WLSE	Weighted Least Square Estimation
ML	Maximum Likelihood
LR	Likelihood Ratio
SSR	Sum of squares of residuals
AIC	Akaike Information Criterion
LRT	Likelihood Ratio test
HM	Hausman and McFaddan

LM	Lagrange Multiplier
LDV	Lag Dependent Variable
IM	Information Matrix
MVB	Multivariate Binomial
W	Wald
MN	Multinomial
SSR	Sum Squares of Residuals
ARI	Appropriate Ranges of Integration
ME	Meat Expenditure
BEV	Beef and Veal
MUL	Mutton and Lamb
T	Take Way,
PIGGR	Pork, Ham and Bacon
PICS	Processed Meat
POUL	Poultry
EISGR	All Other

ABSTRACT

In this thesis our main interest is in constructing a multivariate binomial choice model. In doing so we attempt to circumvent the shortcomings with the traditional argument that the multivariate choice problem can be converted to multinomial one and then to apply all statistical properties of multinomial analysis.

In this thesis we have attempted to demonstrate the construction of a multivariate binomial choice model keeping in view that although univariate choice model is frequently observed, multivariate framework is rare. It may be noted that in practical situation interrelations in choice decisions on marginal utilities often occur. In consumption behavior consumers hold inherent interrelated choice decisions, which is quite natural. But, model exhibition of such behavior is not frequently found. Such modeling is quite a hard task not only from the viewpoint of construction but also from the viewpoint of estimation and hypothesis testing. We present discussion on how such a model differs from multinomial one. We also discuss issues of estimation and tests. We believe that application of such a model in practical situation may definitely provide scopes for more appropriate analysis of decision-making processes. Such a modeling strategy also provokes for seeking and establishing inter-relationship among ingredients of choice decisions.

Micro-economic theory provides a way of specifying and interpreting dis-aggregate models up to a degree which is not possible with aggregate models. Moreover, survey data on the individual decision maker which can provide more precise information and estimation than that obtained from aggregate data, is more readily available now a days than that it was before. As noted by Heckman (1976) "Interest in discrete data has been stimulated by a growing interest in micro-economic problems and growing availability of good micro-economic data." With data on individual decision-maker more precise estimates of the underlying parameters of the model are possible. This fact is more important in estimating econometric models since precision of parameter estimates varies with variance of variables entering a model and decreases with co-variance among variables. Dis-aggregate models are capable of capturing effects that can not be incorporated in aggregate models.

Furthermore, it is likely that the factors influencing a particular choice outcome differ in value and range across the population such that any basis for market segmentations must emanate from detailed knowledge at the level of the decision making unit. It is also plausible that consumers exhibit portfolio behavior, which means that consumers make a combination of one or more consumption items disregarding their interdependencies. An important aspect of any behavioral statistical estimation is the treatment of unobserved determinants of individual decision. It is quite clear that under normality assumption for the error term, we have a very complex mathematical expression and computationally it is very intricate for explicit expression of the parameter estimates.

However, although elegant in formulation as well as from practical viewpoint, choice models pose severe estimation problems in addition to complex testing issues. There have been various attempts to circumvent such problematic issues. In this thesis we have attempted to provide a picturesque idea on such attempts along with their comparative assessment.

In line with the foregoing assertion it is believed that such a modeling strategy will open up frontiers for applied researchers to focuses and analyse practical situations more elegantly and prominently. A data set can be analysed from various viewpoints. Thus, it is always imperative to look for appropriateness of the procedure followed. Hence, it is argued that the present modeling framework bears tremendous importance from practical viewpoint.

CHAPTER I
INTRODUCTION

1.1 MOTIVATION

In this chapter we principally throw light on the rationale behind discrete choice models along with areas where such models are mainly used.

In this chapter we deal with the following aspect:

a) Importance of discrete choice models

b) Out line of the thesis

1.2 IMPORTANCE OF DISCRETE CHOICE MODELS

In the recent past of econometric research and analysis the emphasis on dis-aggregate modeling at the level of the individual decision making unit, rather than at the aggregate level, has become more and more prominent. Microeconomic theory provides a way of specifying and interpreting dis-aggregate models up to a degree, which is not possible with aggregate models. Moreover, survey data on the individual decision maker, which can provide more precise information and estimation than that obtained from aggregate data, is more readily available now a days than it was before. As noted by Heckman (1976), "Interest in discrete data has been stimulated by a growing interest in microeconomic problems and a growing availability of good microeconomic data. As economists attempt to make greater use of their theory to solve practical problems such as estimating the demand for new modes of travel and ascertaining the determinants of the labor supply of women, the analytical fiction of the representative consumer and its econometric analogue-the classical regression model-have become less useful. Increasingly, economists have begun to recognize that the analysis of choices at the extensive margin (i.e., discrete choices) are just as interesting and often of great empirical importance as the analysis of choices at the intensive margin that is treated in traditional analysis."

Selection of the individual as the unit of analysis provides a more general basis for developing and implementing policy planning. An important aspect of all planning is to predict the demand for a particular service. It is a recognized fact that the ultimate aim of any policy planning is to determine the aggregate demand, which is the observed aggregate behavior of individual decision-making by many individuals. We can very rationally assume the heterogeneity in tastes and individual behavior and thus an individual-based study appears to be the most appropriate starting point for studying the consumer and producer behavior. Hence, the modeling of individual behavior can be viewed either explicitly or implicitly at the root of all modeling for aggregate behavior. The principal objective of choice modeling is to make aggregate extrapolative and conditional forecasting for policy makers to assess the market impact of proposed and implemented policies. The predictive outcome obtained by the individual dis-aggregate approach is more accurate and definitive than those obtained by the aggregate approach because we can gain a clearer insight of the circumstances when we take the individual variation into account. With data on individual decision-maker more precise estimates of the underlying parameters of the model are possible. This fact is more important in estimating econometric models since precision of parameter estimates varies with variance of variables entering a model and decreases with covariance among variables. Dis-aggregate models are capable of capturing effects that can not be incorporated in aggregate models.

Furthermore, it is likely that the factors influencing a particular choice outcome will differ in value and range across the population such that any basis for market segmentation must emanate from a detailed knowledge at the level of the decision-making unit. These premises are part of the rationale that argues for the relevance of studying the individual's behavior in the prediction of market demand. That is, the true measurement of the variation in the explanatory variables is a necessary condition for determining the influence of those variables on the behavior of concern. At this level of data analysis, we

have maximum flexibility in the way we combine model output in aggregate forecasting (see Hensher and Johnson, 1981).

In addition to the predictive output of a study of individual choice, we may very often be able to obtain an estimate of the utility derived by an individual which provides a source of information on the benefits associated with a specific project. A realistic study of demand must emphasis the relationship between utility and service, and this is only satisfactorily accommodated in a theoretical and empirical framework, which adopts the individual as the unit of analysis. This is the central theme of modeling the choice behavior.

Since most of the discrete data models, including discrete choice models, used in econometrics can be accommodated within an extended or generalised linear model framework, many of the insights of the regression models carry over to such models. McCullagh and Nelder (1983) have shown that the class of generalised linear models include as special cases linear regression, logit and probit models for quantal responses, log –linear and multinomial response models for count data. Gilbert (1979); Hausman, Hall and Grilliches (1984) have pointed that there is much scope for the application of discrete data models in econometric research.

The origin of many of the discrete choice models in use of econometrics is in Biometrics. However, the economic applications of such models often involve relationships amongst variables that are more complex than those found in biological applications. It is worth mentioning that economic qualitative choice models differ from those of biometrics or other disciplines in their theoretical foundation although they may be alike in their mathematical forms. The range of usefulness of discrete choice modeling depends on the range of their meaningful applications.

The majority of the applications of choice modeling in the past have been research oriented and often as a case study for demonstrating the underlying concept. In many situations, particularly in business and economics, many important problems prompt us to apply discrete choice models to predict

aggregate behavior. Examples of applications of such discrete response models include

1. Labor force participation: Parson (1980)
2. Job location: Duncan (1980)
3. Number of patients per period: Hausman, Hall & Grilliches (1984)
4. Migration: Goldberg and Nold (1980)
5. Transpiration: Hausman and Wise (1978), Train (1986)
6. Housing: Rosen and Rosen (1980)
7. Coffee purchasing pattern: Guadangi and Little (1983)
8. Hospital utilisation: Lee and Lohen (1985)
9. Labor force Participation: Winkkelmann (2000)
10. Choice of Jobs: Greene (2002)

One of the most important developments in the recent applied economics is the availability of large samples of individual-based data (see Amemiya, 1981, Akin and Guilky, 1979). Some economists (Hicks, 1956) hold the view that microeconomic theory has greater relevance for aggregate data, arguing on largely intuitive grounds that the variations in circumstances of individual entity average out to negligible proportions in aggregate. Many public sector decisions should be considered well informed only if the knowledge of the determinants of discrete choices by individuals are available. Moreover, the knowledge of the determinants of individuals decisions is necessary for policy makers in designing plans like public transport program, medical care facilities, child care facilities, education program, income maintenance program, workforce formation program. Planning can be formulated at the both micro and macro levels. Commodity-wise or commodity group-wise policy analysis can render enormous guidance to policy makers and individual-based modeling for such purpose is still more useful.

Changes in the structure of total demand have important implications for prices and returns to different industries and analysis of these changes are also important from a policy viewpoint regarding production and possible

developments within the industry and short and long term forecasts of such developments may be more fruitful using individual-based analysis.

The critical discrimination in the cause and effect relationships of the consumption pattern of different consumption items may increase the efficiency of within industry activities. The planners need to forecast accurately the response of consumption demand with changes in various associated factors. This, in turn, necessitates to investigate behavioral models of individual consumption pattern and thus, we need to define behavioral models which represents the decisions that consumers make when confronted with alternative choices and the parameters of a true behavioral model reflect the motivations of individual choice makers. Although we are concerned with aggregate behavior of the community as a whole it is best understood when we consider individual behavior. Aggregate pattern of demand may be influenced by changes in underlying preferences of individual consumer, changes in the composition of population, changes in health.

A distinction exists between demand and Choice. Choice refers to a decision of the type "what" or "whether" while demand refers to "how much" which is measurable in continuum. The traditional demand analysis is based on some very strict assumptions and variation in those assumptions gives rise to diverse types of demand functions which most often become in operational in practice (see Hansen, 1974). In traditional demand analysis price changes impart an observable impact on the quantity demanded. But, in practice, there are many commodities for which price changes bear no effect at all. For example, religion versus alcohol consumption. Thus, it seems that traditional continuous demand analysis is not a solid step to analyse consumer real choice behavior.

Instead of considering the structure of demand a more realistic attempt may be to investigate how those factors mentioned above as well as other socioeconomic-demographic factors influence the individual probabilistic choice behavior. In this type of modeling we do not have to put any restriction on zero or nonzero consumption as is the case with traditional demand analysis. We

should really formulate a model such that the effects of individual differences in tastes and optimizing behavior on the random error structure become more explicit. The probability has a specific range and thus, the higher the probability of consuming some item the more is the consumer's desire for that item and thus it is more representative demand. The aim of a modeling should be to estimate the effects of changes in the explanatory variables on the likeliness or probability of choice and then to use the estimated model as a predictive and analytical tool to represent the behavior rules relevant for the sampled population.

Several authors analysed choice behavior as probabilistic as opposed to deterministic phenomena on the ground that consumers' preferences are uncertain and are subject to random shocks (see Brown and Deaton, 1972). Probabilistic choice models based on threshold approach, random utility indicators and optimizing behavior are much more relevant to many situations. By adopting probabilistic choice models, unlike the traditional consumer's demand modeling in economics, we attempt to introduce a more realistic and practical assumptions on choice behavior. Admitting the fact that an analyst is incapable of fully representing all dimensions of determinants of preferences, the aim of a probabilistic choice model is to identify the significance of the deterministic components of a utility function.

Individual based choice models can provide very powerful mechanism for assessing the effectiveness of wide range of policies and as policies impart impact on individuals with varying degrees of force, it is important that the policy makers be able to determine the individual specific impact prior to the determination of the aggregate share effects. If the model structuring is sound enough as regards to the representative utility condition, then one can have a very flexible policy-sensitive analytical tool (See Hensher and Johnson, 1981). The measures of direct and cross elasticity obtained from choice models can provide guidance on the appropriateness of various policies taken by the policy makers. It has been shown that individual choice models are also empirically

and theoretically sound mechanisms for obtaining behavioral shadow prices for non-cost variables (see Hensher, 1981).

1.3 OUTLINE OF THE THESIS

In this section we give an overview of the entire work of our thesis. In chapter II we discuss the historical development of choice models along with their advantages and disadvantages in terms of specification estimation, testing and applications. In this chapter we also discuss the position and the importance of multivariate choice models. In chapter III we present the estimation and testing problems associated with choice models. In this chapter we discuss motivation, estimate the parameter and testing for different variables. In Chapter IV we focus on place of multivariate choice models. In chapter V we discuss the construction of a multivariate binomial choice model. In chapter VI we present some concluding research. We also discuss the link between the multivariate and multinomial situations and deal with issues like estimation, testing of different models.

CHAPTER II

CHOICE MODELS IN HISTORICAL PERSPECTIVE

II.1 INTRODUCTORY REMARKS

Basing on the discussion of chapter I we deem it appropriate to conclude that individual-based probabilistic choice analysis for studying consumption pattern rather than traditional continuous demand analysis is more appropriate. This consideration in turn prompts us to focus on previous various types of available choice models along with their associated properties and the present chapter is basically meant for this purpose. First of all, we present the basic notion behind considering choice model along with model structure. Thirdly, we present some literature review on various approaches to deal with discrete data . Secondly, we present a lucid idea on multinomial choice model.

II.2 BASIC NOTION BEHIND CHOICE MODELLING

Specification of choice model is related to probabilistic choice theory, which is used to explain behavioral inconsistencies. Probabilistic mechanism can be used to capture the effects of alternatives in the choice set. The whole idea of choice modeling is to predict the probability that any choice maker with given characteristics will make a particular choice from a given choice set or the same individual will choose a new alternative with a given set of choice characteristics. More generally, we attempt to find a relationship between a set of attributes and the probability of making a particular choice. For modeling probabilistic choice behavior we deal with qualitative endogenous variables which may result from economic behavior, which is intrinsically categorical, or from the classification during observation. Qualitative variables may be binomial or multinomial and multinomial responses can be ordered, unordered or

sequential. By multinomial we mean the choice of one alternative from a set of, say, m, alternatives. When m=2, we have the dichotomous/binary/binomial choice situation. Choice situation can also be multivariate which is concerned with the choice of any number of alternatives, one to m, from the choice set comprising of m alternatives. There is actually similarity in the study of multinomial or multivariate choice problems and this can be demonstrated with the help of quantitative representation of qualitative variables (see Gourieroux, 1984, McFadden, 1985). For example, the multivariate binomial choice may be represented by

$y=(y_1, y_2, \ldots\ldots\ldots\ldots y_m)$ where $\begin{cases} y_k = 1 & \text{if alternative k is chosen} \\ =0 & \text{otherwise ;} \end{cases}$

$(k=1,2,\ldots\ldots\ldots m)$

Here, $0 \leq 0 \leq \sum_k y_k \leq m$ whereas in multinomial model it is $0 \leq \sum_k y_k \leq 1$

Let us represent this idea in a schematic way for a choice set C containing m alternatives as follows:

		Alternatives			
		C_1,	C_2,	C_3 .....C_m	
	v_1:	1,	1,	1, ...1	$\sum y_i = m$
Vectors	v_2:	0,	1,	1, ... 1	$\sum y_i = m-1$
	.....				
	v_{2m} :	0,	0.	0, ...0	$\sum y_i = 0$

Since each of c's can occur in two ways so we have 2m vectors and an individual will choose one of them and thus the multivariate binomial choice problem can be converted to a multinomial one. Hence, any arbitrary

multivariate model can be converted to a multinomial one but there exist some differences between the two which we shall discuss in chapter 4.

A multinomial choice model estimates the probability of making a particular choice and is characterised by the following:

(1) there is a finite set of alternatives,

(2) the alternatives are mutually exclusive

(3) the choice set is exhaustive.

(4) the individual decision maker chooses one and only one alternative from the choice set.

It is worth mentioning that choice situations may not be always qualitative, and alternatives may not be finite, exhaustive or mutually exclusive. Train (1982) has clarified this situation by considering example of the choice of insurance policies offered by a salesperson. In this case the set is not exhaustive since the decision-maker could choose a policy offered by another salesperson. But, a non-qualitative choice situation may be redefined to meet all the criteria which qualitative choice models meet. If we consider the above example, the decision maker's choice between two insurance policies offered by a particular salesman does not show the exhaustive set of three alternatives comprising of either of the two policies or neither and thus the situation is a qualitative one. Qualitative choice models designate a class of models of which probit and logist modes are the most commonly used members. All the choice model have difference in their functional forms for estimating the probability that a decision maker unit chooses a particular alternative. When considering modeling we need to pay attention to:

1. How to specify a theory consistent and statistically manageable model
2. How to estimate and test the parameters of the model and
3. How to select among competing models.

The general form of the qualitative choice models can be represented as

$P_{in} = f(x_{in}, s_n, \beta), \quad i \neq j,$

Where,

P_{in}=the probability that the individual n will choose the alternative i from the choice set C

x_{in} = the attributes of the alternative i faced by n.

s_n = characteristics of the individual n.

β = parameter vector having particular meaning in a particular situation and they are

assumed to be constant across choice makers but may vary across alternatives.

f = the function which relates the probabilities to the observed data.

From this general framework, specific choice models can be obtained by specifying f. The most suitable model for quantal response data will depend on the nature of the behavior of factors governing the response and on the objectives of the analysis. Model building can be "direct" statistical" (picking up a specific form for the model), "Latent variable representation" (mapping from unobservable to observable variable) and based on "economics". Determination of P_{in} can be facilitated by considering the microeconomic theory on optimising agents or of utility or profit maximisation which refers to 'perceived maximization'.

We relevantly mention that a number of different model structures exists. The simplest one is the independent structure in which the attributes defining probabilities are independent across choices. But, when the reverse is true (attributes are not independent across choices) a conditional setting is appropriate viz, sequential and recursive sequential. In sequential there is no feedback loops or mutual interactions among choices and sequential choices are made conditioned on previous ones. In recursive sequential feedback is incorporated through an inclusive value and at each stage the decision is made conditioned on previous choice in totality rather than on a single previous decision. In some situations we may also consider the block recursive sequential decisions. The most general structure is the simultaneous decision in

which no specific assumption about independence or hierarchical structure of decisions are made (see Hansher and Johnson, 1981 for a finer details).

A dichotomy exists between fundamentally probabilistic behavior and what appears to us to be probabilistic because of our shortcomings to completely understand and measure the relevant factors affecting the choice behavior (see Ben-Akiva, 1985). Thus. Luce and Suppes (1965) distinguished between constant and random utility. A constant utility approach assumes a decision maker to behave with choice probabilities defined by probability distribution over alternatives that use utility as parameters. The simplest constant utility model was derived by Luce (1959) as $P(I/C)=P(I/\overline{C}).p(\overline{C}/C)$, ie $\overline{C} \leq C$ which states that the probability of 1 from C is the product probability of i from subset $\overline{C}$ and the probability that the choice lies in $\overline{C}$.

A rational decision maker maximises his utility from an alternative. It is worth mentioning that rationality in choice behavior can be 'perfect' and 'bounded' as distinguished by Simon (1957). The notion of rationality can be ambiguous less it is defined by a specific set of rules in a particular situation. The rationality we deal with in our work refers to transitive preference which means that if choice C_1 is preferred to C_2 and C_2 is preferred to C_3, then C_1 is preferred to C_3.

Manski (1977) formalized random utility approach on the supposition that utilities are unknown to the researcher with certainty and should be treated as random variables. Manski (1973) identified that randomness may occur due to unobserved attributes, unobserved latent variations, measurement errors and proxy variables. All these sources may cause randomness in the distribution of utilities. Hence if we write the utility, U_{in}, of the alternative 1 faced by the individual n as

$$U_{in} = U(x_{in}, s_n, \beta)+e_{in} = V_{in}+e_{in}$$

Where, V_{in} is the observable systematic component of utility which is assumed to be homogenous across individuals and e_{in} is the utility aspect which is unknown to the researcher and is the difference between the observed and the true utility and is also a reflection of idiosyncrasies of individual choice behavior, then

$$P_{in} = P_r(U_{in}>U_{jn}), \qquad i\neq j$$
$$= P_r[(Vin+e_{in})>(V_{jn}+e_{jn})]$$
$$= Pr[(e_{jn}-e_{in})<(V_{in}-V_{jn})]$$

Thus, choice probabilities are analysts' statement of the probability that for any decision maker the utility of an alternative exceeds the utilities of all other feasible alternatives. The case of indifference between u_i and u_j will be discussed later on in this chapter.

In practice, e's are assumed to homoskadastic and uncorrelated as in the linear regression models although in the choice situations they hardly meet this criteria. As an explanation of this point we may refer to McFadden (1975). He interprets the error term as those attributes of the alternatives and individual socio-economic characteristics which are unobservable to the researcher. The assumption that e_i and e_j have the same variance means that the unobserved factors affecting the utility of alternative 1 have the same variation as those affecting the utility of alternative J. But this seldom holds in the real world. As regards to the assumption of un-correlated errors we may remark that we need the difference of e's and thus, the correlated e's do not matter and moreover, it is also very difficult to know the correlation structures. Relevantly we may mention that the theory of individual choice developed in psychology by Thurstone (1927), Luce (1959) assume e's to be white noise.

Different choice models can be obtained by specifying different distributions for the unobservable part of the utility (e_{in}). Two types of general distribution functions have been suggested: one type which contains skewed distribution

and another type which contains varying degree of kurtosis (heavier tails). The suggestion given by Prentice (1976) belongs to the first type while those given by Copenhaver and Mielke (1976) belong to the second type. Standard Normal and Logistic distributions can be treated as special cases. Under the assumption that e_{in} has zero mean Vin indicates the observed part of utility and is often termed as representative or expected or average utility. The choice probabilities possess the properties that $0 \leq p_{in} \leq 1$, $\sum p_{in}=1$, and the relation of p_{in} to the representative utility, holding the representative utilities of other alternatives constant, is sigmoid or S shaped.

When we consider the case of two alternatives we have the binary/binomial choice model. This model is the basic model of choice behavior. However, in our present work we deal with a situation having more than two alternatives for which we need to consider multinomial choice situation and thus, we concentrate on various aspects of multinomial models .

II.3 ESSENCE OF MULTINOMIAL CHOICE MODELS

The extension of the case of more than two alternatives in the choice set yields the multinomial choice model and the most widely used model is multinomial logit model (MNL) in multinomial situation. Its popularity lies in the fact that its parameters are easily estimable.

First of all, let us discuss various routes of deriving multinomial logit and then we proceed to multinomial probit models and here we closely follow Maddala (1983). If we express $V_{in} = X\beta$ which is assumed to linear in parameters

where,

β =Vector of parameters X-vector of explanatory variables which can be linear or non-linear or may be transformed variable following Box-Cox or Box-Tuckey approach, and if we consider logistic distribution for the error term we have the multinomial logit model structure as

$$P_{in} = \frac{X'\beta_i}{1 + \sum_{k=1}^{m-1} e^{X'\beta_k}}$$

where p_i is the probability of choosing ith alternative from a choice set having m salternatives by the nth person.

This model has been used by Thell (1969) to study choices of transportation modes, by Cragg and Uhier (1979) to study demand of Automobiles, Uhier and Cragg (1971) to study asset portfolios, Schmidt and Staruss (1975a) to study occupation choices.

McFadden (1974) developed conditional logit model from random utility maximization. The basic difference between McFadden's conditional logit and multinomial logit is that in MNL choice probability depends on individual characteristics only while McFadden considers the attributes of choice alternatives as well. Thus, if x_{in} denotes the vector of values of alternative i as perceived by individual n, then the probability that individual n will choose alternative i is given by

$$P_{in} = \frac{e^{X'\beta}}{\sum_{k=1}^{m} e^{X'\beta}}$$

In the former formulation P_i has different parameter vector β_i and in the latter case β gives the implicit price for choice attributes. By the latter formulation, for given attributes of a new alternative as perceived by an individual, we can predict the probability that the individual will choose that alternative. In this case, the number of parameters to be estimated is equal to the number of characteristics of alternatives. This model has been derived under that the assumption of extreme-value distribution for the error term and it has a logit structure. To estimate the model we need an identifying normalisation, that is to put one of the parameters equal to zero. It has a well-behaved likelihood function and is computationally comfortable to manage. A convenient feature of the MNL model is that the Hessian of its log likelihood function is everywhere

negative definite (no linear dependence of explanatory variables) and thus any stationary value is globally maximum. Tunali (1988) demonstrated that MNL model can be re-expressed as conditional logit and vice-versa.

In the basic MNL model if x_n is the vector of characteristics of the nth choice maker then the probability that the nth individual will choose ith alternative is

$$P_{in} = \frac{e^{X_n'\gamma_i}}{\sum e^{X_n'\gamma_i}}$$

In this formulation, we need some normalisation and the number of alternatives to be estimated of choice maker's characteristics multiplied by m-1. With this model we can predict the probability that a new individual with the same set of characteristics will choose one of the given alternatives and this model was applied by Schmidt and Strauss (1975a).

We have another type of model called Mixed logit model. In this model we have a latent variable representation $y^*_{in} = x'_{in}\beta+x_n\gamma_i+e_{in}$, where x_n is the vector of characteristics of choice makers and x_{in} is the vector of values of alternative i as perceived by individual n and e is the random error term. The observable variable $y_{in}=1$, if nth individual choose ith alternative and zero otherwise (one can use any pair of real number). The structure of the mixed logit model is,

$$P_{in} = \frac{e^{X_{in}\beta} + x_n\gamma_i}{\sum_{k=1}^{m}(e^{X_{in}\beta} + x_n\gamma_k')}$$

The logit model belongs to the family of share models which define a market share or probability of choosing a particular alternative as the ratio of its utility to the sum of utilities of every alternative in the relevant set.

We can formulate a model where the error term follows a distribution which has some form other than the exponential form h(.)=e(.) but the gain in doing so is marginal and the loss of concavity of the programming structure (which is

not the cass with exponential form) is very costly (MoFadden,1976). Thus, a linear additive function of attributes for the utility suggests the use of simple logit of the exponential form.

While the logic probability has the popularity due to the closed form of cdf, the users of logit probabilities are burdened with the property of Independence of Irrelevant Alternatives (IIA) which states that the raito of the probabilities of two alternatives, i and k, is independent any alternative other then i and k, that is $p_{in}/ p_{ik} = e^{v_{in}} / e^{v_{kn}} = e^{v_{in} - v_{kn}}$

The ratio of the probability of choosing i to that of choosing k is a constant irrespective of the number and compositions of other alternatives in the choice set. Clark(1957) termed this as the constant ratio rule. Luce (1959) views IIA as a probabilistic version of the concept of transitivity.

In some situations IIA property of logit model is inappropriate but if it reflects reality, considerable advantages can be gained. Because of the IIA property, it is possible to estimate the model parameters consistently on a subset of alternatives for each sampled decision maker. In analyzing choice situations in which the number of alternatives is large, estimating on a subset saves time and cost. The IIA property also enables the researcher to make predictions for alternatives that do not currently exist. Train (1982) clarifies this aspect with an example of makes and models of automobiles. The idea lies in the fact that if the full set of alternatives involve currently available makes and models plus the soon-to-be introduced makes and models, then the estimation basd on the currently available information is equivalent to estimation on a subset of alternatives which provides consistent estimates of alternatives. But if the new alternatives involve unique parameters, then this advantage of IIA property disappears.

One useful application of IIA property is to data where preference ranking of alternatives are observed or can be inferred. If the probability of the most

preferred alternative in each choice set satisfy IIA property, it must be of MNL form (see McFadden, 1973). Thus the probability of an observed ranking is,

$$p[1>2>.....>m] = \frac{e^{X_1'\beta}}{\sum_{i=1}^{m} e^{X_i'\beta}} \cdot \frac{e^{X_2'\beta}}{\sum_{i=2}^{m} e^{X_i'\beta}} \cdot \cdots \cdot \frac{e^{X_{m-1}'\beta}}{\sum_{i=m-1}^{m} e^{X_i'\beta}}$$

Therefore, each selection of a next-ranked alternative from the subset of alternatives, not previously ranked, can be treated as an independent observation of choice from a MNL model. This formulation of ranking is due to Marscahk (1960) and was applied by Hausman (1981) to investigate the heterogeneity of parameters across the population. The work of McFadden (1975) has indicated that the IIA property in logit models is not as restrictive as it seems to be. He has shown that any model that specifies choice probabilities including models that do not exhibit IIA property, can be expressed in the form of a logit model as

$$p_{in} = \frac{e^{w_{in}}}{\sum_{j \in C} e^{w_{jn}}}$$

Where w_{in} is some function of observed data.

Logit probabilities with appropriate specification of w_{in} equal the true probabilities.

Any choice model with an appropriate choice of w_{in}, can be put to the logit form which gives rise to the "mother logit' in which the utility function of a particular alternative also depends on the attributes of all other alternatives. The logit model derived from extreme-value distribution and which exhibits the IIA property is a special case of mother logit. In general, w_{in} in the mother logit depends on all observed data including characteristics of alternatives other than i. When w_{in} in the mother logit depends only on characteristics of the decision maker and alternative i, the mother logit becomes standard logit and exhibits the IIA property. Thus logit specification can be used in situations where IIA

does not hold. All that is required is that additional variables related to the alternatives other than the one for which the representative utility is designated are added. The general difficulty is to know what terms to be added to the representative utility to account for the true probability not exhibiting IIA property.

The IIA property may be a weakness if (1) observed and unobserved attributes of utility are not independent of one mother and/or there is correlation of the unobserved components of utility among alternatives. But, as noted by Charles River Associates (1976), IIA is a priori neither desirable nor undesirable, but should be accepted or rejected on empirical grounds depending on the circumstances, (2) prescribed market-share changes when any alternative is altered or the choice is changed by addition or deletion.

Attempts have been made to relax the unrealistic homogeneity assumptions of conditional logit models by modifying the deterministic component as well as random component. McFadden (1974, 1980), Train and McFadden (1978) have modified the deterministic component by explicitly relating it to a function of socioeconomic variables. Empirical work by Hensher (1984), Fischer and Nagin (1981) suggests that transferring the homogeneity in the deterministic component. Florina and Fox (1976), Daganzo and Sheffi (1977) have attempted to modify deterministic component by allowing common subcomponents for some alternatives.

Attempts have also been made to modify the random component. Finney (1971) assumed error distribution to be normal but Currim (1982) demonstrated that without further adjustment the model still is subtect to IIA. Albright, Lerman and Manski (1977), Hausman and Wise (1978) developed random co-efficient Probit model in which deterministic component is modeled as arising from a random normal variable with given mean and variance. Yj is approach is equivalent to Fisher and Nagin's (1981) probit with non independent and nonidentically distributed error term. Daganzo (1979) experimented with negative exponential error distribution with different

variances. Dagazo suggests to allow heterogeneity in the error distribution (independent but not identically distributed) for avoiding IIA. Steckel and Vanhoacker (1988) allow for heterogeneity in the error distribution across individuals and their heterogenous conditional logit model does not conform to IIA property.

Under the assumption that e's in the Random Utility Model follow a multivariate normal distribution, we have the multinomial probit model (MNP). This model was first proposed by Thurstone (1927) and has been applied to psychological choice data by Bock and Jones (1965). Hausman and Wise (1978) have applied the MNP model to a transit choice problem. The use of independent normal distributions for the e's yields the Independent Probit model (see Hausman and Wise, 1978; Currim, 1978). MNP is a much more general model than MNL since it allows e's to have covariance or in other words, alternatives need not be independent and thus IIA is not a general property of MNP. It seems that MNP is to be preferred to MNL as it allows taste variations across decision makers. But, unfortunately the MNP model is preferred for only small number of alternatives since the computations become too complex as can be illustrated by the case of only three alternatives.

To demonstrate this we consider the mapping from the latent to observable variables. Let

$y^*_1 = x'_1\beta - e_1$, $y^*_2 = x'_2\beta - e_2$, $y^*_3 = x'_3\beta - e_3$ and $e=(e_1, e_2, e_3) \sim N(0,\Sigma)$.

Now, $x^*_i\beta - e_i > x'_j\beta - e_j$, if i is preferred to j which is equivalent to $e_i - e_j < x'_i\beta - x'_j\beta$.

Now, the probability that the first alternative is chosen is

$$p_1 = pr(Y^*_1 > Y^*_2, Y^*_1 > Y^*_3)$$
$$= pr(e_1 - e_2 < x'_1\beta - x'_2\beta, e_1 - e_3 < x'_1\beta - x'_3\beta)$$
$$= pr(\varepsilon_{12} < x'_1\beta - x'_2\beta, \varepsilon_{13} < x'_1\beta - x'_3\beta)$$

Now, ε_{12} and ε_{13} follow bivariate normal distribution and p_1 is the cdf of bivariate normal. Similar expression can be derived for p_2 and p_3. Thus, each pi

will involve integration of m-1 dimensional normal density if the choice set has m alternatives.

For $m>4$ the computation becomes very complex and expensive. However, work with fourier transformations may enable computation for $m\geq4$ (see Russel, Farrier and Howell, 1985). The algorithm for MNP model using the Method of Simulated Moments (MSN) developed recently by McFadden (1989) is another revoltionary attempt in this direction and a brief discussion on this approach will be given in chapter 3. The MNL is very often used in empirical applications rather than MNP because of its computational simplicity and because in terms of goodness of fit or prediction accuracy it performs well. Both Probit and Logit type models can capture the consistency between statistical models and the underlying choice-theoretic foundations but these two models differ in flexibility of representation of the essentials of choice structure and flexibility in computation.

Fischer and Nagin (1980) argued that additional computational burden imposed by MNP is compensated for by the broader information it provides. However, for reaching a decision between MNP and MNL models we need a comparison between MNP and a more general logit model and McFadden (1978) put forth an approach which solves the problem with IIA property. By using a distribution which still tractable from for choice probability but also avoids the IIA property. This approach is described below:

To start from a more general form of extreme value distribution for e's i.e., GEV given by, $F(u_1, u_2, \ldots\ldots\ldots, u_m) = e^{-G(e^{-u_1}, \ldots\ldots\ldots\ldots, e^{-u_m})}$

Where $G(w_1, w_2, \ldots\ldots\ldots, w_m)$ is a positive function of $(w_1, w_2, \ldots\ldots\ldots, w_m)\geq0$ and is homogeneous of degree one.

$\lim_{w_i\to\infty} (w_1, w_2\ldots\ldots\ldots, w_m) = \infty$, for $i=1,2,\ldots\ldots\ldots m$. For any distinct $i_1, i_2, \ldots\ldots\ldots i_k$.

$\frac{\delta^k G}{\delta w_{i_1} \ldots\ldots\ldots\ldots \delta w_{i_k}}$ is positive for odd k and negative for even k. Here i_k means that w_i repeats k times.

This distribution yields the GEV model with choice probabilities given by

$$p_i = \frac{G_i(e^{-x_i\beta_1}(\beta_1)\ldots\ldots\ldots\ldots e^{-x_m\beta_m}(\beta_m)e^{-x_i\beta_i}(\beta_i)}{G(e^{-x_i}(\beta_1)\ldots\ldots\ldots\ldots e^{-x_m}(\beta_m)}$$

whereG_i is the partial derivative of G w.r.t. w_i.

This model does not have the IIA property. The special case of $G(w_1\ldots\ldots\ldots w_m)=\sum w_i$. yields the MNL model since e's in this case are i.i.d. extreme-values. The fact that the main difference between the GEV and basic MNL models is that GEV permits correlation of random utility components, whereas basic MNL does not, can suggest that the GEV is likely to share the weakness of the basic MNL model in the presence of random taste variations, omitted variables or grouped data.

An alternative way of generalizing MNL model proposed by Tversky (1972) is the EBA (Elimination by Aspects). EBA views choices as sequential process. It assumes that each alternative is described by a set of aspects and that at each stage of the process an aspect is selected with probability proportional to weight. The selection of the aspect eliminates all the alternatives that do not contain the selected aspect and the selection continues until a single alternative remains. Aspects common to all alternatives do not affect the choice probabilities and do not enter in the decision process.

EBA models become computationally infeasible for large choice sets. For n, aspects, one has to consider 2^n-1 subsets. After realising this problem, Tversky and Sattah (1979) specialised the EBA model to the situations in which the alternatives are represented by a Tree diagram. When the aspects have a tree structure, the EBA model reduces to EBT (Elimination By Tree) or HEBA (Hierarchical Elimination by Aspects). In this case, the decision maker selects a

link from the tree with probability proportional to its length and eliminates all the alternatives that do not include that link. The same process is applied to each selected branch until only one alternative is left. Although HEBA involves fewer parameters than EBA, it has not found many econometric applications because of the difficulty, in defining and measuring objective attributes.

A computationally tractable model suggested by McFadden (1981) is the NMNL(Nested Multinomial Logit Model) which involves the sequential use of the multinomial logit process. He argues that HEBA and NMNL give virtually the identical fit to the data, but in many econometric applications with several explanatory variables it is better to use NMNL rather than HEBA model from the viewpoint of computational case. As an illustration of the NMNL model MCFadden (1978) considers the problem of choice of housing locations. This model allows for a general pattern of dependence among the choices and avoids the IIA property. Thus, in a hierarchical model each transition in the tree is described by a MNL model with one of the variables being an inclusive value which summarises the attributes of alternatives below a node. When the inclusive value coefficients are equal to one, NMNL collapses to MNL. McFadden (1979) conducted that ...." a sufficient conditions for a NMNL to be consistent with the individual utility maximization is that the coefficient of each inclusive value lies between zero and one". Inclusive values give a summary measure of the desiarability of a set of alternatives connected directly to a higher level decision and should be interpreted as a utility for the opportunity set of alternatives associated with choice lower in the hierarchy, regardless of which alternative is selected in actuality.

MNL model is a linear restriction on a sequential model with inclusive value equal to unity. The sequential nested logit model generilises MNL model by making the inclusive value to vary between 0 and 1. It allows pair wise attributes correlation maintaining the assumptions of equal-variance within and between all choices in a hierarchy. Nested logit model is considered as structured logit (Daly and Zachary, 1978, Daly, 1979). GEV nested logit model

incorporates an index of similarity between unobserved attributes of alternatives and non-equal variance. GEV provides conditions under which nested logit model can be derived from a theory of stochastic utility maximization (McFadden, 1979). Ben –Akiva nad Lerman (1979), Williams (1977) and Daly and Zachary (1978) have shown that parameterization of inclusive value is important for making sequential logit to be consistent with utility maximization.

There is also an adhoc extension of MNL model to avoid the IIA property and that is the Dogit model due to Gaudry and Dagenasis (1979). In fact, MNL is a special case of the Dogit Model when certain parameters in the Dogit are set to zero. The Dogit model has a structure:

$$p_{in} = \frac{\exp(x'_{in}\beta) + \theta_i \sum_{i=1}^{m} \exp(x'_{in}\beta)}{(1 + \sum_{i=1}^{m} \theta_i) \sum_{i=1}^{m} \exp(x'_{in}\beta)}$$

where i=1,2,.....m.

pin is the probability that the nth person chooses the ith alternative, $\theta=(\theta_1, ...\theta_m)$ is a vector of parameters to be estimated. This model is flexible enough for the choice of pairs of alternatives to be consistent with the IIA property without imposing the constraint on other pairs of alternatives as done by MNL model. This model does not in general have the IIA property as long as $\theta_1>0$.

As a modification of conditional MNL model Anemiya and Nold (1974) proposed incorporating a random variable U_i with zero mean and constant variance σ^2 as a surrogate for the omitted variable as in the usual regression model and their model may be written as:

$$Pi - \frac{\exp(x_i\beta + u_i)}{\sum \exp(x'_i\beta + u_i)}$$

The stochastic formulation of the model together with the covariance structure weakens the IIA property but the assumption of constant variance may be violated in some situations.

A simple one parameter generalization of Logistic distribution is the Burr family of distribution. The Burr type II (B2) generalizes the logistic distribution. The (closed form) cdf of the trivariate B2 distribution is given by

$$F(e_1, e_2, e_3) = [1 + e^{-e_1} + e^{-e_2} + e^{-e_3}]^{-k}$$

Fry (1988) developed a trivariate binomial choice model by using this trivariate B2 distribution and has shown that the different probabilities can be expressed in terms of the difference operators operating on the cdf of B2 distribution. From the estimation point of view Fry's model appears tractable, but he did not derive any diagnostic tests for the model accuracy.

The alternative ways of representing the choice process are complementary to a varying degree. It is not possible to identify a model to be the best and this is a reflection of variation in the population with respect to the evaluation of attributes in preference ranking of alternatives. There is a number of isolated works concerning the various facets of thresholds in choice. Henshen and Goodwin (1978), Blase(1989) have explored the influence of habits on the trade-off between the generalised costs of alternative modes.

Krishnan (1977) has tested a modified logit model incorporating jnd (just noticeable difference) on the traveller's choice of modes. Willium and Ortuzer (1979) have investigated the probability of integrating behavioral constructs such as habit, threshold and limited information into the random utility framework so that logit and probit models might be sensitive to such influences. Krishnan's contribution is most relevant to the development of operational individual choice model using MNL formulation. He hypothesised that " the satisficer would become a utility maximiser when a sufficiently more attractive alternatives are made available to them". The jnd in Krishnan's logit

model is used to overcome the restrictive assumption that the probability of attaining a state of indifference (when two alternatives are equal) is zero. If indifference exists at all, then i and j will be chosen randomly. Once a choice model is specified, we are faced with the problem of estimating the associated parameters and the choice probabilities.

Now, we pass on to next chapter to discuss issues related to estimation and tests associated with choice models.

II.4 SOME COMPARATIVE APPROACHES TO DEAL WITH DISCRETE DATA MODEL

(i) Qualitative variable may be binomial or multinomial and multinomial responses can be ordered, unordered or sequential.

(ii) Actual similarity in the study of multinomial or multivariate choice problem which can be demonstrated with the help of quantitative representation of qualitative variables.

(iii) If a set of alternatives occur in two ways and individual will choose one of them then the multivariate binomial choice problem can be converted to a multinomial one but there exist some differences between the two.

(iv) The ML procedure is based on individual observation and can be used with categorical or continuous data which developed by R.A. Fisher in 1922. Joint density function of sample values which regarded as a function of unknown parameter is called Likelihood Function. ML method consists in finding such values of parameter which maximizes the Likelihood Function. The values of parameter which maximizes likelihood function is called maximum likelihood estimate of the parameter. They are found by the method of differentiation under the following regularity conditions.

- Range of the observed variable is independent of parameter.
- Parameter estimate is continuous.
- Likelihood function is twice differentiable function of the parameter.

- Maximum value of likelihood function is attained at an interior value of parameter.

(v) WLS requires adhoc adjustment of cells including all positive responses or no positive responses.

(vi) If the number of observation is small, WLS procedure is likely to produce empty cells.

(vii) Manski Shows that maximum score estimates are generally not unique since more than one value of parameter estimates may yield the same % right values but non- uniqueness is reduced along with the increase in sample size.

(viii) A multivariate discrete choice model specifies the joint probability distribution of two or more discrete dependent variables.

(ix) Specialized by Grizzle (1970) is multivariate logit model. It ignored the probability structure of the error terms and thus. their results may not be too efficient in moderate or small samples. It does not generate correlation estimates for Joint distribution responses.

(x) Multivariate probit provides the desired correlation structure.

(xi) McFadden's GEV model / NMNL model is useful in multivariate situation because the alternatives can be naturally classified according to the outcome of one of the variables such that each of the classes contains similar alternatives. It is advantage that fewer parameters need to estimate but their correlation structure may be more complex then it is assumed to be.

(xii) Nerlove and Press (1973) Log linear model depends on independent variables- particularly main- effect parameters. This creates certain unbearable restriction in some cases.

VARIABLES

Here we provide an account of variables types used in LDV models.

Changeable values of any characteristics which is indicated by any letter is called variables. In econometrics variables are termed as endogenous and exogenous variables e.g. dependent and independent variable. Endogenous variables are those determined within the economic system and exogenous variables are given from the out side the system. In a broad sense almost all variables are endogenous and the only exogenous variables one can think of are weather, cyclones etc. As for example, studying the demand for gasoline by households, we can treat the quantity demanded as endogenous and income ands prices are exogenous. However, as we lengthen the time period of our observations, these variables will also become endogenous. In general, the greater the level of aggregation whether it be over time periods or over individuals cross section units the more exogenous variables will have to be treated as endogenous. Endogenous variables are classified as (i) target variable and (ii) non-target variable. Target variables are those we like to influence and non-target variables are those we do not care about i.e, employment and price may be target variables.

Exogenous variables are classified as (i) instruments and (ii) non-instrument variables. An instrument is an exogenous i.e, specifically manipulated so as to achieve some targets. Government expenditure, taxes and subsides are instrumental variable. If we consider, $Y= b_0+b_1X + u$.

In this model, X= Independent/exogenous/explanatory variable and
Y= Dependent /endogenous /explained /study variable

Dummy-endogenous variable model: Let us consider problem is to estimate the returns to college – going education. Suppose that we are given data on the variable y (income) for n individuals; n_1 of them have college

education and n_2 of them have not college education. We are interested in estimating the average increment in income

due to going to college.

One simple solution is to regress y-on a dummy variable D defined as-

D = 1, if the individual has college education

= 0, otherwise.

The co-efficient of D will measure the average increase in income due to going to college.

If $\bar{y}_1$ is the mean income of those with college education.

$\bar{y}_2$ is the mean income of those without college education

$\therefore$ Coefficient of D= $\bar{y}_1 - \bar{y}_2$

Suppose that all individuals have the same opportunities available to them and that an individual goes to college if he thinks his income will be higher by having college education. Otherwise he does not.

Let , y_1*= expected income with college education

y_2*= expected income without college education.

The individual goes to college if y_1*>y_2*

then the dummy variable is D= 1 if y_1*> y_2*

= 0 other wise

suppose that y_1*= $\alpha_1 + u_1$

y_2*= $\alpha_2 + u_2$

where, u_1 & u_2 have a bivariate normal distribution with mean zero and a covariance matrix

$$\begin{vmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{vmatrix}$$

we are to estimate α_1, α_2, σ_1^2, σ^2_2, and σ_{12}

The difference $\alpha_1 - \alpha_2$ = measure the average increase in income due to college education. Since u_1 & u_2 are normally distributed, so, u_1 - u_2 is also normally distributed with variance $\sigma^2 = \sigma^2_1 + \sigma^2_2 - 2\sigma_{12}$

Probability that an individual goes to college is

$p(y_1^* > y_2^*) = p(u_1 - u_2 > \alpha_1 - \alpha_2)$

$$= 1 - \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \text{ ; where, } z = \frac{\alpha_2 - \alpha_1}{\pi}$$

Note that, if $\alpha_2 = \alpha_1$ the estimate of z is zero or $\alpha_2 = \alpha_1$

Thus the incremental returns to college education are zero.

Endogenous variables

The variables whose values are determined by the simultaneous interaction of the relations in the model is called endogenous variable. As for example, Let us consider an income determination model as, $C_t = \alpha + \beta y_t + \mu_t$ and $y_t = C_t + Z_t$. Here, C_t and y_t are endogenous variables.

Exogenous variables

The variables whose values are determined outside the model is called exogenous variable. As for example, Let us consider an income determination model as, $C_t=\alpha+\beta y_t+\mu_t$ and $y_t=C_t+Z_t$.

Here, Z_t is endogenous variables.

Lagged variable

In economic phenomenon, generally , a cause often produces its effect after a lapse of time; thise lapse of time(between cause and its effect) is called lag and lagged variables refers to the time-lagged values of the dependent or independent variables.

Sometimes the values of the current period of the dependent variables depend on the lagged values of some of the explanatory variables or the lagged values of the dependent variables. The variables with lagged values are called lagged variables. Lagged variables means simply the lagged independent variables.

The variables are to be lagged variables which are use to cover the inclusion on the right hand side of the regression equation of lagged values of the exogenous variables X's and lagged values of the dependent variable Y. The general form of a distributed lagged model with only lagged exogenous variables is, $Yt=\beta_0+\beta_1X_t+\beta_2X_{t-1}+.....+\beta_sX_{t-s}+u_t$

where, X_{t-1} , X_{t-2}, ...X_{t-s} are to be lagged variables.

for example, advertisements expenditure done by a firm may influence consumer's demand for its commodity over a successive period . Also demand in a particular period may have the effects advertisements done in various period.

Stochastic explanatory variables

In econometric model, when the explanatory variable X is random, variable is called stochastic explanatory variables.

Let us consider the linear regression model, $Y=\beta X+u$, we have been assumed that the explanatory variable X is not random variable. This assumption is rarely valid in econometrics where the observations are seldom generated by a controlled experiment. Also, in the case where the explanatory variables are lagged dependent variables, they can not be considered as nonrandom.

If X is a random (or stochastic) variable, the derivation of exact finite sampling distributions becomes difficult and we often have to use asymptotic results; i.e., what we consider is the properties of estimators as the sample size grows very large.

Some Indicative discussion is provided below.

Asymptotically Unbiased ness

If $\hat{\theta}_n$ is an estimator of θ based on a sample of size n, θ_n is called an asymptotically unbiased estimator of θ if $\lim n \to \infty E(\hat{\theta}_n) = \theta$. This definition is not useful if $E(\hat{\theta}_n)$ does not exist. In other words, $\hat{\theta}_n$ is an asymptotically unbiased estimator of θ if the mean of the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ is zero. $\hat{\theta}_n$ is an asymptotically unbiased estimator of θ if $AE(\hat{\theta}_n)= 0$ (AE= asymptotic expectation)

The same is the case with the asymptotic variance. It can be defined as, $\lim \xrightarrow{n \to \infty} E\left[\sqrt{n}(\hat{\theta}_n - \theta)\right]^2$ if we consider limits of variances or $AE\left[\sqrt{n}(\hat{\theta}_n - \theta)\right]^2$.

Consistency

Suppose, $\hat{\beta}_n$ is the estimator of β based on sample size n. Then the sequence of $\hat{\beta}_n$ estimators is called a consistent sequence if for any arbitrarily small positive numbers ε and η there is a sample size n_0 such that $P\left[\left|\hat{\beta}_n - \beta\right| < \varepsilon\right] > 1 - \eta$ for all $n>n_0$

e.g., by increasing the sample size n, the estimator $\hat{\beta}_n$ can be made to lie arbitrarily close to the true value β with probability arbitrarily close to 1. We write this as plim $\hat{\beta}_n$ =β (plim is probability limit). In practice we often drop the subscript n on $\hat{\beta}_n$ and also drop the words "sequence of estimators" and merely say $\hat{\beta}$ is a consistent estimator for β.

Sufficiency

A sufficient condition for $\hat{\beta}$ to be consistent is that the bias and the variance should both tend to zero as the sample size increases. This condition is often useful to check in practice, but it should be noted that the condition is not necessary. There are also some relations in probability limits that are useful in providing consistency. These are

plim(c_1y_1=c_2y_2) =$c_1$plim(y_1)+ c_2 plim(y_2), where c_1 and c_2 are constants,

plim(y_1y_2)=plim(y_1)plim(y_2) and $\text{plim}\left(\frac{y_1}{y_2}\right)=\frac{p\lim(y_1)}{p\lim(y_2)}$, provided plim($y_2$) ≠0

Also if plim y=c and g(y) is a continuous function of y, then plim g(y)=g(c).

Another concept that is commonly referred to is that of limiting distribution. If we consider a sequence of random variables y_1, y_2, y_3, ... with corresponding distribution functions F_1, F_2, F_3, ..., this sequence is said to converge in distribution to a random variable y with distribution function F if $F_n(x)$ converges to F(x) as $n\to\infty$ for all continuity points of F(x). This distribution function F(x) is called the limiting distribution of this sequence of random variables. It is not necessary true that the moments of F are the limits of the moments of F_n as $n\to\infty$. In fact, often it is the case in econometrics that F_n does not have moments but F does. As an example, suppose x is a random variable with mean $\mu\neq0$ and variance σ^2 and P(X=0)=c>0. Suppose a random sample x_1, x_2, ...x_n is drawn from the

population and $\bar{x}_n$ is the sample mean based on a sample of size n. If we consider the sequence of random variables $y_n = \frac{1}{\bar{x}_n}$, then E($y_n$) does not exist because there is a positive probability that $x_1 = x_2 = \ldots\ldots = x_n = 0$ and thus $y_n = \infty$ when $\bar{x}_n = 0$.

However, it can be shown that $\sqrt{n}(1/\bar{x}_n - 1/\mu)$ has a limiting distribution which is normal with mean zero and variance σ^2/μ^4 so that y_n is asymptotically normally distributed with mean $1/\mu$ and variance $\sigma^2/n\mu^4$. Because moments may not exist for F_n.

Proxy Variables

Often the variables we measure are not really what we want to measure. It is customary to call the measured variable a "proxy" variable – it is a proxy for the true variable. For instance, in an investigation of the effect of education on income, what we want to measure is years of education. Instead we measure years of schooling, which is a proxy for years of education. Sometimes the proxy variable is treated as just the true variable with a measurement error. For example, the "permanent income hypothesis" of Friedman where the true variable we want to use is "permanent income' Y_p. Instead we use the proxy "measured income" Y. It is assumed that $Y=Y_p+Y_t$, where, Y_t is the "transitory income", is assumed to have the following properties:

$E(Y_t)=0$

$Cov(Y_t,Y_p)=0$,

i.e.,Y_t is an error term with zero mean and is uncorrelated with the true variable Y_p.

Consider the regression model,

$y=\beta z+u$

where, instead of the true variable z, we measure the proxy p=z+e.

In addition to the usual assumption cov(u,z)=0,it is often assumed that cov(e,z)=cov(e,u)=0,i.e., the measurement error e is uncorrelated with both the error in the regression equation u and the "true" value of the variable z.

Testing Techniques

Here we discuss some testing procedures which can be useful for model application.

The likelihood- Ratio (LR) Test

Most of the tests in common use can be derived from an important Principle due to Neyman and pearson (1928) called the likelihood-ratio principle. The LR test play a role in the theory of testing similar to that of the maximum-likelihood method in the theory of estimation.

Simply stated, the likelihood ratio

$$\lambda = \frac{\max L\ (\theta) over\ \ values\ \ of\ \theta\ specified\ \ by\ Ho}{Mox\ L\ (\theta)\ over\ all\ \ values\ \ of\ \theta}$$

Where $L(\theta)$ is the likelihood function. clearly λ lies between o and I because the denominator, which is an unrestricted maximum of $L(\theta)$ has to be $\geq$ the numerator, which is a restricted maximum of $L(\theta)$. we reject Ho if $\lambda \leq -c_{\alpha}$ where c_{α} is a constant defined so that the type I error is α.

The intuitive argument for the test as follows

If we are comparing the plausibility" of one value of θ against another, given a sample (y_1, y_2, -----,y_n). we would be choosing that value of θ which gives the likelihood a larger value. If we can not find an appreciably larger value of the likelihood function. by searching through the values of θ other than those specified by Ho: than we are can reason that the most plausible value θ belongs to the set specified by Ho: Hence we accept Ho:

The Spearman rank-correlation test

Test is the simplest one, which may be applied to either small or large samples, and may be outlined as follows:

i. The regression equation of y on x is given by,

$y=b_0 + b_1x + \mu$

and we obtain the residuals, e's which are estimates of the μ's.

ii. We order the e's (ignoring their sign) and the X values in ascending or descending order and we compute the rank correlation co-efficient,

$r=1-\frac{6\sum d_i^2}{n(n^2-1)}$ where di =difference between the ranks of corresponding pairs of X and e, and n = no. of observations in the sample.

Here the hypothesis which is to be tested.

H_0: μ's are homoscadastic.

H_i: μ's are hateroscadastic

Test statistic is,

$t=\frac{r\sqrt{n-k}}{\sqrt{1-r^2}}$ ~t with n-k d.f

Where k is the no. of parameter in the model. If this t is significant then we may suggest that the μ's are heteroscadastic.

If we have a relationship with more explanatory variables we may compute the rank correlation coefficient between μi and each one of the explanatory variables separately.

The Goldfeld and Quandt test

This test is applicable to large samples. The observations must be at least twice as many as the parameters to be estimated. The test assumes normality and serially independent μ_i's. Here the hypothesis is

H_0: μ's are homoscadastic.

H_i: μ's are hateroscadastic

The steps involved in Goldfeld and Quant test may be outlined as follows:

Firstly, We order the observations according to the magnitude of X's the explanatory variables.

Secondly, we selected arbitrarily a certain number (C) of central observations, which we omit from the analysis. The remaining (n-c) observations are divided into two sub-samples of equal size (n-c)/2, one including the small values of X and the other including the larger values of X.

Thirdly, we fit separate regressions to each sub-sample, and we obtain the sum of squared residual from each of them.

$\sum e_i^2$ = residual sum of squares from the sub-sample of low values of X, with $\frac{(n-c)}{2} - k$ d.f., where k is the total number of parameters in the model.

$\sum e_{2i}^2$ = residual sum of squares from the sub-sample of high values of X, with $\frac{(n-c)}{2} - k$ d.f., where k is the total number of parameters in the model.

Test statistic is formed as,

$$F = \frac{\sum e_2^2 / [\{(n-c)/2\} - k]}{\sum e_1^2 / [\{(n-c)/2\} - k]} = \frac{\sum e_2^2}{\sum e_1^2} \sim F(\nu_1, \nu_2) \text{ where, } \nu_1 = \nu_2 = \frac{n-c}{2} - k$$

The observed F is compared with the theoretical value of F with (ν_1, ν_2) d.f at a chosen level of significance. If the observed F< the theoretical F, we accept that the μ's are homoscadastic; otherwise the μ's are heteroscadastic.

The Glejser test for heteroscadasticity

This test may be outlined as follows:

i. We perform the regression of Y on all the explanatory variables and we compute the residuals e's.

ii. We regress the absolute value of e's on the explanatory variables.

The actual form of this regression is usually not known, so that are may experiment with various powers of X. For example , $|e| = a_0 + a_1 X_i^2$

or, $|e| = a_0 + a_1 X_i^{-1}$

or, $|e| = a_0 + a_1 X_i^{1/2}$ and so on.

We choose the form of regression which gives the best fit in the light of the correlation co-efficient and standard errors of the coefficient a_0 and a_1.

Heteroscadasticity is in the light of the statistical significance of a_0 and a_1 ;that is we perform any standard test of significance($\hat{b}_i$,t, F) for these coefficient ,and if they are found significantly different from zero we accept that the μ_i's are hetroscadastic

Likelihood Ratio, Wald and Lagrange Multiplier Test Principles

We consider three general procedures based on the likelihood ratio (LR), Wald (W) and lagrange multiplier (LM) tests. Although these tests generally differ in computational complexities, they are asymptotically equivalent. In particular, they share some asymptotic optimality properties such as consistency and best local power when the parameters specified under the alternative hypothesis differ from the null values by a term to the order of o ($1/\sqrt{n}$) .As it is typical of most econometric models applied in practice that specification tests with known small sample properties are not available, the LR, W and LM tests are especially valuable.

We shall start by defining the framework of a parametric model which will be analyzed under the maximum likelihood principle. The LR, W and LM statistic will be defined and their asymptotic distribution established. These statistic also have asymptotically equivalent but computationally different forms. We will discuss some of these variation and comment on their relative computational

advantages. Comparison of the tests in small samples will be discussed later when we consider their applications to specific models and hypothesis.

We now define the LR, W and LM test statistics for testing Ho. Some geometrical interpretation of these tests will be provided, which would help to understand the relationship among the tests. Discussion of the asymptotic properties of the tests will be postponed .

The test statistics are defined as

LR = $L\left[l(\hat{\theta}) - l(\bar{\theta})\right]$

W = θ $\hat{}$ $=n\,h(\hat{\theta})'\left[H(\hat{\theta})V(\hat{\theta})^{-1}H(\hat{\theta})'\right]^{-1}h(\hat{\theta})$

LM= $d(\tilde{\theta})'I(\tilde{\theta})^{-1}d(\tilde{\theta}) = d(\tilde{\theta})'V(\tilde{\theta})^{-1}d(\tilde{\theta})/n$

$\bar{\theta}$ $\tilde{\theta}$

If H0: is true, we would expect $\hat{\theta}$ and $\bar{\theta}$ to be close, and like wise l($\hat{\theta}$) and l($\tilde{\theta}$). Hence, comparing l($\hat{\theta}$) and l($\tilde{\theta}$) gives the LR test. An alternative is to consider h($\hat{\theta}$), which is expected to be close to zero if Ho is true. Hence W test is based on the significance of the departure of h($\hat{\theta}$) from zero. Finally, if Ho is true, d($\tilde{\theta}$) is expected to be close to zero, by virtue of the fact that d($\hat{\theta}$)=0. Another way to look at this is to note that d($\tilde{\theta}$)=H($\tilde{\theta}$)$'\tilde{\lambda}$. Given that H($\tilde{\theta}$) has full rank, d($\tilde{\theta}$)=0 is equivalent to $\bar{\lambda}$=0, ie, the Lagrange multipliers vanish. Thus, the test based on the score vector, as originally proposed by Rao (1948), is identical to the test based on the Lagrange multipliers suggested by Aitchison and Silvey (1958) and Silvey (1959). As the Lagrange multipliers can be interpreted as the implicit costs (shadow price) of imposing the restrictions, it is not surprising that a test of the restrictions can be conducted by examining the sample values of the multipliers.

Asymptotic Distributions of LR, W and LM:

Under the regularity conditions R1 to R3, and when Ho is true, the LR, W and LM statistics are asymptotically equivalent and distributed as χ_r^2 .

We have LM= $d(\tilde{\theta})' I(\tilde{\theta})^{-1} d(\tilde{\theta})$

$$= \frac{\bar{\lambda}'}{\sqrt{n}} H(\tilde{\theta})' V(\tilde{\theta})^{-1} H(\tilde{\theta})' \frac{\bar{\lambda}}{\sqrt{n}}$$

which is asymptotically equivalent to $\left(\tilde{\lambda}/\sqrt{n}\right)' R(\theta)\left(\tilde{\lambda}/\sqrt{n}\right)$. Thus LM$\rightarrow \chi_r^2$.

For W, we expand h(θ) about $\hat{\theta}$ to obtain, under Ho, $\sqrt{n}h(\hat{\theta}) + \sqrt{n}H(\theta)(\theta - \hat{\theta})$

=0

Hence, W = $h(\hat{\theta})'\left[H(\hat{\theta})I(\hat{\theta})^{-1}H(\hat{\theta})'\right]^{-1}h(\hat{\theta})$

=n $h(\hat{\theta})'\left[H(\hat{\theta})V(\hat{\theta})^{-1}H(\hat{\theta})'\right]^{-1}h(\hat{\theta})$

=n $h(\hat{\theta})'R(\theta^{-1})h(\hat{\theta})$

=n $(\hat{\theta} - \theta)' H(\theta)R(\theta)^{-1}H(\theta)(\hat{\theta} - \theta)$

$$= \frac{\bar{\lambda}'}{\sqrt{n}} R(\theta)^{-1} \frac{\bar{\lambda}}{\sqrt{n}}$$

=LM

The last but second equality above comes from the fact $\frac{\bar{\lambda}}{\sqrt{n}} =$

$-R(\theta)^{-1}H(\theta)(\hat{\theta} - \theta)$.

Finally, expanding I($\tilde{\theta}$) about I($\hat{\theta}$) we obtain

$$I(\tilde{\theta}) = I(\hat{\theta}) + \frac{1}{2}(\tilde{\theta} - \hat{\theta})\frac{\partial^2 I(\theta)}{\partial\theta\partial\theta'}(\tilde{\theta} - \hat{\theta}).$$

Which implies

$$LR = (\hat{\theta} - \theta)' I(\theta)(\hat{\theta} - \theta).$$

We have,

$$\sqrt{n}(\hat{\theta} - \tilde{\theta}) = V(\theta)^{-1}H(\theta)'R(\theta)^{-1}H(\theta)V(\theta)^{-1}\frac{d(\theta)}{\sqrt{n}}$$

$$\sqrt{n}(\hat{\theta} - \tilde{\theta}) = V(\theta)^{-1}H(\theta)'\frac{\bar{\lambda}}{\sqrt{n}}$$

Therefore,

$$LR = \sqrt{n}\left(\hat{\theta} - \tilde{\theta}\right)' V(\theta)\sqrt{n}(\hat{\theta} - \tilde{\theta}).$$

$$= \frac{\tilde{\lambda}''}{\sqrt{n}} H(\theta)V(\theta)^{-1} H(\theta)' \frac{\tilde{\lambda}}{\sqrt{n}}$$

=LM

thus LR, W and LM are asymptotically equivalent under Ho, and they all tend to χ_r^2 in the limit.

The above theorem assumes that the null hypothesis is correct. When the null hypothesis is false the test typically reject it with probability approaching one when n approaches infinity (i.e, the test is consistent). As it remains important to examine the power functions of the tests in the limiting case, we consider alternative hypothesis which are very near to the null hypothesis. These alternatives are called local alternatives.

Thus we consider,

$$H_1: h(\theta) = \frac{\partial}{\sqrt{n}}.$$

For some constant vector $\partial, \partial'\partial < \infty$ obviously as n→∞, H_1 coincides with Ho.

Under the local alternatives, expanding $h(\hat{\theta})$ around $h(\theta)$ gives $\theta\sqrt{n}h(\hat{\theta}) = \partial + \sqrt{n}H(\theta)(\hat{\theta} - \theta)$. Thus $\sqrt{n}h(\hat{\theta}) \xrightarrow{D} N(\partial, R(\theta))$, which implies the W statistic is asymptotically distributed as a non-central χ_r^2 with r degrees of freedom and non-centrality parameter $\mu = \partial' R(\theta)^{-1}\partial$ (denoted as χ_r^2). Under H_1, the zero sub vector on the RHS replaced by $-\partial$. Hence, there is an additional term of $-V(\theta)^{-1} H(\theta)' R(\partial)^{-1}\partial$ on the RHS, and the asymptotic mean of $\bar{\lambda}/\sqrt{n}$ is $-R(\theta)^{-1}\partial$. Repeating the argument in the proof of theorem it can be shown that LR and LM are both asymptotically distributed as χ_r^2.

Thus under the null as well as the local alternative hypothesis all three tests are asymptotically equivalent, which implies asymptotic comparison cannot provide any basis for using one test in preference to another.

μ is a finite constant, the LR, W and LM tests have asymptotic power strictly less than one for testing against local alternatives. Further discussions on the asymptotic equivalence and optimality of the tests can be found in Engle (1984). From now on when we refer to the asymptotic equivalence of two tests, the statement is made with respect to the null and local alternatives.

CHAPTER III

ESTIMATION AND TESTING ISSUES ASSOCIATED WITH CHOICE MODELS

III.1 Motivation

Although choice model is very nice by itself, it is associated with severe estimation and testing problems. We focus on such issues in this chapter. First of all, we discuss estimation issues followed by those of tests.

III.2 Estimation Issues

A. Maximum Likelihood (ML) method

We now discuss the estimation problems in two situations. One when the sampling process is exogenous and the one where it is endogenous or choice-based.

If the sample is random or stratified random with the strata defined on factors that are exogenous to the choice being analyzed, then the probability of person n choosing the alternative i is $\prod_{i \in C} p_{in}^{y_{in}}$ where y_{in} is one if the person n chose alternative i and zero otherwise. Since, the decision maker's choice is assumed to be independent of that of others,

we have $L = \prod_{n \in N} \prod_{i \in C} p_{in}^{y_{in}}$ where N is the set of decision makers. This expression gives the probability of each person's chosen alternative multiplied across all persons in the sample. As a function of the parameter vector β, it can be considered as the likelihood function (L). The Maximum Likelihood Estimation (MLE) of β is consistent and efficient (McFadden,1973). Given the MLE, we can demonstrate that the estimated values of the alternative-specific constants are those that equate the average probability for each alternative with the share of

the decision makers in the sample who chose that alternative. In the case when the number of alternatives is too large, it necessitates the researcher to consider only a subset of alternatives for estimating the parameters. In this case the estimates will be consistent but will be inefficient due to the loss of information. In some situations, a sample drawn on the basis of exogenous factors would include fewer people than who have chosen a particular alternative.

Estimation of model parameters with samples drawn at least partially on the basis of decision maker's choices is fairly complicated and varies with the exact form of the sampling procedure. If a researcher is using a purely choice-based sample and includes an alternative-specific constant in the representative utility for each alternative, then estimating the model parameter as if the sample was exogenous produces consistent estimates for all model parameters except the alternative specific constant. The estimated constant β_i, for alternative i, is related to the true constant β_i^* as $E(\beta_i) = \beta_i^* - \log(A_i/S_i)$ where A_i is the % of decision makers in the population that chose i and S_i is the % of the people in the choice-based sample that chose alternative i. If the population shares are known then a consistent estimate is the estimated constant plus $\log(A_i/S_i)$.
For a specified level of precision, choiced-based sampling can reduce the sample size and cost. Indeed for infrequently chosen alternatives choiced-based samples are recommended. However, Manski and Lerman(1977) and Manski and McFadden (1981) have shown that the usual MLE's are inconsistent for choiced-based samples. Manski and Lermain (1977) proposed a weighted likelihood function which yields consistent estimates. For random-choice-based samples the standard ane weighted MLE's are identical but for a stratified choice-based sample two MLE's are different. In our case we do not have choice-based sample and thus we discuss the estimation procedures that are relevant in our case. Several estimation procedures for choice models have been discussed in the literature (see Flath and Leonord, 1979; Green, Carmone,

and Wachsprest, 1977). Two estimation techniques are widely used, namely, the Maximum Likelihood (ML) and the Weighted Least Squares (WLS) methods. In the ML method the estimator maximises the log of Likelihood Function(1). It can be shown that 1 of the commonly used choice models is globally concave and therefore the solution is unique if it is bounded. An implication of this fact is that any iterative procedure which converges to a stationary point also converges to a global maximum (Amemiya,1981). It is also well known that MLE's are consistent, asymptotically normal and have the minimum variance. McFadden (1968) has shown that an unique maximum exist for MNL model except in very special condition not usually encountered in practice.

B. Weighted least squares (WLS) Method

WLS or Minimum chi-square method involves the use of log of odds as the dependent variable and the attributes as explanatory variables. The WLSE's can be shown to be asymptotically equivalent to the MLE's. Computationally manageable WLSE have been discussed in the literature (see Green, Carmone and Wachspress, 1979, Nakanishi and Cooper,1982). The ML method, however, has several advantages over WLS. WLS requires many observations per cell. Grouping of observations that are not exact replications to achieve the cell frequencies required for the application of WLS makes the estimators inconsistent and may make it seriously biased (McFadden,1974a). Moreovr, for more than 3 or 4 independent variables it is often impossible to group the data. The ML procedure is based on individual observation and can be used with categorical or continuous data. WLS requires adhoc adjustment if cells include all positive responses or no positive responses. MLE does not need any such adjustment for small sample ($n<30$) both the methods may be inappropriate. For small samples the properties of MLE are not understood while the disadvantages of WLS are magnified (McFadden, 1974a). If the number of observations is small, WLS procedure is likely to produce empty cells. In many of the choice modeling situations, such as marketing research, the analysis is

done at the individual level and the number of observations available is generally small, thus, WLS is inappropriate. As mentioned above, it is important to distinguish between many observations and few observations per cell. Presence of many observations per cell can improve the estimates (see Amemiya, 1981). The ML method is usable for many observations as well as for few observations per cell and it is its another advantage.

In the context of our present problem, we consider that ML method to be the most appropriate since we do not have many observations per cell. For the multinomial logit model with a sample of size N we have the likelihood function as,

L= $\prod_{n=1}^{N}\prod_{i\in C} p_{in}^{y_{in}}$ Where $y_{in}=1$ if the person chooses alternative i and zero otherwise

$= \prod_{n=1}^{N}\prod_{i\in C}\left\{\exp(x_{in}'\beta)/\sum e_{in}^{x'}\beta\right\}^{y_{in}}$ and the log-likelihood function is

$$I=\sum_{n=1}^{N}\sum_{i\in C} y_{in}\left[x_{in}'\beta - ln\sum_{i\in C} e^{x_{in}'}\beta\right] \text{ and}$$

$$\frac{\partial I}{\partial \beta} = \sum\sum y_{in}\left[x_{in} - \frac{\sum e^{x_{in}'\beta x_{in}}}{\sum e_{in}'\beta}\right]$$

For multinomial probit model we put $p_{in}=\phi(x_{in}'\beta$ where ϕ is the cumulative normal density.

C. Non-parametric method

One class of estimation approach similar to ML but which optimises a different function is the non-parametric estimation technique. In this case no specific distributional assumption of the error terms is adopted and it is hypothesised that the distribution belongs to some class having some very general properties. Manski (1975) proposed maximum score estimation technique which selects parameters so that the fraction of the sample who chose the alternative

with the greatest systematic components of the utility is maximised. Manski shows that maximum score estimates are generally not unique since more than one value of parameter estimates may yield the same % right values but non uniqueness is reduced along with the increase in sample size.

The most relevant condition for maximum score consistency is that for each individual in the population the choice probability must be monotonically increasing with the systematic components of utility, that is, greater V_{in} is, the greater is p_{in}. Thus, each individual's disturbances need not come from the same distribution as long as all the disturbances have this property.

Another estimation approach slightly more complicated than the above one due to Coslett (1983) is to consider all individual disturbances to have come from the same unknown distribution. Then, to derive estimates of choice probabilities as function of $V_{in} - V_{jn}$ along with estimates of β. Actually, Coslett's method is MLE when both choice models and the parameters are unknown. This method is computationally quite difficult.

After estimating a selected model, we need to perform various types of tests. Thus, we pass on to discuss the available testing procedures in the discrete choice models in the next section.

III.3 Testing Issues

Once a model has been estimated, the next relevant question arises how well the model fits to the observed data. The desirable properties of the estimation procedure are realised only when the selected model fits the observed data well. At this point we may quote, Frisch (1951)," ..... As we dig into the foundation of any economic .....model, we will always find a line of demarcation which we cannot transgress unless we introduce another type of test of significance, a test of applicability of the model itself". The large number of specification errors that affect the probabilistic choice-models and the large potential magnitudes of the resulting errors imply that diagnostic testing should play an important role in the development of these models. Model mis-

specification may lead to non-robust parameter estimates which suffer from asymptotic bias leading to erroneous inference and thus, specification tests for choice models are of interest from two points of view.

Firstly, such tests are needed to assess the impact of omitted variables on the model fit. Secondly, the empirical applications of the model necessitate the testing of the restrictive assumption like IIA property in MNL. In discussing such tests it is worth mentioning that the performance of the tests is a function of the particular model structure and data, and that testing procedures differ in their computational requirements and such computation becomes burdensome when the parameters are poorly identified.

Several statistics are available for evaluating choice model fit to the observed data (see Train, 1986, Amemiya, 1981). A scalar criteria can be used for deciding the goodness of fit of the model. Of course, many economic-theoretic and statistical factors must be simultaneously taken into consideration when one chooses a model. A list of scalar criterion is given here.

Likelihood Ratio (LR)

Sometimes a Likelihood Ratio (LR) index is used to test how well the choice model fits the available data. Basically, this statistic tells how the model with estimated parameters performs as compared to a model in which all the parameters are zero except the intercept term. This comparison is made on 1 at the estimated parameters and at zero parameters.

The statistic is given by

$$\rho = 1 - \frac{I(\beta)}{I(o)}$$

where I refers to the log of L. If the estimated parameters do no better, then $I(\beta)=I(o)$ and $\rho=0$ which is the lowest value. At the other extreme, if the model could predict decision maker's choice perfectly, then $I(\beta)=1$ since the probability of observing the choices that were actually made is one; and since $\log 1=0$, then $I=0$ at the estimated parameters. Thus, the highest value of ρ is

one. Naturally, higher values of ρ are preferred since it indicates better fit. Unlike R^2, which indicates % of variation in the dependent variable that is explained by the estimated model. LR index gives the % increase in the 1 above the value taken at zero parameters. However, two models estimated on samples that are not identical or with a different set of alternatives for any sampled decision maker, can not be compared via their LR index.

Another approach to assess the goodness of fit is "The percent perfectly predicted". For this statistic we identify for each sampled decision maker the alternative with the highest probability based on the estimated model and determine whether or not this was the alternative which was actually chosen by the decision maker. Then, we compute the proportion of the sampled decision makers for which the highest probability alternative and the chosen alternative are the same. This percent is taken as an indicator of perfectly predicted proportion. However, this statistic is based on the idea that the decision maker is predicted by the researcher to chose the alternative for which the model gives the highest probability. But, this notion is opposite to the fact that the researcher does not posses enough information to predict decision maker's choice.

Number of Wrong Predictions

There are other criteria such as Number of Wrong Predictions which are similar to the one above and are appropriate when there is an all or no loss function. One disadvantage with such criteria is that if we deal with an event which happens with a high probability or with a low probability, most models will do well by this criterion.

Sum of Squares of residuals (SSR)

Sum of Squares of residuals (SSR) is a criterion which does not suffer from the disadvantage of the previous one. But its use cannot be supported for choice

model because choice model is heteroscedastic model since the error terms don't have constant variance although it is assumed to be so.

Sum of Squares of residuals Weighted by probability is preferred to the wn-weighted SSR because, firstly, if the estimated probabilities are replaced by true probabilities, the minimisation of the above criterion with respect to the parameter, β, yields an asymptotically more efficient estimator than with wnweighted SSR, and secondly, it seems reasonable to attach a higher loss to the error made in predicting a random variable with a smaller variance since such a random variable should be easier to predict. The criterion R^2 has the same disadvantage as with SSR. An appealing model fit criterion is the Akaike (1973) Information Criterion (AIC) defined by AIC= -1+k where k is the number of parameters to be estimated and 1 is the log L. The model with the lowest AIC is to be chosen. This criterion is particularly appropriate for comparing models with different numbers of parameters.

Another information theoretic approach proposed by Hauser (1978) has been applied by Hauser and Urban (1977) and Silk and Urban (1978). Basically, in this testing approach the accuracy of the model is calculated by comparing the empirical information with the expected information, that is, for the choice set $C=[c_1, c_2,.....c_m]$ and for the explanatory variable vector x, the information I(C,X) is equal to its expected value E[I(C,X)].

As no single criterion can be considered to be the best in every situation, the researcher has to employ more than one criterion to decide the model fit. As Feinberg (1977) stated, "unfortunately there is no all purpose best method of model selection". But the law of parsimony holds that when confronted with a choice between competing explanations, the best explanation is that which adequately explains the existing data in the simplest terms.

The use of discrete choice models typically involves the analysis of data obtained from surveys. These data are likely to have outlying response values

as well as extreme values of the independent variables which may have an impact on the model fit. We discuss this issue in the next section.

III.4 DIAGNOSTICS OF INFLUENTIAL OBSERVATIONS

The estimation by the method of Maximum Likelihood (ML), which is the usual procedure adopted in discrete choice models, is very sensitive to sample outliers due to the iteration-involved in its computation. The model building process can be highly influenced by peculiarities in the data. It is necessary to detect data points that are likely to be more influential in fitting a model because inferences made from a model may depend heavily on certain response values or covariant values. Hence, diagnostic measures to detect such extreme observations and quantifying their effects are of paramount importance. Such diagnostics based upon regression diagnostics originally proposed by Cook (1977) have been advanced in the binary choice model by Pregibon (1981) and in the multinomial case by Lesaffre and Albert (1989).

After ascertaining the model fit, we need to perform test of hypothesis regarding the parameters involved in the selected model. Let us discuss some of the commonly used testing procedures in choice models.

III.5 TESTS FOR LINEAR RESTRICTIONS ON PARAMETERS

In order to test the hypothesis H_0: $A\beta=b$ where A is a $q\times k$ matrix, k is the number of elements in parameter vector and b is a vector of $q\times 1$ known constants, $q\leq k$ and q rows of A are linearly independent, one can apply Wald test statistic which, for $q=1$, becomes a squared N(0,1) test statistic. In the case when $q>1$, the likelihood Ratio test (LRT) statistic is applicable. Examples of the use of such tests are Chapman and Staeli (1982), Flath and Leonard (1979) and Malhotra (1982a).

The most important testing procedure in the area of discrete choice analysis is the test for IIA property in MNL model and we focus on the available tests in the next section.

III.6 TESTS FOR IIA PROPERTY IN MNL MODELS

The specification tests of Logit models proposed in the literature are diverse (see for example Hausman and McFadden, 1981; McFadden, Tye and Train, 1977, Engle, 1984; Ruud, 1984).

The diagnostic tests of IIA property in MNL models flourished very little till the work of Hausman and McFadden (1984). The test for the IIA property originally proposed by McFadden, Tye and Train (1977) was modified by Hausman-McFadden (1984) and by Small and Hsiao (1985). Another such test is to estimate a mother logit along with a multinomial logit model and then to test the hypothesis that all the coefficients of the extra variables in the another logit are zero. Such test can be performed by means of a LRT. If the hypothesis is rejected, then the multinomial specification is rejected.

The original testing procedure for the IIA property devised by McFadden, Tye and Train (1977) which is an asymtotic Likelihood Ratio test is defined as

$-2[L_A(\hat{\beta}) - L_A(\tilde{\beta})]$ where

$L_A(\hat{\beta})$ = Log likelihood for subset A model evaluated at MLE obtained from the full set Choice set C

$L_A(\tilde{\beta})$ = Log likelihood for the subset A model evaluated at MLE obtained from the subset A.

This test is asymptotically a χ^2_q where q is the dimension of $\tilde{\beta}$ or in other words, the number of identifiable sub vectors from A.

Small and Hsiao (1982) point out that this test statistic is not a proper LRT statistic because the assumption involved in it that $\hat{\beta}$ is nonstochastic is identical to assuming that $\hat{\beta}$ and $\tilde{\beta}$ are uncorrelated and it may not hold

good. Such an assumption results in a LRT biased towards not rejecting the null hypothesis of IIA model structure. Small and Hsiao proposed a corrected version of this test statistic. In the case a scalar difference between the covariance matrices does not exist, Small and Hsiao suggest using an exact test.

Hausman and McFadden (1981), henceforth HM, proposed two specification tests. The first is an application of the Hausman (1978) specification test to the idea proposed by McFadden, Tye and Train (1977). The idea behind the test is that multinomial specification can be tested by comparing parameter estimates obtained from the complete choice set with estimates obtained from conditional choice from a restricted choice set. If the parameters come out to close to each other, the multinomial restriction can be accepted. The computation of this test statistic is quite straight forward and can be performed with logit program.

The second test proposed is to consider the more flexible nested logit model (McFadden,1981) as an alternative model to test the original multinomial logit model. Since the multinomial logit is a special case of general nested logit model, classical testing procedures such as **Wald, LRT, LM** can be applied.

The first test proposed by Hausman and McFadden (1984) tests the hypothesis that the IIA holds against the alternative that it does not. This test gives an useful diagnostic for failure of IIA. McFadden (1987) shows that the **HM test** is equivalent to a score test.

The second test for the IIA property given by Hausman and McFadden (1984) is to embed the MNL model in a more general **Nested Multinomial Logit (NMNL)** model which exhibits a hierarchical pattern of similarities of alternatives and to perform the **LM** test for the restriction of MNL. NMNL refers to a family of response models which includes MNL and permits a pattern of similarity across alternatives. Hausman and McFadden give some exact sample properties of this test.

An LM test for IIA in MNL model taking Dogit model as an alternative was performed by Tse (1987) while such a test taking MNP as alternative was attempted by Daganzo (1970). The test statistic of Tse is a two-sided one but it can be extended to one-sided locally best mean most powerful test following King and Wu (1989).

It is recognized that MLE which is the most usual estimation procedure for discrete choice models is highly sensitive to model mis-specification and omitted variables are the important source of such misspecification. Thus, diagnostic check for omitted variables is an important test and we discuss this aspect in the next section.

III.7 TESTS FOR OMITTED VARIABLES

The gap between theory and practice can be narrowed down in two major ways. Firstly, by the continued development of robust estimation methods and testing procedures that requires fewer untenable assumptions. Secondly, by the development of diagnostic tools that identify aspects of a problem that does not conform to the hypothesized modeling process. Abnormal residuals may indicate the presence of anomalous values, which need further investigation. When the fitted model is incorrect, the distribution of the unobserved errors and hence that of the residuals will change. The study of the residuals enables one to infer any incorrect assumption concerning the error term. The residuals can be transformed to have a desired covariance structure which is needed in the tests of hypothesis. In the connection we may relevantly quote Herschel.

"Most of the phenomena which nature presents are very complicated; and when the effects of all known causes are estimated with exactness and subducted, the residual facts are constantly appearing in the form of phenomena altogether new, and leading to the most important conclusions." (Herschel, 1830).

Diagnostic tests for functional misspecification in choice models can be performed by testing the significance of an auxiliary regression of residuals on

included as well as potentially excluded variables. The Lagrange Multiplier (LM) test for omitted variables enables one to diagnose the structure of deviations from the model through testing the various plausible sources of misspecification of the model. McFadden (1987) gives an extension to the MNL model of a result obtained by Engle (1984) that a LM test for omitted variables in a binary logit can be calculated by an auxiliary regression. Both the **HM and LM tests for** IIA property have been shown to be equivalent to a test for omitted variables.

In Hausman and McFadden's second category of test a generalised form of MNL model, namely, NMNL, has also been shown equivalent to the test of omitted variables. However, we need to take into account that this test may also face a problem with the nature of nesting which becomes more problematic when the number of alternatives is excessively large.

In addition to the testing procedures discussed above sample moment conditions are used for constructing diagnostic tests for discrete choice models is provided by Pagan and Vella (1988) and the sample moment condition is given by

$\hat{\tau} = N^{-1}\sum \hat{F}_i'[\hat{F}_i(1-\hat{F}_i)](y_i - \hat{F}_i)$ where N is the number of observations,

F_i=the probability function, F_i' =the first derivative of F_i, F_i' evaluated at the MLE under the alternative choice model, y_i=1 in case of occurrence and zero otherwise. Choice of F_i' provides diagnostic tests for the probit or logit models.

Among other conditional moment restrictions for generating diagnostic tests a procedure that has seen wide application in the mode or brand choice literature is due to horowitz (1985) and it involves the comparison between the everagee predicted probability of choosing a particuslar brand with the observed relative frequency of that choice and it tests the adequacy of the model. The test statistic in such case is given by, $\hat{\tau} = N_j^{-1}\sum_{i\in I_j} \hat{F}_i - N_j^{-1}\sum_{i\in I_j} Y_i$ where

N is the sample size, i=1,2,.....N, I_J=j^{th} class of individuals, N_J= number of individuals in the j^{th} class.

Another type of test for general misspecificationis the **information matrix (IM) test** introduced by White (1982) and was adopted by matrix (see also lancaster, 1984, davivtion and mackinon, 1984). Orme (1986) provides a simple calculational procedure for IM test statistic for logit and probit models. We have so far discussed the issues like specification, estimation and testing of quantal choice models in connection with unidimensional choice. Situations also exist where we need to take account of the joint response of multidimensional quantal choices and such is the case of our choice problem. Thus, we pass on discuss such problem in the next chapter.

CHAPTER IV

WHY MULTIVARIATE CHOICE MODELS ?

IV.1 INTRODUCTORY REMARKS

When we have several jointly dependent variables each one of which is polytomous, we deal with multivariate polytomous situation of which multivariate binomial is a special case having jointly dependent variables each one of which is a dichotomous variable. In many situations we may need to analyse the structure of the relationship between response patterns of various categories or of various systems for a particular choice maker or subject with multivariate generalisation of the univariate analysis. In general, multivariate analysis is applied to study the inter-relationships of various systems considered together. A multivariate discrete choice model specifies the joint probability distribution of two or more discrete dependent variables. McFadden (1976) clarified such situation by an example of travel modes and possible destinations. Situations where multivariate modeling is necessary may include portfolio of household appliances, sequence of outcomes of a collective bargaining process, labor force participation describing a sequence of binomial decisions over time. Portfolio of meat consumption which we deal with in our present work indicates a situation where multivariate choice analysis is very much relevant and necessary. In this connection we need to focus on previous empirical works with multivariate analysis and we do so in the next section.

IV.2 SOME MULTINOMIAL QUALITATIVE CHOICE MODELS:

Types of outcomes:

1. Ordered, discrete outcomes:

    i. The number associated with each outcomes has some quantified meanings

    ii. Number of outcomes can be finite.

As for example: 0 = less than a high school degree

1 = high school degree, less than a bachelor's degree.

2 = at least bachelor's degree.

In this example, a higher value is associated with more schooling and the rang of possible outcomes is limited to three.

iii. Number of outcomes could also be unbounded

As for example: Number of health clinics in a country. In this example, the range of outcomes is the set of non- negative integers in "count" data.

2. unordered, discrete outcomes:

   a. the number associated with each outcome is used to identify the outcome but has no quantitative meaning

   b. As for example: marital status

   Occurrence of outcomes: 0= never married

   1= currently married

   2= previous married (divorced, windowed or separated)

In this case three is no quantitative significance to the numbers that are used to identify outcomes

Multinomial probit model:

i. Assume that the errors associated with each choice are jointly normally distributed with arbitrary covariance structure.

ii. Probability of observing j is described by a (j-1) variate normal distribution.

iii. For j greater than , the multivariate normal can be difficult to evaluate.

iv. Errors are not assumed to be independent across choices, the multinomial probit model does not suffer from the IIA property.

v. Identification of all variance and covariance term is also difficult.

vi. Multinomial probit model is available in a ML.

Multinomial logit model:

Error distribution:

i. Assumme the errors across choices are independently and identically distributed such that

$F(u_{ij}) = \exp(e^{-u_{ij}})$

ii.The conditional probability of observing choice j is

$$\text{Prob}(y_i=j/x_i) = \frac{e^{x_i\beta_j}}{\sum_{k=1}^{J} e^{x_i\beta_k}}$$

iii. This is the probability for the multinomial logit model.

iv. Usually normalize $\beta_1 = 0$

Coefficients:

i. The interpretation of the coefficients is complicated .The coefficients need to be interpreted as the marginal effect of the observed variable on the difference in utilities between j and 1(more generally the normalized choice).like the other qualitative models, the coefficients can not be interpreted as regression coefficients)

ii. Formula for the marginal effects is

$$\frac{dp_j}{dx_{im}} = p_j\left[\beta_{jm} - \sum_{k=1}^{J} p_k\beta_{km}\right]$$

iii. Because of the differences between the coefficients and marginal effects both should be reported.

Independence of irrelevant alternatives:

i. One properties of the multinomial logit model is that ratio of any two choice probabilities does not depend on any other choice probabilities;

$$\frac{p_{ij}}{p_{ik}} = \frac{\exp(x_i\beta_j)}{\exp(x_i\beta_k)}$$

ii. This implies that if a choice were removed, the determinants of the probability ratios of the remaining choice would be unaffected.

iil. Property results from the assumption that the errors for choices are independent of one another.

Stata code:

i. Basic code is the same as a binomial logit.

ii. By default, base category is most frequent category, can change this by specifying base category.

Binary Choice Model:

In this chapter, we discus Choice Behaviour of individual when two alternative are available and one must be chosen. The binary decision by the i^{th} individual can be conveniently represented by a random variable y_i that takes the value 1 if one choice is made and the value 0 if the other is made. Let P_i represents the probability that y_i takes the value 1 while it may be of interest to estimate the probability P_i. The 1^{st} is based on the maximization of expected utility by an individual decision maker. As some of that the utility derived from a choice is based on the attributes of the choice, which are specific to the individual decision maker, the individual's socio-economic characteristics and a random disturbance. Let U_{i1} and U_{i0} denote the utilities of the two choices, x'_{i1} and x'_{i0} vectors of characteristics of the alternatives as perceived by individual i and w'_i a vector of socio-economic characteristics of the i^{th} individual. Then, assuming linearity,

$$U_{i0} = \bar{u}_{i0} + e_{i0} = \alpha_0 + z'_{i0}\delta + w'_i\gamma_0 + e_{i0}$$

$$U_{i0} = \bar{u}_{i1} + e_{i1} = \alpha_{i1} + z'_{i1}\delta + w'_i\gamma_1 + e_{i1}$$

Thus, $y_i=1$ if $U_{i1}>U_{i0}$ and $y_i=0$ if $U_{i0}>U_{i1}$. Consequently,

$$\Pr(y_i=1)= \Pr(U_{i1}>U_{i0})= \Pr[(e_{i0}-e_{i1})<(\alpha_1-\alpha_0) +(z_{i1}-z_{i0})'\delta+w'_i(\gamma_1-\gamma_0)]$$

$$=F(x'_i\beta).$$

where, $x_i'=(1, (z_{i1}-z_{i0}), w_i', \beta'=((\alpha_1-\alpha_0), \delta', (\gamma_1-\gamma_0)')$, and F is the cumulative distribution function (CDF) of $(e_{i0}-e_{i1})$.

Note that if an individual characteristics has the same effect on expected utility, then the corresponding elements of γ_0 and γ_1 are equal and variable falls out of the model. Also the presence of an intercept employs that the choice has effects on utility a part from the attributes. The kinds of choice model one obtains depends on the choice of F. The most common choices in economic applications are:

The linear probability model: $F(x_i'\beta) = x_i'\beta$

The probit model: $F(x_i'\beta) = \int_{-\infty}^{x_i'\beta} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$

The logit model: $F(x_i'\beta) = \frac{1}{1+e^{-x_i'\beta}}$

A second motivation for such model is that it is possible to define an unobservable random index for each individuals that defines their propensity to choose an alternative. If that unobservable index is $y_i^* = x_i'\beta + e_i$ then the binary choice is defined by assuming a pdf for e_i and letting the random variable $y_i= 1$ if $y_i^* >0$ and $y_i^* \leq 0$, where the choice 0 is arbitrary. This is equivalent to assuming only the sign and not the numerical values of y_i^* is observed. Finally one may model the probability of choices directly as a function of a set of explanatory variables and by pass the utility maximization or threshold arguments .

The difficulties of using standard regression procedures when a quantal choice model is adopted are easily illustrated by a simple example. Suppose that we interested in the factors that determine whether or not an individual attend in a College a certain year. We can represent the choice of any individual by using a dummy variable, which takes the value 0 or 1 depending upon whether the individual does not or does attend College respectively. Furthermore, suppose that the decision is model to depend upon income and nothing else but random effects. Then this model can be represented as $y_i = \alpha + \beta x_i + e_i$;

i = 1, 2, ...,T

where y_i is are be a random variable that takes the values 1 if ith individual goes to college and 0 otherwise, x_i is the ith individuals income, e_i is the random disturbance and α and β are unknown parameters. If p_i is the probability that $y_i=1$ and $(1-p_i)$ is the probability that $y_i=0$ then

$E(y_i) = p_i = \alpha + \beta x_i$ if $E(e_i)=0$.

A Specification has several readily apparent problems. We want to make the usual assumption that the random variable e_i has mean 0 i.e., $E(e_i)=0$ for all i we must face the fact that while y_i can take but two values, 0 and 1 the systematic portion of the right hand side can take any value. This means that e_i can take only two values given x_i namely $\alpha + \beta x_i$ and $1- (\alpha + \beta x_i)$. Furthermore, if $E(e_i)$ is to be 0, they must takes this values with probability $\alpha + \beta x_i$ and $1- (\alpha + \beta x_i)$ respectively. Since $\alpha + \beta x_i$ can takes the values >1 or <0, the probabilities above can be >1 or <0. Given the above result, the usual assumption that $E(e_i^2)=\sigma^2$ is no longer tenable. The Bernoulli character of y_i implies a variance for e_i of $(\alpha + \beta x_i)(1- \alpha - \beta x_i)$. Since the variance of e_i depends on i, the e_i are heteroscedastic and a part from the 1st problem mention the use of OLS will result in inefficient estimates and imprec*ise* predictions if $E(y_i/x_i)$ is confined to the unit interval, prediction outside the unit interval can be produced for values of x outside the sample range even if the co-efficient estimates are dividd by minimizing the SS residuals subject to the

condition that within sample prediction lie in the unit interval. The fitted relationship will be vary sensitive to the values taken by the explanatory variables special where their bunched. The usual test of significance for the estimated co-efficient do not apply, estimated standard error are not consistent.

In this model a change in the explanatory variable has the same effect on pi whether the initial values is near 0 or near 1. In many economic applications the effects of a change in near .5. Since the y_i's are not normally distributed no method of estimation i.e. linear in the y_i's in general efficient. i.e. linear in y_i's such as LS or GLS can be improve upon.

Multinomial Choice Model

A typical quantum Choice model considers subject who face $j \geq 2$ alternatives and must choose one of these. In this sense, these models are generalizations of those presented in Binary Choice Model. Let yij be the binary variable that takes the value 1 if the jth alternative j=1,2----j is chosen and 0 othewise.

Let $p_{ij} = Pr(y_{ij}=1)$

then $\sum_{j=1}^{J} y_{ij} = \sum_{j=1}^{J} p_{ij} = 1$

and given a sample of T individuals the Likelihood function is

$l = \prod_{i=1}^{T} p_{i1}^{y_{i1}} \ldots\ldots\ldots p_{ij}^{y_{ij}}$. Each observation is assume to be drawn from independent but not identical, multinomial distribution; hence the name multinomial choice model. The number of alternatives that face different individuals might vary so that j could be indexed by i, but we will not pursue that generalization. This model is made into a behavioral one by relating the selection probabilities to attributes of the alternatives in the choice set and the attributes the individuals making the choice.

In Binary choice model can be motivated by assuming that individuals maximize utility and that the utility ith individual derives from the choice of the jth alternative can be represented as

$$U_{ij} = \overline{U}_{ij} + e_{ij} = x'_{ij}\beta + e_{ij}$$

where x_{ij} is a vector of variables representing the attributes of the jth choice to the ith individuals, β is a vector of unknown parameter and e_{ij} is a random disturbance. The presence of the random disturbance e_{ij} can be rationalized as reflecting intrinsically random choice behaviour and measurement or specification error. In particular it may reflect unobserved attributes of the alternatives. Having specified a utility function, each individual is assumed to make selection that maximized their utility. The probability that the 1st alternative is chosen is

p_{i1}=Pr($U_{i1}>U_{i2}$ and $U_{i1}>U_i$ ----------and $U_{i1}>U_{ij}$)

$$=\Pr(e_{i2} < \overline{U}_{i1} - \overline{U}_{i2} + e_{ij} \text{ and } e_{i3} < \overline{U}_{i1} - \overline{U}_{i3} + e_{ij} \text{ and } \ldots\ldots e_{ij} < \overline{U}_{i1} - \overline{U}_{ij} + e_{ij})$$

similar expressions hold for all others p_{ij}'s and the e_{ij}'s become well defined probabilities once a joint density function is chosen for the eij. It is convenient to look above equation in differenced form as

$$P_{i1}==\Pr(e_{i2} - e_{i1} < \overline{U}_{i1} - \overline{U}_{i2} \text{ and } e_{i3} - e_{i1} < \overline{U}_{i1} - \overline{U}_{i3} \text{ and } \ldots\ldots e_{ij} - e_{i1} < \overline{U}_{i1} - \overline{U}_{ij})$$

The transformation reduced by one the number of integrals that must evaluated in order to determine the p_{ij}'s. It also explained why distributions that are closed under subtraction, or produce convenient distribution under subtraction are popular candidates for the joint density of the e_{ij}.

As the formulated above, the multinomial probit model is less restrictive than the multinomial logit model, but the generality is gained at considerable computational expanse. The multinomial logit model assumes that the parameters of the random utility functions are constant across the individual in the population and that the random utilities are independent, even for alternative that are similar. The multinomial probit relaxed both of these assumptions. By adopting a random utility function of the form test parameters

are assumed to be vary normally across the population of decision makers and utility can be correlated. Hausmen and Wise (1987) assume the random parameters β_{ij} are uncorrelated across i and that the e_{ij} are uncorrelated across i and j. Thus u_{ij} and u_{ik} are correlated because the same β appear for all the alternatives. The computational procedure the adopt is feasible for a model with up to 5 alternatives. They compare the general model, and a special case that assume the variances of the random co-efficient to be zero to multinomial logit. They conclude that in their example the result of the logit and independent probit are similar. Their general model fits the best procedure significance different estimates and forecasts concerning a new alternative. Albright, Lerman and Manski's (1997) model is more general and they consider alternative procedure for evaluating the integral of the normal pdf. Their procedure can handle models with upto 10 alternative. They applied their model to the Hausmen and Wise data and concluded that the logit and probit results did not differ by much. Furthermore, they could not obtained accurate estimates of the covariance matrix of β_i, the increase the value of the log likelihood using probit rather than logit did not seem substantial relative to the decrease in d.f, and the probit model required much more computational time per iteration than did the logit model. Thus while they demonstrated the feasibility of the multinomial probit model, they could not justify its additional cost, at least for the data set they used.

In addition to literature on the choice between multinomial logit and probit, additional work has appeared on a variety of related topics. McFadden (1977,1982) introduces a non independent logit model that eliminates some of the weakness of the multinomial logit model described above. Misspecification in the contex of choice models has been considered by several authors. Lee(1982) considers the effect of omitting a relevant variable from a multinomial logit model.Rud(1983) considers the effects of applying ML specified. Parks(1980) extends the procedure of Amemiya and Nold(1975) to allow for an additional disturbance to reflect misspecification of the random

utility function in the contex of a sample with repeated observations. Fomby and Pearce (1982) present standard errors for the selection probabilities, partial derivatives and the expected response in the multinomial logit model. Zellner (1983) presents a Bayesian analysis of asimple multinomial logit model.

IV.3 EMPIRICAL EXPERIENCE WITH MULTIVARIATE ANALYSIS

Logit analysis of several qualitative variables by WLS was attempted of Grizzle, Starmer and Koch (1969). They ignored the probability structure of the error terms and thus, their results may not be too efficient in moderate or small samples. Their model has been specialised by Grizzle (1970) to a multivariate logit. Although this model is computationally attractive, it has the disadvantage that it does not generate correlation estimates for joint distribution of responses. In that case, the next recourse is multivariate probit which provides the desired correlation structure. Ashford and Swoden (1970) specialised a bivariate probit. Although they noted that extension of bivariate probit is infeasible, 3 or 4-variate multivariate probit should be manageable in this age of sophisticated computerizing facilities. Dutt and Lin (1975), Dutt (1976), Milton (1972), Nelson (1976) provide methods of computing normal orthant probabilities but these methods do not give the derivatives of these probabilities. Domencich and McFadden (1975) applied a bivariate independent logit by using special parameterization to travel mode-time situation. This parameterization is useful only when some independent variables affect only one of the binary variables. Domencich and McFadden developed a two-step estimation procedure for their model, which was again proved by Amemiya (1978) to be less efficient than MLE. Amemiya (1981) demonstrated how to use McFadden's GEV model to bivariate case. McFadden's NMNL (GEV) model is useful in a multivariate situation because the alternatives can be naturally classified according to the outcome of one of the variables such that each of the classes contains similar alternatives. GEV models have the advantage that

fewer parameters need to estimated but they have the disadvantage that the true correlation structure may be more complex than it is assumed to be.

A log-linear parameterization of four variate model was applied by Goodman (1972) to study the choice of camp of soldiers. One disadvantage with log-linear model is that the higher order interaction term is set to zero and for m variates one needs 2^{m-1} parameters. Nerlove and Press (1973) proposed to make the parameters in a log-linear model to depend on independent variables-particularly main-effects parameters. This creates certain unbearable restriction in some cases. They have developed an analysis of variance framework for multivariate multinomial logit analysis. Heckman (1975) made comprehensive treatment of simultaneous equations with quantal response analysis while Schmidt and Strauss (1975) considered the simultaneous version of multivariate logit models and Kuang, Liang and Zeger (1989) used logistic regression model for multivariate time series data. Amemiya (1974) developed binary probit model on tolerance analysis which is not applicable to random utility models. Fry (1990) developed a trivariate binomial model using Burr distribution and he has developed estimation procedure basing on difference operator proposed by Meyer (1979). This method is applicable only to the case where cdf has a closed from. Attempts have also been made to deal with multivariate analysis by deliberately ignoring the multivariate structures, see for example, Silverman and Durden, (1976), Bartel (1979).

There is fundamental difference between multivariate probit and multivariate logit or log-linear models. In the former case marginal probabilities are specified first and a joint probability consistent with marginal probabilities is found while in the latter case, joint or conditional probabilities are specified at the beginning. In the former, marginal probabilities have a simple form in the latter (see Amemiya, 1981). The univariate similarity between logistic and normal distribution disappears in the multivariate case because there is no multivariate distribution with unconstrained correlation structure for which marginal distribution is logistic (see Gumbel, 1961). Due to fundamental

differences among multivariate models researchers were interested to investigate their comparisons and one such example is Morimune (1979) who compared Ashfor-Sowden bivariate probit with Nerlov-Press log-linear models using Cox's test and its modification and his evidence supported probit by both the tests.

Several authors like Amemiya (1981), McFadden (1976) have hinted that multivariate problem can be treated as one of univariate multinomial response problem. But multivariate choice situation may pose further complicacies and considerations. Specification of probabilities by taking into account of special feature of a multivariate model should concern us and the specification of probabilities differs in the case of multivariate and converted multinomial models. There is really the need for developing practical models for selection probabilities which exhibit multivariate structure and there has been little attention paid to work on estimation and testing for such structural models. We note that our approach in the present work was never followed before as per hour knowledge is concerned. Our next task is to construct our desired model and we do this in the next chapter.

CHAPTER V

CONSTRUCTION OF A MULTIVARIATE BINOMIAL (MVB) CHOICE MODEL

V.1 INTRODUCTORY REMARKS

In a multidimensional choice situation we should take the fact into account that in consuming, the consumers are likely to exhibit tendency towards considering interdependencies among marginal utilities of choice items. It is also plausible that consumers exhibit portfolio behavior, which means that consumers make a combination of one or more consumption items disregarding their interdependencies. In this chapter we are going to address these two issues in terms of model specification, estimation and tests. Before we pass on to the construction of the MVB choice model for our purpose, we feel it necessary to present general interpretation of a MVB choice situation and we do it in the next section.

V.2 BASIC IDEA BEHIND MULTIVARIATE BINOMIAL CHOICE

When we face a situation in which multidimensional choices determine the consumption pattern, which is true in our case, we need to analyse the joint response of several dependent variables. The term multidimensional or multivariate choice indicates a related set of choices each of which is either multiple or binary.

Let us consider the general multivariate polytomy prior to considering the specific MVB case. Suppose, we have a vector of m variables, say, $y=\{y_1, y_2, ...y_m\}$which may take any one of $I_1, I_2, ...I_m$ unordered alternative values. We intend to analyse the joint response of these variables by relating the joint probability to some explanatory variables. Finney (1964, P325) has discussed the statistical analysis of multiple response variates and he concluded

that a direct approach to the problem is very laborious and tedious. Thus, attempts have been made to convert multivariate quantal data to a univariate quantal form.

In our example mentioned above, we have $Q=\prod I_i$ possibilities. Now, for the case of two dichotomous variables we have $Q=2^2=4$ possibilities. It is evident to us that the m-variate polttomy with polytomies of order I_1, I_2, ...I_m is equivalent to a univariate polytomy of order Q. The probabilities that some polytomous variable z_k takes on the 1^{st}, 2^{nd}, ... or the Q^{th} value are given by

$P_1=P_{1,1,\ldots\ldots\ldots 1}$

.

.

.

$P_Q = P_{I_1 I_2 \ldots\ldots\ldots I_m}$

Conversely, any polytomy can be written in terms of a multivariate dichotomy. The probabilities for a set of m jointly dependent binary variables may always be re-expressed in terms of a single polytomy with 2^m values and any polytomy can be expressed in terms of joint occurrence of binary variables. Of course, we may need to impose some restrictions on the probabilities since the polytomy may not always be of order exactly a power of 2. Thus, the next higher power of 2 must be used and probabilities of certain combinations need to be set to zero since total probability must be unity (see Nerlove and Press, 1973 for further details).Now, turning to the MVB case, we have the observed vector Y each element of which, say, y_i takes 1 if the corresponding unobserved latent variable y_i^* exceeds some threshold value and 0 otherwise, where the latent variable is expressed as a function of some known explanatory variables (X_i) and parameters (β_i) to be estimated plus some random error term.Thus, we can write $y_{in}^*=X'_{in}\beta_i+e_{in}$ where e is the random error term and the vector X may include the alternative specific attributes as well as choice maker specific

characteristics and the parameters may vary across alternatives but not across choice makers because in that case we can not identify all the parameters. We form an indicator variable

$y_i = \begin{cases} 1 \\ 0 \quad otherwise \end{cases}$ if $y_{in}^* = X'_{in}\beta_i - e_{in} > 0$ where i=1,2, ...m, n=1,2,.....N.

In total we shall have $M=2^m$ vectors only one of which is applicable to a particular choice maker. The probability of the ith choice is $p(y_i=1)=p(y_i^*>0)$ where y_i^* indicates the relative desirability of $y_i=1$ as compared to $y_i=0$ and the multivariate choice probability is given by $p(y_1=1$ or 0, $y_2=1$ or 0, ...$y_m=1$ or 0). The proportion of choice makers for whom Y is optimal is provided by the probability p(Y). Now, we are in a position to consider the construction of our desired model and we do it in the next section.

V.3 CONSTRUCTION OF A MULTIVARIATE BINOMIAL CHOICE MODEL

It is important to note at the beginning that in building a choice model we may follow two paths: firstly, we may follow a path from the theory (economic) of choice behavior to the properties of latent variable models and the structures of their associated probabilities. Secondly, we may select a response probability model and then try to establish that the model meets sufficient conditions for deriving from the theory of optimisation. In connection with utility-maximisation behavior it is also relevant to mention here that the conditions for a response probability model to be derivable from a population of utility maximisers are sufficiently restrictive since the conditions of response probabilities for a single optimizer may not be consistent with those of the entire population. We do not go into the details of that debate and we follow the first path in our present work. We form a latent variable as a function of deterministic and random components to indicate the maximum utility derived from making a particular choice. Finally, we note that we put strict restriction on the observed part of utility, that is, we assume that it is a linear-in-parameters function where the expected utility with respect to unobservable of the indirect utility function for a

given alternative does not possess all the properties of an indirect utility function (see McFadden, 1981).

We consider a choice set containing 4 alternative choices viz, C_1, C_2, C_3, C_4 and we assign 1 for a particular alternative if it is chosen by a particular choice maker, and 0 otherwise. A particular choice maker may choose any number of alternatives from the choice set and thus we have a 4-variate dichotomous situation. This 4-variate dichotomy results in 16 possibilities which can be converted to a univariate (V) polytomy of order 16 with $v_1=(1,1,1,1)$, $v_2=(1,1,1,0)$, ...$v_{16}=(0,0,0,0)$, such that an individual chooses one of these 16 values according to multinomial law giving rise to $0 \leq \sum_j Y_j \leq 4$ where y_j indicates the j^{th} element in v_i, i=1,2,.....16. We can demonstrate this situation schematically as follows:

This scheme illustrates that all the possible combinations of values attached to the 4 variables each one of which is binary give rise to a single variable V having 16 values. Each one of these 16 values does not carry equal importance for a particular individual and the choice of one of them depends on the 'worth' attached to it by a particular individual. Only one of this 16 values applicable to an individual. Now, an individual may consider the interdependencies among the elements present in choosing one of v,s and thus, we should consider estimating the joint response probabilities and in doing so we need to adopt multivariate threshold approach.

Alternatively, an individual may disregard the interdependencies among consumption items and may consider the choice of one of 16 possible values governed by multinomial law. Thus,we may consider to modelling the same consumption behavior for a prticular data set from two points of view which we discuss below.

Individual choice maker is assumed to maximise utility from an alternative and thus, the formulation of the utility funtion is of central importance in individual choice modelling.The question of representing the sources of utility derived

from an alternative is resolved by assuming that an individual defines the utility in terms of alternatives and the characteristics of choice makers. The preferences of choice makers in the population can be described by a combination of a nonrandom or deterministic component which reflects representative tastes, and a random component which reflects idiosyncrasies in tastes of choice makers. Thus, in the first case, we define an underlying latent variable.

$y_{in}^{*}=X'_{in}\beta_i - e_{in}$ to indicate the maximum level of the indirect utility attainable from the ith choice by nth choice maker where, i=1,2,3,4, n=1,2, ...N,

x= vector of explanatory variables,

β=parameter vector which is supposed to vary across alternatives.

In the case of multivariate choice situation, we do not compare the utility of one choice item with another, rather we consider the maximum utility attainable from a particular item. Thus, we adopt the "threshold" approach since an individual will choose a particular item if its underlying perceived utility exceeds some threshold which, without loss of generality, can be taken to be zero. Here, we presume that an individual looks at the utility associated with each choice at first and then forms a portfolio. We assume that the parameters may vary across alternatives of the choice set since a particular individual with a given set of characteristics will have different propensity to consume different choices or in other words, we try to capture the differential impact of explanatory variables on different choices. We assume constancy of the parameters (β) over choice makers which means that only the coefficients and not the level of attributes of alternatives are independent of n. We want to set a link between stochastic factors surrounding agent decision-making and the structure of response probabilities. Thus, a particular individual will pick up (we supers the subscript n),

C_1 if $y^{*}_1= X'\beta_1 - e_1>0$ or $e_1< X'\beta_1$,

C_2 if $y^{*}_2= X'\beta_2 - e_2>0$or $e_2<X'\beta_2$, ...(1)

C_3 if $y^*_3 = X'\beta_3 - e_{\backslash 3} > 0$ or $e_3 < X'\beta_3$,

C_4 if $y^*_4 = X'\beta_4 - e_4 > 0$ or $e_4 < X'\beta_4$,

For interdependent case, we need to model explicitly the joint probability and the probability that a particular individual will choose all the choices can be obtained once we complete the multivariate model by specifying the joint probability $p(y_1=1, y_2=1, y_3=1, y_4=1)$ which is equivalent to

$$p(v_1) = \int_{-\infty}^{x'\beta_1} \int_{-\infty}^{x'\beta_2} \int_{-\infty}^{x'\beta_3} \int_{-\infty}^{x'\beta_4} f(.)de_1 de_2 de_3 de_4 \quad \text{.....(2)}$$

where, f(.) is the 4-dimensional density function chosen for the error term.

Here, we have structural dependencies arising from substitutability or complementarily relation among choices, We assume that utilities from different choices are interdependent and there is commonality of attributes of different choices. Thus, we need to have a joint distributional assumption of the error term. Each of the 16 probabilities characterizes the joint occurrence or non-occurrence of the four dichotomous variable y_1, y_2, y_3, y_4 and are represented parametrically. This probabilistic model can be used to predict the probability that a particular new individual with the same set of characteristics but not included in the sample will consider the dependencies among the choices in the same way as considered by individuals included in the existing sample. Once we have estimated the joint probabilities we in a position to estimate the marginal and conditional probabilities as well which are necessary for marginal and conditional predictions.

The next task we are faced with is the distributional assumption about the error term. The normality assumption has the advantage that it provides a good approximation to many multivariate distributions and any linear function of normal variates is also normal. The normality assumption also provides the estimates of full correlation structure which is not provided by many other distributional assumptions such as logistic distributional assumption.

For a typical choice maker n, let $p(v_i) = p_{in}$ and the corresponding likelihood function for a sample of size N is $L=\prod_{i}\prod_{n} p_{in}^{y_{in}}$ where y_{in} is the response indicator satisfying

$$y_{in}=\begin{cases}1\\0 \quad otherwise\end{cases}$$ if i is chosen by n .....(3)

and the log likelihood function

$$I=\sum_{i}\sum_{n} y_{in}\log P_{in}$$

$$==\sum_{i}\sum_{n} y_{in}\log\left[\int\int_{ARI}\int\int f(.)de_1de_2de_3de_4\right]$$

Where ARI Indicate the Appropriate Ranges of Integration.

i=1, 2, ...16, n=1,2,.....N.

Note that the ranges of each of 4-dimensional integration in the likelihood function will vary over i and a typical one is given by equation (2) of this section.

The associated score vector is $\partial I/\partial\beta_k = \sum\sum y_{in}\frac{\partial p_{in}}{\partial\beta_k}.\frac{1}{P_{in}}$ .....(4)

where, k=1, 2, 3, 4.

Note that under the assumption of normal distribution of the error term the ranges of integration in the above expression involve the parameters to be estimated.

In the second situation, we consider V_1, V_2, ...V_{16} as 16 bundles of goods or 16 portfolios and we intend to compute the probability that an individual will choose one of them which globally maximises the utility and in that case we disregard the individual element present in a particular v_i. This becomes a multinomial choice problem and thus we are in a position to utilise the standard known results for MLEs and their properties (see Amemiya, 1985). But, one

drawback of this conversion is that $M=2^m$ increases rapidly with m and makes multinomial problem quite intractable. In this case our parameter vector should be different from that of the previous case since in this case parameters are no more associated with a single item but with a portfolio and so is the random error term.

We assume the new parameter vector $\gamma_i = (\gamma_{i1}, \ldots\ldots\ldots \gamma_{ik})$ where k is the number of explanatory variables, are identical across choice makers but may vary across alternatives and we denote the random error component by u. Thus, under the assumption that the utilities of alternatives are not certain, but rather are random variables determined by specific distribution, we have the latent variable representation

$$\left.\begin{aligned} y_1^* &= X'\gamma_1 + u_1 \\ y_2^* &= X'\gamma_2 + u_2 \\ &\cdot \\ y_{15}^* &= X'\gamma_{15} + u_{15} \\ y_{16}^* &= X'\gamma_{16} + u_{16} \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

We have a multinomial problem of choosing one vector v_i from 16 possible vectors. Here, we have a maximisation argument and thus, we define a response indicator variable

$$yi \begin{cases} = 1 \\ = 0 \quad otherwise \end{cases} \text{ if } y^*_i = \max \{y^*_1, y^*_2, \ldots\ldots\ldots\ldots y^*_{16}\}$$

In this case, the response probabilities indicate the proportion of choice makers which optimise the utility at ach of 16 portfolios and y^*_i is the maximum utility for portfolio i obtained by optimising all other dimensions of choice.

It may appear that we can apply the MNL model for estimating selection probabilities of choosing one of 16 possibilities but unfortunately it is not appropriate in our present case because IIA property does not hold good due to

overlapping attributes (see section II.2). We may apply the MNP model which relaxes the assumption of IID for the error term and allows for covariance structure of the alternatives and thus, under the assumption that $u=(u_1, u_2, ...u_{16})'$ follows multivariate normal distribution the probability that v_1 is preferred to others is given by

$$p(v_1) = p(y^*_1>y^*_2, y^*_1>y^*_3, ...y^*_1>y^*_{16})$$

$$= p(X'\gamma_1 + u_1 > X'\gamma_2 + u_2 .....X'\gamma_1 + u_1 > X'\gamma_{16})$$

$$= p(\varepsilon_{1,2} < X'\gamma_1 - X'\gamma_2, ..., \varepsilon_{1,16} < X'\gamma_1 - X'\gamma_{1,16})$$

$$= \int_{-\infty}^{X'\gamma_1 - X'\gamma_2} ..... \int_{-\infty}^{X'\gamma_1 - X'\gamma_{16}} f(\varepsilon_{1,2}, ...\varepsilon_{1,16}) d\varepsilon_{1,2} .....d\varepsilon_{1,16} .....(6)$$

where f(.) denotes a 15-dimensional density and thus, we need a 15-dimensional integration for each one of the probabilities $p(v_1)$, ...$p(v_{16})$.

In this case, the probabilistic model can enable us to predict the probability that a new individual with the given set of characteristics will pick up a particular portfolio out of the 16 available portfolios.

For a typical household n, let $p(v_i)=p_{in}$ and the likelihood function for a sample of size N in this case is

$$L=\prod_i \prod_n p_{in}^{y_{in}}$$

and the log-likelihood function is given by

$$I=\sum\sum y_{in}\left[\int_{ARI}\int.....\int f(.)d\varepsilon\right] .....(7)$$

i=1,2, ...16, n=1,2, ...N.

The associated score vector is given by

$$\partial I / \partial \gamma_1 = \sum\sum \frac{y_{in}}{P_{in}} . \frac{\delta p_{in}}{\delta \gamma_i}$$

The two models constructed above can be used to analyse the same choice behavior but there exist some remarkable differences between the two. The traditional argument that the statistical properties of MN analysis are valid when we convert multivariate analysis to multinomial one is not always practical and is dubious. We discuss the plausible differences between the two approaches in the next section.

V.4 HOW DOES MULTIVARIATE MODEL DIFFER FROM MULTINOMIAL ONE?

Although a multivariate problem can be converted to a multinomial one and vice-versa, in practice several differences appear between the two.

First, of all ,it is clear from our previous discussion that by converting multivariate case to multinomial one we basically attempt to model the same choice behavior from two points of view. In the multivariate case we postulate the choice probability on the assumption that an individual at first considers the utility associated with each choice separately and then forms the portfolio i.e., we introduce randomness at first and the forms a portfolio. From the joint probability of multivariate case one is able to estimate the marginal probabilities also which are sometimes useful in reducing the number of alternatives. From the joint probabilities one can derive conditional probabilities also which may be useful in conditional predictions. If the estimates of conditional probabilities are estimated through the joint probabilities, we can gain in statistical efficiency and can be less sensitive to specification errors (see Ben-Akiva, 1974). This is because we can incorporate restrictions across conditional probabilities and thereby use more information to estimate some coefficients in joint probability estimates.

Now, for a better explanation of the link between the two models let us consider a specific case of finding the probability of choosing meat category 1 i.e., $p(y_1=1)$. In the multivariate situation $y_1=1$ occurs in 8 of 16 vectors each one having 1 as the first element. Thus,

$p(y_1=1)=p(v_1)+p(v_2)+ \ldots\ldots +p(v_8)$

$$= \int_{-\infty}^{X'\beta_1}\int_{-\infty}^{X'\beta_2}\int_{-\infty}^{X'\beta_3}\int_{-\infty}^{X'\beta_4} f(.)de_1de_2de_3de_4 + \ldots. + \int_{-\infty}^{X'\beta_1}\int_{X'\beta_2}^{\infty}\int_{X'\beta_3}^{\infty}\int_{X'\beta_4}^{\infty} f(.)de_1de_2de_3de_4 \ldots..(8)$$

= sum of eight integrations each one of which involves 4-dimensional integration.

In the multinomial case the same probability is given by,

$p(y_1=1) = p(v_1) +p(v_2)+ \ldots\ldots + p(v_8)$

$$= \int_{-\infty}^{X'\gamma_1 - X'\gamma_2} \ldots\ldots\ldots\ldots \int_{-\infty}^{X'\gamma_1 - X'\gamma_{16}} f(\varepsilon_{1,2}, \ldots\ldots\ldots \varepsilon_{1,16}) d\varepsilon_{1,2} \ldots\ldots d\varepsilon_{1,16}$$

+

.

.

.

+

$$= \int_{-\infty}^{X'\gamma_8 - X'\gamma_1} \ldots\ldots\ldots\ldots \int_{-\infty}^{X'\gamma_{81} - X'\gamma_{16}} f(\varepsilon_{8,1}, \ldots\ldots\ldots \varepsilon_{8,16}) d\varepsilon_{8,1} \ldots\ldots d\varepsilon_{8,16} \ldots\ldots\ldots\ldots (9)$$

=sum of 8 integration's each one of which involves 15-dimensional integration.

Is there a straight forward link between β's and γ's?

The answer is not obvious.

When we convert a multivariate case to multinomial one we assume that a particular individual considers the utility derived from a particular portfolio and then selects a particular portfolio of its choice and with multinomial probabilities we are not in a position to derive conditional and marginal probabilities and coefficients although we may need them in certain circumstances. Besides that, the two approaches differ in the interpretations of the parameter estimates. If we look at the multivariate probability structure in equation (2) of the present chapter we notice that the parameters β's in this case are related to the propensity of consuming each meat category separately while in the second

case the parameters γ's refer to the propensity to consume a particular portfolio.

Similarly, if we consider the random error terms in two cases, it is evident to us that e_i indicates the peculiarities or oddities of a particular individual with respect to the choice of ith choice. But μ_i indicates the peculiarities in choice behavior with respect to the ith portfolio. Thus, if we consider the particular choice, say , v_1 in both cases, it leads us to ask a question such as whether $\mu_1=e_1+e_2+e_3+e_4$. But no explicit relation between e and μ is plausible.

Moreover, in a multivariate problem we have a covariance structure among 16 portfolios introduced by the common appearance of different choices in more than one portfolio.

Besides that, we observe that it is not possible to construct a specific test to test whether one model is nested in the other.

Thus, we may conclude that although multivariate problem can be easily converted to multinomial one, in empirical applications we need to distinguish between the two.

It is quite obvious to us that the probability statement for the same event given in (2) and (6) are different in their formulations, dimensions and range of integration. So, we face the problem of deciding upon the appropriate estimation technique and interpreting the results so obtained in two cases. We discuss the appropriateness of different estimation technique and interpreting the results so obtained in two cases. We discuss the appropriateness of different estimation techniques for our problem in the next section.

V.5 ASPECTS OF ESTIMATION AND TESTS ASSOCIATED WITH MULTIVARIATE BINOMIAL CHOICE MODEL

An important aspect of any behavioral model involving statistical estimation is the treatment of unobserved determinants of individual decision. It is quite clear that under normality assumption for the error term, we have a very complex mathematical expression and computationally it is very intricate for

explicit expression of the parameter estimates in both (2) and (6) above. In order to estimate the multivariate probabilities in (2) we may consider 4-dimensional integral or we may also consider the method of difference operator developed by Meyer (1979, ch. 10).

Although in the case of converted multinomial the estimation poses a very complicated task and it involves higher dimensional integration than before, the recent development of an algorithm by McFadden (1989) for MNP involving 5-20 alternatives can lessen our computational burden. McFadden (1989) has developed the Method of Simulated Moments (MSM) as a modification of the classical method of moment estimates for discrete response models which replaces the response probabilities with estimators obtained by Monte Carlo simulation. Several variants like, simple frequency simulator, polynimial kernel, Stern frequency simulator, conditional chi-square frequency simulator have been proposed.

Lerman and Manski (1981) suggest a Monte Carlo procedure for estimation probabilities with large m for MNP but we need to face with impractical number of Monte Carlo draws which is a disadvantage for this procedure. Daganzo (1980) has developed an MLE method for MNP model using a normal approximation to maxima of normal variates suggested by Clark (1961). This approach has the disadvantage that accuracy of the approximation cannot be refined with increasing sample size and for unequal variances of the components the method is inaccurate.

Now, for our present purpose we plan to consider the second situation, i.e., to introduce randomness in the utility function after converting multivariate to multinomial and to exploit the method of simulated moments developed by McFadden although it may strike some one that instead of reducing the dimension of integration we are attempting to enlarge it to estimate the probabilities. The reason behind favouring such approach is the recent development of the algorithm for the conditional chi-square simulator combined with the Deak construction of antithetic directions by McFadden which seems to

be very fast and very accurate. Here we present a brief description of this simulation method.

Let, c= {1, 2, ...,m} to be a set of mutually exclusive and exhaustive alternative.

A latent variable model for response from c is defined by

$$U_i = \alpha x_i ; i \in c$$

Where, α is a row vector of individual weights

x_i is a column vector of measured attributes of alternative.

Response I-is observed if $u_i \geq u_j \ for \ j \in c$

Also, let d_i- denote a response indicator =1(one), for observed response

= 0 (zero), otherwise.

In economic application, the latent variable u_i often have the interpretation of utility or profit and $p_c(i/\theta, X_c)$ is the choice probability for a population of observed characteristics of the alternatives and of the decision-makers, with αx_i interpreted as an approximation to a general economic function of observed characteristics and of the deep parameter θ. Alternative–specific dummy variables may be included in x_i; the associated components of α can be interpreted as alternative- specific additive disturbances.

let n= 1,2,------,N index a random sample from the population, yielding observations (d_{c_n}, x_{c_n}) with d_{c_n} = (d_{1n},-----, d_{mn}) and x_{c_n} =(x_{1n}, x_{mn})

The log- likelihood function of the sample is

$$L(\theta) = \sum\sum_{n=i \ i \in c} d_{in} \ I_n P_c(i/\theta, X_c) \quad .....(1)$$

The associated score is

$$\frac{\delta L(\theta)}{\delta \theta} = \sum\sum w_m [d_m - p_c(i/\theta, x_{c_n})] .....(2)$$

where, $w_{in} = \delta \ln p_c(i/\theta, x_{c_n}) / \delta\theta$ ..... (3)

Equation (2) is derived using the identity,

$$0 = \sum_{i \in C} \partial P_C(i|\theta, X_C)/\partial\theta = \sum_{i \in C} [\partial \ln P_C(i/\theta, X_C)/\partial\theta] P_C(i|\theta, X_C) \quad \text{.....(4)}$$

The primary impediment to practical maximum likelihood estimation of θ for the MNP model is computation of the (m–1) dimensional orthant probabilities for $u_{C\text{-}i}$ to obtain $P_C(i/\theta, X_C)$. Direct numerical integration is practical for $m \leq 4$ using a method of Owen (1956), modified by Hausman and Wise (1978), or expansion due to Dutt (1976). Otherwise, unless α has a factor –analytic covariance structure with less than four factors, it is usually impractical to carry out the large number of numerical integrations required to iteratively solve (2). Lerman and Manski (1981) suggest a Monte Carlo procedure for estimating $P(i/\theta, X_C)$ that can be applied to MNP models with large m, but find that it requires an impractical number of Monte Carlo draws to estimate small probabilities and their derivatives with acceptable precision. Daganzo (1980) has developed approximate maximum likelihood estimators for MNP using a normal approximation to maxima of normal variates suggested by Clark (1961). This approach has the drawbacks that the accuracy of the approximation can not be refined with increasing sample size and the method can be inaccurate when components have unequal variable (Horowitz, Sparmonn, and Dagango, 1981).

V.6 COMPUTATIONAL ISSUES AND STATISTICAL EFFICIENCY

Practical use of the MSM estimator requires that the Monte Carlo simulation of the probabilities and their derivatives be practical and well behaved that easily calculated moderately efficient instruments W be available that iterative algorithms to compute the

estimators be fairly stable and efficient and that estimators for the asymptotic covariance matrix of the estimators be computable.

The conventional method of moment estimator of a k×1 parameter vector θ (with domain θ) in the discrete response model p_c (i/θ , X_c) satisfies

$$\theta_{mm} = \arg\min_{\theta} (d - p(\theta))' W' W (d - p(\theta))$$

where d—p (θ) denotes the mN×1 vector of residuals $d_m - p_c(i)\theta, X_{cn})$ stacked by observation and by alternative within observation, and where W is a K×mN array of instruments of rank K≥k. The instruments may depend on θ, but are evaluated at some fixed θ_0 in forming first –order conditions for solution of (4). The instrument array (3), evaluated at θ*(or at a consistent estimator of θ*), yields a method of moments estimator asymptotically equivalent to the maximum likelihood estimator forθ, and hence asymptotically efficient. If computation makes exact calculation of the efficient instruments impractical, (3) nevertheless provides a template for instruments that with relatively crude approximations to p and its mN×k array of derivatives p_θ will yield moderately efficient estimators. In the remainder of this section, will assume W is a given fixed instrument array.

Under mild regularity assumptions, sufficient conditions for classical method of moments estimation to be CAN are following:

(i) The instruments T are asymptotically correlated with the score; i.e, the array $\bar{R}$ = Lie $N^{-1}wp_\theta(\theta^*)$ is of maximum rank.

(ii) The conditional expectation of the residual d-P (θ), given the instruments, is zero if and only if $\theta = \theta^*$.

The **Method of Simulated Moments** (**MSM**) avoids the computation of p (θ) required for (6), replacing it with a simulator $f(\theta)$ that is (asymptotically) conditionally unbiased, given W and d, independent across observations, and well behaved in θ. An example frequency simulator calculated from the latent variable model by independently drawing, for each observation, one or more vectors η from the density g (η), and then for any trialθ calculating $U_{in}=\alpha(\theta,\eta)x_{in}$, and counting the frequency with which the u_{in} for each alternative is maximized. The MSM estimator is given by any argument θ_{sm} satisfying,

$$(d - f(\theta_{sm}))'W'W(d - f(\theta_{sm}) \leq \inf_{\theta\in\theta}(d - f(\theta))WW(d - f(\theta)) + 0(1)$$

This definition assures the existence of a MSM estimator even if the in minimum cannot be attained.

Sufficient conditions for the MSM estimator to be CAN are given formally. They involve the same regularity assumptions and conditions on instruments as classical method of moments estimators and in addition place two critical restrictions on the simulator $f(\theta)$

(iii) The simulation bias $B(\theta)= N^{-1/2}W(Ef(\theta)-P(\theta))$, where the expectation is conditioned on W and d, is either zero, or else satisfies.

$$\sup_{\theta \in \theta} |B(\theta)| = 0(1)$$

(iv) The simulation residual process $\xi(\theta) = N^{-1/2}W(f(\theta)-Ef(\theta))$ uniformly stochastically bounded and equi-continuous in θ;

i.e

$$\sup_{\theta \in \theta} |\zeta(\theta)| = O_p(1)$$

$$\sup_{\theta \in \theta} |\zeta(\theta)-\zeta(\theta^*)| = o_p(1),$$

where, for a given $\delta>0$, $A_N= \{\theta | N^{1/2} \|\theta-\theta^*\| \le \delta\}$.

If $f(\theta)$ is an unbiased simulator of $p(\theta)$ for θ all and random numbers used to compute $f(\theta)$ are independent of the observed responses of and any simulation used in the construction of the instrument array, then (iii) is satisfied. The simulation residuals $\zeta(\theta)$ are by construction the normalized sum over observations of independent identically distributed terms that are independent of d and uniformly bounded, with $E(\zeta(0)/W)=0$ for each 0. Then $\zeta(\theta)$ is an empirical process in θ that by a standard central limit theorem is pointwise asymptotically normal. Elementary arguments establish that if f(θ) is smooth for θ in a compact domain then can also be satisfied by simulators that

have "well-behaved" discontinuities, such as the simple frequency simulator. To satisfy a simulator must avoid ``Chatter" as 0 varies; this will generally that the Monte Carlo random numbers used to construct $f(0)$ not be redrawn when 0 is changed.

CHAPTER VI

EMPERICAL RESULTS

V1.1 INTRODUCTION

In this chapter we present and discuss our estimation results obtained from MSM procedure. For estimation purpose we have used data set retrieved from Australian household survey archive. We had to develop an algorithm and used LIMDEP package for the purpose. However, we have estimated the model after performing cluster analysis of the data. So at the start we provide conceptual framework of cluster Analysis.

VI.2 ASPECTS OF CLUSTER ANALYSIS

There are two main reasons for the widespread application of cluster sampling. Although the first in tension maybe to be use the elements as sampling units, it is found in many surveys that it would be prohibitively expensive to construct such a list .In many countries there are no complete and up-to-date lists of the people, the houses, or the farms in any large geographic region. From maps of the region, however, it can be divided into a real units such as block in the cities and segments of land with readily identifiable boundaries in the rural parts. In the united states these clusters are often chosen because they solve the problem of constructing a list of sampling units.

Even when a list of individual houses is available, economic considerations may point to the choice of a larger cluster unit. for a given size of sample ,a small unit usually gives more precise results than a large unit.

A rational choice between two types or sizes of unit may be made by the familiar principle of selecting the unit that gives the smaller variance for a given cost or the smaller cost for a prescribed variance.

Cluster analysis has been employed as an effective tool in scientific inquiry. One of its most useful roles is to generate hypothesis about category structure. An algorithm can assemble observations into groups which prior misconceptions

and ignorance would otherwise preclude. An algorithm can also apply a principle of grouping more consistently in a large problem than can a human. Such methods can be particularly valuable in cases wherein the ultimate outcome is something like "Eureka! we never would have thought to associate X with A, but once you do the solution is obvious." A series of interconnected hypothesis may suggest models for the mechanisms generating the observed data. Put another way, cluster analysis may be used to reveal structure and relations in the data. It is a tool of discovery.

A most novel and unique application of cluster analysis is described by Bartels et al (1970). These authors take quite literally the role of cluster analysis as being one of discovering "natural classes."

The uses of cluster analysis described so far in this section are rather general in nature and may be applicable in almost any context. Adding another dimension to the discussion, it is useful to undertake a broad review of the various study disciplines where clustering has been used and to consider some of the entities subjected to cluster analysis. The many fields of study are grouped below into six major areas to help the reader organize this great variety of applications.

i. In the life sciences (biology, botany, zoology, entomology, cytology, microbiology, ecology, paleontology) the objects of analysis are life forms such as plants, animals, insects, cells, microorganisms, communities of interdependent members and the fossil records of life in other times. The operational purpose of the analysis may range from developing complete taxonomies to delimiting the subspecies of a distinct but varied species.

ii. Closely related to the life sciences are the medical sciences (psychiatry, pathology and other specialties focusing on clinical diagnosis) where the objects of a cluster analysis may be diseases, patients (or their disease profiles), symptoms and laboratory tests. The operational emphasis here is on discovering more effective and economical means for making positive diagnoses in the treatment of patients.

iii. The behavioral and social sciences (psychology, sociology, criminology, education, anthropology, and archeology) have provided the setting for an extraordinary variety of cluster analysis applications. The following entities have been among the many objects of analysis: training methods, behavior patterns, factors of human performance, organizations, human judgements, test items, drug users, families, neighborhoods, clubs and other social organizations, criminals and crimes, students, courses in school, teaching techniques, cultures, languages, artifacts of ancient peoples, and excavation sites. Factor analysis traditionally is a strong competitor to cluster analysis in these applications.

iv. The earth sciences (geology, soil science, geography, regional science, remote sensing) have included applications of cluster analysis to land and rock formations, soils, river systems, cities, counties, regions of the world and land use patterns.

v. Some engineering sciences (pattern recognition, artificial intelligence, systems science, cybernetics, and electrical engineering) have provided opportunities for using cluster analysis. Typical examples of the entities to which clustering has been applied include handwritten characters, samples of speech, finger prints, pictures and scenes, electrocardiograms, wave forms, radar, signals, and circuit designs. Application in engineering have been relatively few in number to date; any marked increase in the use of cluster analysis in this realm probably will require some dramatically new problem formations.

vi. The information, policy and decision sciences (information retrieval, political science, economics, marketing research, operations research) have included application of cluster analysis to documents and to the terms describing them, political units such as counties and states, legislators, votes on political issues, such issues themselves, industries, consumers, eras of good (and bad) times, products, markets, sales

programs, research and development projects, political districting, investments, personnel assignments, credit risks, plant location and floor plan designs. This realm of study probably is the source of the most innovative applications of cluster analysis in the last few years.

The literature of cluster analysis is both voluminous and diverse. Significant work may be found in almost any field of endeavor.

Literature review on Cluster Analysis

We are interested in homogeneous clusters, which indicate high internal coherence and external non-coherence. Thus, we need an appropriate clustering method, which can provide us with more homogeneous clusters of consumers. Now, one point is in order and that is the appropriateness and computational feasibility of the clustering method we choose for our purpose. We believe that a hierarchical clustering method is a suitable candidate for such purpose. This procedure has the advantage and attraction that it does not need a pre-specified distance between objects to be clustered as in necessary in a non-hierarchical clustering procedure. It sometimes becomes too hard a task to large data set like ours. So, we adopt the hierarchical clustering method and avoid all complex problems associated with non-hierarchical procedures.

Although hierarchical method is mainly suitable for studying evolutionary processes, it is equally good for other fields of research (Lorr, 1983) for mixed data sets which is our case. We note that we do not need a strict natural clustering as a biologist does.

Our related issue arises regarding the choice of an appropriate hierarchical clustering method. Here, we need to consider the computational feasibility of the method we apply. We have concerned with two aspects of the method we choose viz,

1) Performance of the method in forming homogeneous clusters
2) computational feasibility.

More than one clustering method may provide a more extensive understanding of the data structures and enable us to identify whether each method is equally sensitive to the special features of the data. Thus, if two or more methods provide identical results, we may choose one for our purpose. An elegant discussion on different clustering methods are available in Anderberg, 1973; Hartigan, 1975; Everitt, 1983; Funkhauser, 1983.

Now average linkage and Centroid methods are easily manageable computationally by using algorithm(s) in existing package like SAS. So, we have decided to use these two methods for our purpose and have used the algorithm available in the SAS package.

Some issues need to be resolved by a researcher before and after performing cluster analysis. We now discuss such issues. One such issue may be that a researcher needs a non-random sub-sample from a bigger random sample for a particular structural model in mind. This raises the question whether to take non-random sub-sample first and then to do clustering or to do clustering first and to dc take non-random sub-sample. Such issue is definitely linked with the problem of selectivity of bias and test for it. Another issue is the model selection strategy cluster wise.

However, we are now faced with a question like whether we can apply the clustering methods to a non-randomly chosen sub-set of the total observations if necessary for some purpose. This gives rise to the issue of selectivity bias. So, we need to resolve this issue first before we proceed further. Thus, we pass on to the next phase of our research in the next section, which involves the investigation of selectivity bias.

Cluster analysis has been employed as an effective tool in scientific inquiry. One of its most useful roles is to generate hypothesis about category structure. An algorithm can assemble observations into groups which prior misconceptions and ignorance would otherwise preclude. An algorithm can also apply a principle of grouping more consistently in a large problem than can a human. Such methods can be particularly valuable in cases wherein the ultimate outcome is something like "Eureka! we never would have thought to associate X with A, but once you do the solution is obvious." A series of interconnected hypothesis may suggest models for the mechanisms generating the observed data. Put another way, cluster analysis may be used to reveal structure and relations in the data. It is a tool of discovery.

A most novel and unique application of cluster analysis is described by Bartels et al (1970). These authors take quite literally the role of cluster analysis as being one of discovering "natural classes."

The uses of cluster analysis described so far in this section are rather general in nature and may be applicable in almost any context. Adding another dimension to the discussion, it is useful to undertake a broad review of the various study disciplines where clustering has been used and to consider some of the entities subjected to cluster analysis. The many fields of study are grouped below into six major areas to help the reader organize this great variety of applications.

i. In the life sciences (biology, botany, zoology, entomology, cytology, microbiology, ecology, paleontology) the objects of analysis are life forms such as plants, animals, insects, cells, microorganisms, communities of interdependent members and the fossil records of life in

other times. The operational purpose of the analysis may range from developing complete taxonomies to delimiting the subspecies of a distinct but varied species.

ii. Closely related to the life sciences are the medical sciences (psychiatry, pathology and other specialties focusing on clinical diagnosis) where the objects of a cluster analysis may be diseases, patients (or their disease profiles), symptoms and laboratory tests. The operational emphasis here is on discovering more effective and economical means for making positive diagnoses in the treatment of patients.

iii. The behavioral and social sciences (psychology, sociology, criminology, education, anthropology, and archeology) have provided the setting for an extraordinary variety of cluster analysis applications. The following entities have been among the many objects of analysis: training methods, behavior patterns, factors of human performance, organizations, human judgements, test items, drug users, families, neighborhoods, clubs and other social organizations, criminals and crimes, students, courses in school, teaching techniques, cultures, languages, artifacts of ancient peoples, and excavation sites. Factor analysis traditionally is a strong competitor to cluster analysis in these applications.

iv. The earth sciences (geology, soil science, geography, regional science, remote sensing) have included applications of cluster analysis to land and rock formations, soils, river systems, cities, counties, regions of the world and land use patterns.

v. Some engineering sciences (pattern recognition, artificial intelligence, systems science, cybernetics, and electrical engineering) have provided opportunities for using cluster analysis. Typical examples of the entities to which clustering has been applied include handwritten characters, samples of speech, finger prints, pictures and scenes, electrocardiograms, wave forms, radar, signals, and circuit designs.

Application in engineering have been relatively few in number to date; any marked increase in the use of cluster analysis in this realm probably will require some dramatically new problem formations.

vi. The information, policy and decision sciences (information retrieval, political science, economics, marketing research, operations research) have included application of cluster analysis to documents and to the terms describing them, political units such as counties and states, legislators, votes on political issues, such issues themselves, industries, consumers, eras of good (and bad) times, products, markets, sales programs, research and development projects, political districting, investments, personnel assignments, credit risks, plant location and floor plan designs. This realm of study probably is the source of the most innovative applications of cluster analysis in the last few years.

The literature of cluster analysis is both voluminous and diverse. Significant work may be found in almost any field of endeavor. The purpose of this thesis is to illustrate the diversity of sources and give at least some general guidance for further reading.

First of all, the terminology differs from one field to another. "Numerical taxonomy" is a frequent substitute for cluster analysis among biologists, botanists and ecologists, while some social scientists may prefer "typology". In the fields of pattern recognition, cybernetics and electrical engineering the terms "learning without teacher" and "unsupervised learning" usually pertain to cluster analysis. In information retrieval and linguistics "clumping" refers to a particular kind of clustering. In geography and regional sciences one may find the term "regionalization". Graph theorists and circuit designers prefer "partition" as a term describing a collection of clusters. Anthropologists have a problem they call "seriation" which consists of grouping ancient excavations into time strata. These terms will help identify a paper as being related to cluster analysis.

The diversity of terminology does not stop here. It extends throughout problem formulations, algorithm descriptions and reporting of results. The cause is probably a mixture of professional jealousy, relative isolation among the fields, and genuine differences of viewpoint. For the novice, the disarray is bewildering and confusing; ultimately it is highly duplicative since the same idea is discovered repeatedly and published in a variety of journals.

Turning to books for a comprehensive survey is not very rewarding. Sokal and Sneath (1963) ranks as the classic work, at least in part because there is no other book that tries to be a complete reference work. But aside from its age, the book is diminished in value for the general user by the fact that a major portion of the text is devoted to philosophical issues unique to biological and evolutionary problems. Jardine and Sibson (1971) also direct their attention to the problems of the biological taxonomist. Their presentation is neither introductory nor elementary; the reader needs a working knowledge of the Sokal and Sneath book as well as considerable mathematical sophistication. Fisher (1968) describes a new method he devised and discusses its application to the agregation problem of economics. Cole (1969) gives a collection of papers well worth reading. Tryon and Bailey (1970) describe their B C TRY system and illustrate its application with three well-studied data sets. Mathematically-oriented psychologists will be most comfortable with this book not only because of the psychological character of the applications but also because the methodologies are extensions of classical factor analysis. These five volumes exhaust the published books in the field. Only Sokal and Sneath try to provide anything approaching a general reference work. Any real education in cluster analysis must then involve extensive reading in learned journals.

HIERARCHICAL CLUSTERING METHODS

Hierarchical clustering methods, which build a tree from branches to root often are called agglomerative methods. Less common are the divisive hierarchical methods which begin at the root and work toward the branches.

Once a tree is constructed for n entities, the analyst may choose from as many as n sets of clusters. Taking the example of Figure 2.3, the possible sets of clusters are as shown in Table 2.1. Notice that these clusters are nested: from the agglomerative view, when two entities are merged they are joined together permanently and become a building block for later merges; from the divisive view, when a group of entities is split into two parts, the parts are separated permanently and may be treated independently for the remainder of the analysis. Herein lie both the strength and weakness of hierarchical methods: by taking early decisions as permanent the number of possibilities that need be examined is reduced greatly as compared with complete enumeration: but this same convention precludes discovering early mistakes or capitalizing on later opportunities

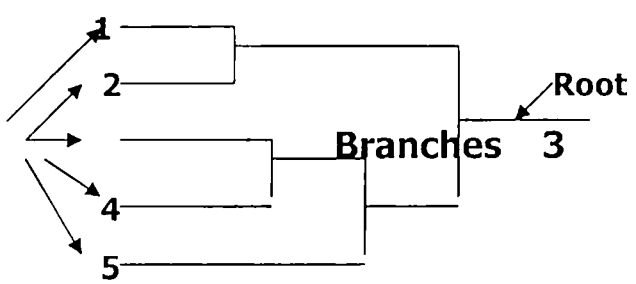


Figure 2.3 Tree for hierarchical clustering.

TABLE 2.1

CLUSTERS

Number of clusters	Clusters
5	(1), (2), (3), (4), (5)
4	(1, 2), (3), (4), (5)
3	(1, 2), (3, 4), (5)
2	(1, 2), (3, 4, 5)
1	(1, 2, 3, 4, 5)

While there may appear to be an overwhelming abundance of hierarchical clustering methods treated in the literature, all the methods seem to be alternative formulations or minor variations of three major clustering concepts:

1. Linkage methods.
2. Centroid methods and
3. Error sum of squares or variance methods.

All of these methods are suitable for clustering data units. However, the linkage methods are the only techniques in this thesis, which may be used for clustering variables. These various methods have individual computational characteristics so that several different approaches to the problem of computation are of interest. A presentation of a general procedure for agglomerative hierarchical clustering and then explores the details of three computational approaches for implementation. Particular variations of the three major hierarchical concepts are described in conjunction with the applicable computational approaches. All these methods are implemented through computer programs.

The Central Agglomerative Procedure

The agglomerative hierarchical methods are treated as variations on a single major technique. Most agglomerative methods can be implemented efficiently within this framework.

$$
\begin{array}{lllllll}
s_{21} & & & & & & \\
s_{31} & s_{32} & & & & & \\
s_{41} & s_{42} & s_{43} & & & & \\
\cdot & \cdot & \cdot & \cdot & & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \\
s_{n1} & s_{n2} & s_{n3} & \cdot & \cdot & \cdot & s_{n(n-1)}
\end{array}
$$

Figure 2.4 Lower triangular similarity matrix.

Let s_{ij} be the similarity between entities i and j as defined by one of the association measures. Assuming that the similarity is symmetric (i.e., $s_{ij}=s_{ji}$), the complete schedule of similarities for all $\binom{n}{2}=\frac{1}{2}n(n-1)$ possible pairwise combinations of entities may be arrayed in a lower triangular similarity matrix, as an Figure 2.4. All the methods assume that the s_{ij} entries are nonnegative. This limitation is of consequence only for correlation and the cosine of the angle between vectors; the distinction between positive and negative association cannot be utilized in these clustering methods. A simple remedy is to use the absolute value or the square of the measure if it can assume negative values. The primary purpose for studying variables and association measures were to provide the means for creating such a similarity matrix. Once the matrix is defined, the process of clustering entities is almost trivially simple. The general procedure for agglomerative clustering on a data matrix is as follows:

1. Begin with n clusters each consisting of exactly one entity. Let the clusters be labeled with the numbers I through n.
2. Search the similarity matrix for the most similar pair of clusters. Let the chosen clusters be labeled p and q and let their associated similarity be s_{pq}, p>q.
3. Reduce the number of clusters by 1 through merger of clusters p and q. Label the product of the merger q and update the similarity matrix entries in order to reflect the revised similarities between cluster q and all other existing clusters. Delete the row and column of S pertaining to cluster p.
4. Perform steps 2 and 3 a total of n—1 times (at which point all entities will be in one cluster). At each stage record the identity of the clusters which are merged and the value of similarity between them in order to have a complete record of the results.

Different agglomerative methods are implemented by varying the procedures used for defining the most similar pair and for updating the revised similarity matrix:

Hierachical clustering on similarity matrix may be carried out using at least three different computational approaches; each approach has its own unique advantages and limitations.

SINGLE LINKAGE

The method of single-linkage cluster analysis is the simplest of all hierarchical techniques. At each stage, after clusters p and q have been merged, the similarity between the new cluster (labeled t) and some other cluster r is determined as follows:

1. If s_{ij} is a distance-like measure

$$s_{tr}=\min(s_{pr},s_{qr}).$$

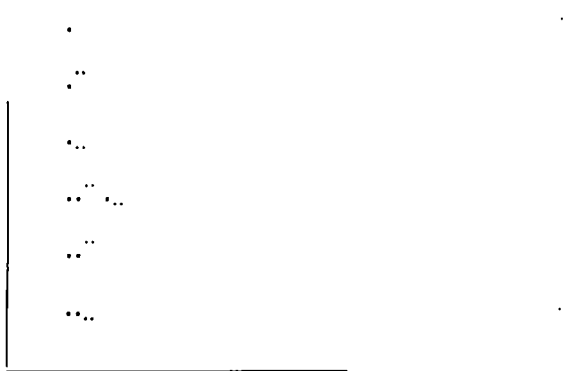
The quantity s_{tr} is the distance between the two closest members of clusters t and r. If clusters t and r were to be merged, then for any entity in the resulting cluster the distance to its nearest neighbor would be at most s_{tr}

2. If s_{ij} is correlation-like measure

$$s_{tr}=\max(s_{pr},s_{qr}).$$

The quantity s_{tr} is the similarity between the two most similar entities in clusters t and r. If clusters t and r were to be merged, then for any entity in the resulting cluster there would be at least one other entity in the same cluster such that the pair would have a similarity at least as large as s_{tr}.

The method is known as single linkage because clusters are joined at each stage by the single shortest or strongest link between them. For example, Fig. 2.3a shows two clusters with their mutually closest members linked. Moving these clusters slightly closer together would make this pair the most similar in the entire array. This example illustrates the fact that single linkage clustering is incapable of delineating poorly separated clusters. On the other hand, if the two clusters are moved farther apart then single-linkage clustering will distinguish between them quite well. For any cluster of two or more entities produced by the single-linkage method, every member is more similar to some other member of the same cluster than to any other entity not in the cluster. The single-linkage method is one of the very few clustering techniques which can outline non ellipsoidal clusters (Figure 2.5 b shows a U-shaped cluster as delineated by the single-linkage method).



The tendency to give long serpentine clusters is sometimes called "chaining"; this property is often criticized because entities at opposite ends of a cluster may be markedly dissimilar.

Since the updating process involves choosing only he minimum or maximum, single-linkage clustering is invariant to any transformation which leaves the ordering of the similarities unchanged, that is, any monotonic transformation. Jonson (1967) and Jardine *et al.* (1967) make a variety of theoretical ovservations on this property. From a practical point of view, this property means that some measures of association are equivalent to each other when used with this method.

This method of clustering appears in the literature in many different forms. Zahn (1971) gives a most enlightening discussion of the intuitive aspects of single-linkage clustering and provides a profusion of two-dimensional examples. Johnson (1967) describes a single-linkage clustering algorithm which inspired the design of the stored matrix hierarchical algorithm given here. The method also has interesting connections with the problem of finding the minimum spanning tree of a graph.

COMPLETE LINKAGE

The complete-linkage method is related closely to the single-linkage method and is no more difficult to execute. At each stage, after clusters p and q have been merged, the similarity between the new cluster (labeled t) and some other cluster r is determined as follows:

1. If s_{ij} is a distance-like measure

$$s_{tr}=\max(s_{pr}, s_{qr})$$

The quantity s_{tr} is the distance between the most distant-members of clusters t and r. If clusters t and r were merged, then every entity in the resulting cluster would be no farther than s_{tr} from every other entry in the cluster. The value of s_{tr} is the diameter of the smallest sphere which can enclose the cluster resulting form the merger of clusters t and r.

2. If s_{ij} is a correlation-like measure

$$s_{tr} = \min(s_{pr}, s_{qr})$$

The quantity s_{tr} is the similarity between the two most dissimilar entities in clusters t and r. If clusters t and r were to be merged, then every entity in the resulting cluster would have a similarity of at least s_{tr} with every other entity in the cluster.

The method is called complete linkage because all entities in a cluster are linked to each other at some maximum distance or minimum similarity. Such a cluster is called a "maximally connected sub graph" in graph theory. In contrast to the single-linkage method, the interpretation of the clusters can be made only in terms of the relationships within individual clusters; there is no particularly useful interpretation involving the differences between clusters. Like the single-linkage method, complete-linkage cluster analysis is invariant to monotonic transformations of the similarity measure. Johnson (1967) discusses this property and relates it to the ultrametric inequality.

AVERAGE LINKAGE WITHIN THE NEW GROUP

In the single linkage method each cluster is characterized by the lowest link needed to connect any member of the cluster to some other member of the cluster; in the complete linkage method each cluster is characterized by the longest link needed to connect every member of a cluster to every other member. Instead of relying on extreme values as in these two cases, it may be of interest to characterize a cluster by the average of all links within it. The s_{ij} entries in the initial similarity matrix may be built up as the sum of similarities associated with all pair wise combinations formed by taking one entity from cluster i and the other from cluster j. Before any merges have occurred each cluster consists of just one such entity and there is only one such pair of entities for each pair of clusters. When clusters p and q are merged, the sum of pairwise similarities between the new cluster (labeled t) and some other cluster r is updated as $s_{tr}=s_{pr}+s_{qr}$ to maintain this form for the similarity matrix. Also,

let SUM_i be the sum of all pair wise similarities among entities within cluster i and let N_i be the number of entities in cluster i. Before the first merge $SUM_i=0$ and $N_i=1$ for all clusters. When clusters p and q are merged, the new cluster is labeled t and the description for the cluster is updated as

$$SUM_i = SUM_p + SUM_q + s_{pq}$$
$$N_t = N_p + N_q.$$

Then, when searching for the most similar pair, the average within group similarity for the cluster formed by merging the candidate pair of clusters i and j is, $\frac{SUM_i + SUM_j + S_{ij}}{(N_i + N_j)(N_i + N_j - 1)/2.}$

The denominator in the last expression is the number of distinct pair wise combinations that may be formed from the entities in the merger of i and j.

This method of clustering is not dependent on extreme values for defining clusters and accordingly it is not possible to make any statements about the minimum and maximum similarity within a cluster. However, as a practical matter this method frequently gives results that are little different from those obtained with the complete linkage method.

AVERAGE LINKAGE BETWEEN MERGED GROUPS

An alternative method based on average linkage consists of evaluating the potential merger of clusters i and j in terms of the average similarity for links between the two clusters. If the similarity matrix is treated, then s_{ij} is the sum of pairwise similarities between clusters i and j. The number of such between group pairwise similarities is the product of N_i and N_j, where N_i is the number of entities in cluster i. Then the average between group similarity for clusters i and j is $\frac{S_{ij}}{N_i N_j}$.

The principal difference between this method and the sums of within group pair wise similarities, SUM_i and SUM_j, are ignored. Lance and Williams (1967a) label

this method as "group average" and provide a different computing expression. As might be expected, this method gives results which are not radically different from those obtained with the method.

CENTROID METHODS

In statistical analysis the mean is often the basic summary statistic for a set of data. The familiar t-test and the analysis of variance technique both are used to identify differences between groups by testing for differences between their means. It then should be appealing to cluster hierarchically by merging at each stage those two clusters with the most similar mean vectors or centroids. The method is easy to implement with a stored similarity matrix only for squared Euclidean distance. Lance and Williams (1967a) give the update equation fo this case as, $S_{tr}= \frac{N_p}{N_p+N_q}S_{pr}+\frac{N_q}{N_p+N_q}S_{qr}-\frac{N_pN_q}{(N_p+N_q)}S_{pq}$, Where p and q are the labels for the clusters just merged, t is the label adopted for the new cluster, and r is any other existing cluster. This equation could be used with any similarity measure for either variables or data units; however, the results would lack a useful interpretation and could be quite strange unless S_{ij} is the squared Euclidean distance between the centroids of cluster i and j which treats the centroid method but in a context amenable to a wider variety of distance functions. This method of clustering has enjoyed some considerable popularity in the community of biologists. Sokal and Sneath (1963, pp. 183-185) describe the so-called "pair group" methods which are variations on the centroid approach. The essential difference is that the centroids of the merged group are weighted to compute the centroid of the new group and these weights are not necessarily proportional to the number of entities in each group. Another variation on the same approach is the "median method" first proposed by Gower (1967, p.626). The general idea is that the centroids are weighted equally regardless of how many entities are in the respective clusters. When S_{ij}

is a distance-like function, the update equation for the median method is

$$S_{tr}=\frac{1}{2}(S_{jr}+S_{qr})-\frac{1}{4}S_{pq}$$

When S_{ij} is a correlation-like function, the update equation is,

$$S_{tr}=\frac{1}{2}(S_{pr}+S_{qr})+\frac{1}{4}(1-S_{pq}).$$

Lance and Williams (1967a, p.375) note that the line between the centroid of some cluster r and the centroid of the cluster formed by the merger of clusters p and q lies along the median of the triangle formed by p, q, and r, hence the name for the method.

A characteristic of the centroid method and its variants is that the similarity value associated with the merger of the most similar clusters may rise and fall from stage to stage. For example, when the similarity measure is a distance, the distance between the centroids of some pair may be less than that between another pair merged at an earlier stage; the following sequence of distances between cluster centroids could occur in three successive mergers: $s_{21}=8$, $s_{34}=9$, $s_{31}=7$. The resulting tree is shown in Figure 2.6a; the last merger occurs at a lower level than either of the preceding two so that depicting the tree requires crossovers. With only a few reversals the tree begins to look like a wiring diagram for a color television set. Reversals occur because cluster centroids can migrate as mergers take place. Referring to Fig 2.4b, if clusters with centroids p and q are merged, the centroid of the resulting cluster t will be closer to point r than were either p or q. This problem is not shared by any of the other hierarchical methods in this chapter.

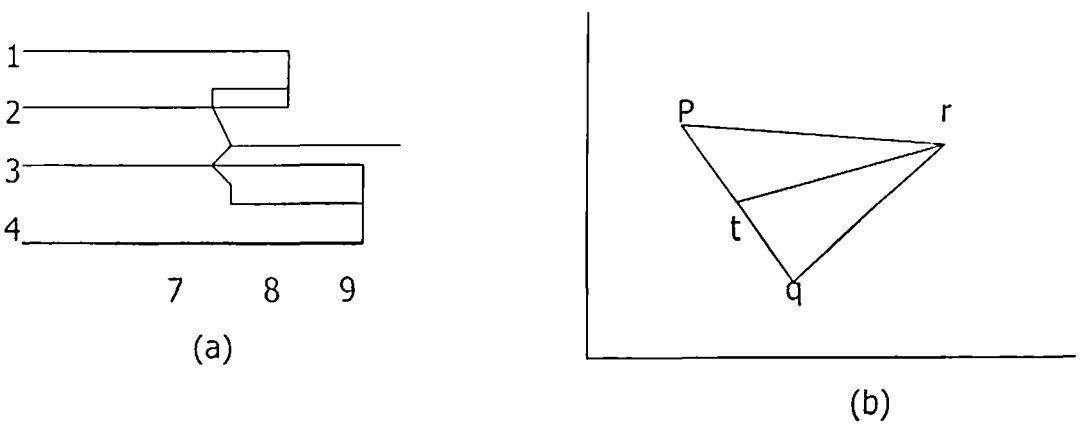


Figure 2.6 Centroid reversal phenomenon.

THE WARD METHOD

Ward (1963) and Ward and Hook (1963) describe a very general hierarchical clustering method in which the merges at each stage are chosen so as to maximize an objective function which reflects the investigator's purpose in the particular problem. Ward illustrated his method with an error sum of squares objective function, and the example came to be better known than the more general procedure. This particular method originally was implemented in the form described in the form described in hierarchical clustering methods. However, Wishart (1969a) showed how the Ward algorithm could be implemented through updating a stored matrix of squared Euclidean distances between cluster centroids.

Define the following quantities:

x_{ijk} = score on the i^{th} of n variables for the j^{th} of m_k data units in the k^{th} of h clusters,

$$\bar{x}_{ik} = \sum_{j=1}^{j=mk} x_{ijk} / m_k ,$$

= mean on the ith variable for data units in the kth cluster,

$$E_k = \sum_{i=1}^{i=n} \sum_{j=1}^{j=mk} (x_{ijk} - \bar{x}_{ik})^2 = \sum_{i=1}^{i=n} \sum_{j=1}^{j=mk} x_{ijk}^2 - m_k \sum_{i=1}^{i=n} \bar{x}_{ik}^2$$

= error sum of squares for cluster k; sum of Euclidean distances from each data point in cluster k to the mean vector of cluster k; within group squared deviations about the mean for cluster k;

$$E = \sum_{k=1}^{k=h} E_k$$

= total within group error sum of squares for the collection of clusters.

To begin there are m data units, each a cluster unto itself; consequently the membership and the mean of each cluster coincide so that $E_k=0$ for all clusters. The Ward objective is to find at each stage those two clusters whose merger

gives the minimum increase in the total within group error sum of squares E. Suppose clusters p and q are chosen to be merged and the resulting cluster is denoted as t. Then the increase in E is,

$\Delta E_{pq}=E_t - E_p - E_q$

$$= \left[\sum_{i=1}^{i=n}\sum_{j=1}^{j=m_t} x_{ijt}^2 - m_t\sum \bar{x}_{it}^2\right]-\left[\sum_{i=1}^{i=n}\sum_{j=1}^{j=mp} x_{ijp}^2 - m_p\sum_{i=1}^{i=n}\bar{x}_{ip}^2\right]-\left[\sum_{i=1}^{i=n}\sum_{j=1}^{j=mq} x_{ijq}^2 - m_q\sum_{i=1}^{i=n}\bar{x}_{iq}^2\right]$$

$$= m_p\sum_{i=1}^{i=n}\bar{x}_{ip}^2 + m_q\sum_{i=1}^{i=n}\bar{x}_{iq}^2 - m_t\sum_{i=1}^{i=n}\bar{x}_{it}^2 \quad \text{..... (2.20)}$$

since the terms involving x^2_{ijk} all cancel. The mean on the ith variable for the new cluster is found from the relation $m_t\bar{x}_{it} = m_p\bar{x}_{ip} + m_q\bar{x}_{iq}$; squaring both sides of this equation gives $m_t^2\bar{x}_{it}^2 = m_p^2\bar{x}_{ip}^2 + m_q^2\bar{x}_{iq}^2 + 2m_pm_q\bar{x}_{ip}\bar{x}_{iq}$.

The product of the means can be re-written as

$$2\bar{x}_{ip}\bar{x}_{iq} = \bar{x}_{ip}^2 + \bar{x}_{iq}^2 - (\bar{x}_{ip} - \bar{x}_{iq})^2,$$

so that

$$m_t^2\bar{x}_{it}^2 = m_p(m_p + m_q)\bar{x}_{ip}^2 + m_q(m_p + m_q)\bar{x}_{iq}^2 - m_pm_q(\bar{x}_{ip} - \bar{x}_{iq})^2.$$

Noting that $m_t=m_p+m_q$ and dividing both sides of the last equation by m_t^2,

$$\bar{x}_{it}^2 = \frac{m_p}{m_t}\bar{x}_{ip}^2 + \frac{m_q}{m_t}\bar{x}_{iq}^2 - \frac{m_pm_q}{m_t^2}(\bar{x}_{ip} - \bar{x}_{iq})^2. \quad \text{..... (2.21)}$$

Substituting Eq. (2.2) into Eq. (2.1) then gives

$$\Delta E_{pq} = \frac{m_pm_q}{m_p + m_q}\sum_{i=1}^{i=n}(\bar{x}_{ip} - \bar{x}_{iq})^2. \quad \text{..... (2.22)}$$

Thus, the minimum increase in the error sum of squares is proportional to the squared Euclidean distance between the centroids of the merged clusters. This result is different from the centroid clustering methods in that it involves weighting of the distance between centroids when computing the distances. The error sum of squares function E is non decreasing and the method is not subject to reversals.

Equation (22) is sufficient to define the algorithm when using the stored data approach as in hierarchical clustering methods. However, for the stored similarity matrix approach an update equation still needs to be developed. Letting t denote the result of merging clusters p and q, and letting r denote any other cluster, the increase in E that would result from the potential merger of clusters r and t is obtained from Eq. (2.3) as,

$$\Delta E_{rt} = \frac{m_r m_t}{m_r + m_t} \sum_{i=1}^{i=n} (\bar{x}_{ir} - \bar{x}_{it})^2 \quad \text{..... (2.23)}$$

By substituting $\bar{x}_{it} = (m_p \bar{x}_{ip} + m_q \bar{x}_{iq})/m_t$, $m_t = m_p + m_q$, and the squared terms in the summation may be written

$$(\bar{x}_{ir} - \bar{x}_{it})^2 = \frac{m_p}{m_t}(\bar{x}_{ir} - \bar{x}_{ip})^2 + \frac{m_q}{m_t}(\bar{x}_{ir} - \bar{x}_{iq})^2 - \frac{m_p m_q}{m_t^2}(\bar{x}_{ip} - \bar{x}_{iq})^2.$$

Substituting in Eq. (2.4) and gathering terms

$$\Delta E_{rt} = \frac{1}{m_r + m_t} \sum_{i=1}^{i=n} [m_r m_p (\bar{x}_{ir} - \bar{x}_{ip})^2 + m_r m_q (\bar{x}_{ir} - \bar{x}_{iq})^2 - \frac{m_r m_p m_q}{m_p + m_q}(\bar{x}_{ip} - \bar{x}_{iq})^2],$$

and using Eq. (2.3) again

$$\Delta E_{rt} = \frac{1}{m_r + m_t}[(m_r + m_p)\Delta E_{rp} + (m_r + m_q)\Delta E_{rq} - m_r \Delta E_{pq}]. \quad \text{..... (2.24)}$$

It would be simplest to construct the similarity matrix such that $s_{ij}=E_{ij}$ and then use Eq. (2.5) as an update equation. Since the initial clusters each consist of one data unit, Eq. (2.3) shows that the initial entries in such a similarity matrix are $E_{ij}=\frac{1}{2}d_{ij}^2$, half the squared Euclidean distance between data units i and j. However, it is not necessary to actually divide the entries in the initial matrix of squared Euclidean distances by 2; the updates of Eq. (2.5) may be carried out using the squared distances and when the increment to the error sum of squares function is added, merely use half the computed value. The entire algorithm then involves the following steps:

1. Store the lower triangular similarity matrix of squared distances. Treat s_{ij} as E_{ij}
2. **At each stage merge the two clusters which as a pair give the smallest increment to the error sum of squares. Let these clusters be labeled p and q with the resulting cluster labeled t.**
3. Update entries in the similarity matrix according to Equation (2.24). Increment the error sum of squares function by using the replacement relation

$$E=E+\frac{1}{2}\Delta E_{pq}$$

4. Repeat steps 2 and 3 for m-1 stages where m is the number of data units.

As with the centroid methods, the algorithm operates directly on the similarity matrix which is just an array of numbers; hence, the entries in this matrix could be computed using any association measure for either variables or data units. However, it is not known what properties the resulting clusters would have unless the similarity is the squared Euclidean distance.

Hierarchical methods do not gurantee an optimal solution in terms of the clustering criterion; a set of h clusters produced by the Ward method may or may not give the minimum possible sets of h clusters formed from the m data units. However, the Ward solution is usually very good even if it is not optimal on this criterion. The Equation (2.22) gives the increase in the error sum of squares due to the merger of clusters p and q; this latter equation requires only the storage of the mean vector and the number of data units in each cluster. Since the clustering begins with each data unit as a cluster of one member, the storage requirement is equal to the size of the data set plus a vector containing the number of data units in each cluster. However, the constant recompulation of cluster means gives the opportunity for accumulation of round-off error

which could be important in large problems. Accordingly a slightly altered method of computation is suggested. Define he following quantities:

$$T_{ik} = \sum_{j=1}^{j=m_k} x_{ijk} = m_k . \bar{x}_{ik}$$

= total of scores on ith variable for data units in the kth cluster,

$$S_k = \sum_{i=1}^{i=n} \sum_{j=1}^{j=m_k} x_{ijk}^2$$

= sum of squared scores on all variables for all data units in the kth cluster.

Then, the error sum of squares for cluster k may be written as

$$E_k = S_k - \sum_{i=1}^{i=n} T_{ik}^2 / m_k$$

The increase in the total error sum of squares due to the merger of clusters p and q to form the new cluster t is,

$$\Delta E_{pq} = E_t - E_p - E_q$$

$$= S_t - \sum_{i=1}^{i=n} T_{it}^2 / m_t - E_r - E_q$$

$$= S_p + S_q - \sum_{i=1}^{i=n} (T_{ip} + T_{iq})^2 / (m_p + m_q) - E_p - E_q,$$

since $m_t = m_p + m_q$, $S_t = S_p + S_q$, and $T_{it} = T_{ip} + T_{iq}$. This expression could easily be reduced further or put in many other forms; however, accumulating S_k, E_k, m_k, and $\{T_{ik}: i=1,\ldots\ldots, n\}$ for each cluster primarily involves simple addition and avoids the repeated multiplication and division required when using cluster means. This formulation also permits an easy implementation of the other methods given in the following sections. In particular, the other methods can be utilized to control the clustering out at the same time the Ward criterion can be computed easily to give a common base of comparison for results from the different methods.

Assessing Cluster Analysis Results

In general, procedure used to evaluate groups determined by cluster analysis are of two types. Procedures of one type test are resultant groups against the null hypothesis that the group were determined randomly. Procedures of the second type are based on the accuracy that the cluster analysis method in use has found the optimal partition and given an appropriate null hypothesis, compare the resultant clusters against the distribution of optimal partitions.

To illustrate the differences between the two types of procedures assume that there is some F-statistic which can be used to assess resultant clusters. If the null hypothesis states that the points representing the units are distributed randomly throughout a multivariate space procedures of the first type would compare a given partition's observed F-statistic with the F-distribution generated from all possible partitions. Procedures of the second type, however, would compare the observed F-value with the theoretical distribution of maximum F-values.

Some researchers e.g. Arnold 1979, Milligan and Mahajan (1980) argue that only tests of the second type should be used because most, if not all, cluster analysis algorithms find relatively homogeneous groups. This argument, however, has two limitations. First, it is not clear that all algorithms usually find "good" partitions. Instead, it appears that many algorithms perform well for certain types of data and may perform very poorly for other types of data (Cormack, 1971). Second tests similar to those proposed by Arnold (1979) and Milhgan and Mahajan (1980) are usually based on the results of a specific cluster analysis algorithm (Arnold 1979; Hubert 1974), require the enumeration of all possible partitions, or impose some other significant costs or assumptions (Friedman and Rubin 1967; Lee 1979; Ling 1973; Marriott 1971).

Tests of the first type are relatively easy to apply (especially those described here) and provide a useful first step for analyzing cluster analysis results. Though the results from most cluster analysis algorithms may "pass" such tests, those results which do not would certainly be suspect. In some cases (e.g.,

when a researcher uses cluster analysis as a data reduction technique and it is not concerned with the presence or absence of natural grouping), these tests may suffice by themselves (Anderberg,1973).

A general approach is suggested for testing whether resultant clusters differ significantly from randomly determined groups of the same size. Although the approach is similar to that suggested by McClain and Rao (1975), a differs in two important aspects. First, the approach is sufficiently general to allow the definition of numerous alternative statistics which might be of interest to a researcher. Second, for any of these alternative statistics, expressions are given for deriving the parameters of the required permutation distribution. These expressions, based on an approach suggested by Mantel (1967), result in an easily implemented method which can be used to test whether a resultant partition significantly differs from a randomly determined partition.

CLUSTER ASSESSMENT STATISTICS

To define the assessment statistics, let us assume that there is a set of N units $\{u\}_{i=1}^{N}$ where each i^{th} unit is described by a vector of C descriptive characteristics. Furthermore, we assume that these characteristic vectors have been reduced to a square symmetric matrix of non-negative measures (d_n) which express a degree of similarity between all pairs of units μ_i and μ_j. In this study, we assume that these similarity measures are related to some definition of distance between the points which represent units.

Given that a cluster analysis algorithm, using the similarity measures (d_{ij}), has identified P mutually exclusive and exhaustive subsets of the N units, our goal is to analytically assess these partitions. To define assessment statistics, we consider the goal of many cluster analysis procedures (Cormack 1971): find groups which are both internally cohesive (homogeneous) and externally isolated (heterogeneous). To measure internal homogeneously, we use some

measure of average distance within groups: to measure external heterogeneity, we use some measure of average distance between groups.

Because there is no unique way to measure average within or between group distance, three different methods are used here to define the following statistics. In each case, a different set of weights is used to define the average within-or between – group distance.

We will use w_r to denote the r^{th} measure of within group homogeneity and b_r to denote the r^{th} measure of between-group heterogeneity. Because we usually want to minimize the within-group homogeneity and maximize the between-group heterogeneity. We will denote the r^{th} overall assessment statistic by g, and define it as $g_r=w_r - b_r$

such that, given a number of groups, a smaller value of g_r denotes a more significant partition. Though this general function is used to define the statistics in this study, note that any linear combination of within-group homogeneity and between-group heterogeneity could be used. Other measures (e.g. the ratio of w_r/b_r) can be treated by other approaches (Overall and Klett 1972).

To describe the statistics precisely, the following notation is used. The indices of the units in each group are indicated by the index sets $\{G_rf_{i=1}\}$ for example, if units u_2, u_3 and u_5 are grouped in the second group, $G_2=\{2,3,5\}$. In addition, n_k equals the number of units in the k^{th} group (e.g. $n_2=3$).

The g_1 Measure

The first measure of within group homogeneity (w_1) can be defined by initially computing the average pair wise distance between units in each group and then finding the unweighted mean of these values. In this case, w_1 is defined as,

$$w_1 = \frac{1}{p}\left\{\sum_{i=1}^{p} \frac{2}{n_i(n_i-1)}\left[\sum_{i,j=Gir,\ i<j} d_{ij}\right]\right\}$$

In similar fashion, an unweighted average between group distance (b_1) can be defined. In this case, the average distance between units in all pairs of groups is found; the mean of these (p/2) average values is then used to define b_1. More precisely, the value of b_1 is defined as

$$b_1 = \frac{2}{p(p-1)}\left\{\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\frac{1}{n_i n_j}\left[\sum_{i=1}\sum_{j=1} d_{ij}\right]\right\}$$

Assuming that both within-group homogeneity and between-group heterogeneity are important. We define the assessment statistic g_i as

$g_1 = w_1 - b_1$

The g_2 measure

Other measures can be defined similarly by considering alternative definitions of within-group homogeneity and between-group heterogeneity. For example, if each average within-group distance is weighted by $(n_1-1)/2$, the overall mean within-group distance (w_2) becomes

$$w_2 = \sum_{i=1}^{p}\frac{1}{n_i}\left[\sum_{i,j=G_i,i<j} d_{ij}\right]$$

If squared Euclidean distances are used to define the similarity measures $\{d_{ij}\}$. w_2 is equal to the within-group sum of squared distances (or the pooled within-group sum of squares). This definition was used by Fisher (1958), Vinod (1969), Rao (1971), Jensen (1971), and Bellman (1973), among others, who developed cluster analysis algorithms to find the partition which minimizes w_2.

An alternative definition of w_2 is provided by considering the (N×C) data matrix X. Letting the origin be the center of gravity, the total scatter matrix T is defined as:

$T = X'X$

Which can be sub divided into the within-group scatter matrix (W) and the between-group scatter matrix (B). Where T=W–B.

and

$$W=\sum_{i=1}^{n}\left[\sum_{i\in G_i}(R_i-C_i)'(R_i-C_i)\right]$$

and

$$B=\sum_{i=1}^{i=n} n_i C_i' C_i.$$

and R_{ik} =i^{th} row vector in matrix X and C_i=mean vector for group G_i.

Friedman and Rubin (1967) point out that a useful scalar measure in this case is the trace, where trace (W) equals the total within-group sum of squares defined by $w_2=\sum_{i=1}^{p}\frac{1}{n_i}\left[\sum_{i,j\in G_i, i<j} d_{ij}\right]$ and trace (B) equals the total weighted (by n_i) distance from each unit to the origin. Letting b_2 equal trace (B), the g_2 measure can be defined as

g_2=0=trace(W) – trace (B)

However, because trace(W) + trace (B) = trace(T) and trace(T) is constant for any set of units, the g_2 measure can be reduced to w_2 in this case.

It should be noted that other scalar assessment measures have been suggested based on W, T or B. For example, Overall and Klett (1972) discuss the measure defined by the ratio of the determinants W/T and show how this measure which equals Wilks Λ may be approximated by either an F or χ^2 distribution. An other measure is defined by Hotelling's trace criterion trace, $[W^{-1}B]$. Marriott (1971) suggested the measure p^2W. Other measures have been used by Hartigan (1975) and Arnold (1979), among others.

If squared Euclidean distances are not used. There are several alternatives for defining g_2. For example, w_2 could be redefined as the average within-group distance to the centroid and b_2 could be redefined as the average distance between group centroids. However, because squared Euclidean distance is widely used. g_2 is defined here by the computationally simpler expression in

$$w_2=\sum_{i=1}^{p}\frac{1}{n_i}\left[\sum_{i,j\in G_i, i<j} d_{ij}\right]$$

The g_3 Measure

A third measure can be defined by weighting the average pair wise distances by the number of object pairs $\binom{nk}{2}$ in a group. In this case,

$$w_3 = \sum_{k=1}^{p}\left[\sum_{i,j\in i_k} d_{ij}\right] \Big/ \sum_{k=1}^{p}\binom{n_k}{2}$$

and $$b_3 = \left\{\sum_{k=1}^{p-1}\sum_{j=k+1}^{p}\left[\sum_{i,j\in i_k (i<j)} d_{ij}\right]\right\} \Big/ \left[\binom{N}{2}\sum_{i=1}^{p}\left[\binom{n_k}{2}\right]\right]$$

and $g_3=w_3-b_3$.

Although other measures could be defined in similar fashion (e.g. by varying the weights used to compute w3 and b3) only the measures defined by equations by

$w_2 = \sum_{i=1}^{p}\frac{1}{n_i}\left[\sum_{i,j=(i_i,i<j} d_{ij}\right]$ and $g_3= w_3-b_3$ are used.

DISTRIBUTION OF THE ASSESSMENT STATISTICS

To eliminate bias caused by the number of groups as well as test the null hypothesis that the groups are randomly determined, one can derive to permutation distribution for the measures and expressions for the expected value E(g) and variance var(g) of each r statistic. Basically two approaches have been proposed. One approach suggested by Graves and Whinston (1970) and studied by Huben and Levin (1977) is to implicitly consider all possible permutation of the N units and compute the values of $E(g_r)$ assuming that all permutations are equally likely. The difficulty with this approach is that it ignores the transitivity requirement inherent in the cluster analysis problem

(that is, if units A, B, and C are grouped together and units C and E are grouped together, then units A, B, and E must also be grouped together).

A more reasonable approach is to assume that the number of groups (P) and group size (n_2) are held constant (thereby preserving any underlying population structure) and implicitly enumerate all possible partitions having these values of P and n_k. To avoid enumerating all possible partitions, however, a method described by Knox (1964) and Mantel (1967) who studied disease clustering in time and space, can be used to find directly the parameters of the permutation distribution.

To derive these expressions, we will assume that there is a square symmetric matrix of order N with elements {t}, and define a statistic as

$$Z = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} t_{ij} d_{ij}$$

(using the similarity measures d_{ij}). Given any values of t_{ij} and d_{ij}, however, expressions for the expected value of Z, E(Z), and the variance of Z, Var(Z), are

$$E(Z) = \frac{1}{\binom{N}{2}} \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} t_{ij} \right] \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij} \right]$$

and

$$Var(Z) = \frac{1}{4N(N-1)} \left[2\beta_d \beta_t + \frac{4H_d H_t}{N-2} + \frac{K_d K_t}{(N-2)(N-3)} \right] - \{E(Z)\}^2$$

Where

$$\beta_d = \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij}^2$$

and

$$H_d = \sum_{i=1}^{N} \left(\sum d_{ij} \right)^2 - \beta_d$$

$$K_d = \left(\sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij} \right)^2 + 2\beta_4 - 4 \sum_{i=1}^{N} \left(\sum_{i=1}^{N} d_{ij} \right)^2 .$$

$$G_d = \left(\sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij} \right)^2$$

Now, $\beta_t = \sum_{i=1}^{N} \sum_{j=1}^{N} t_{ij}^2$

and

$$H_t = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} t_{ij} \right)^2 - \beta_t$$

$$K_t = \left(\sum_{i=1}^{N} \sum_{j=1}^{N} t_{ij} \right)^2 + 2\beta_t - 4 \sum_{i=1}^{N} \left(\sum_{i=1}^{N} t_{ij} \right)^2.$$

$$G_t = \left(\sum_{j=1}^{N} \sum_{j=1}^{N} t_{ij} \right)^2$$

Mantel (1967) and Hubert and Levin (1977) examined the distribution of Z for various tij values. In both studies, the authors concluded that a normal distribution with parameters E(Z) and Var(Z) could be used to represent the empirical distribution of Z if N is sufficiently large. However, as no guidelines apparently exist for determining an adequate size of N. Hubert and Levin recommended that the Chebyshev inequality be used as a conservative approach.

The g_1 Measure

Recall that $g_1 = W_1 - b_1$ and that w_1 and b_1 are defined as the mean unweighted within-group pair wise distance and between-group distance, respectively. Then, if

$$t_{ij} = \begin{cases} \dfrac{2}{n_k(n_1 - 1)P} \quad for\ i \varepsilon G_t\ and\ j \varepsilon G_t\ (i \neq j) \\ \dfrac{2}{n_k n_1 P(P-1)} \quad for\ i \varepsilon G_t\ and\ j \varepsilon G_{kt}\ (k \neq 1) \\ 0 \quad for\ i = j \end{cases}$$

and Z is defined by expression $Z = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} t_{ij} d_{ij}$, then Z=$g_1$ and the expected value of g_1[E(g_1)] and variance of g_1[Var(g_1)] are defined by

$$E(Z) = \frac{1}{\binom{N}{2}} \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} t_{ij} \right] \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij} \right]$$ and

$$Var(Z) = \frac{1}{4N(N-1)} \left[2\beta_d \beta_t + \frac{4H_d H_t}{(N-2)} + \frac{K_d K_t}{(N-2)(N-3)} + \frac{K_d K_t}{(N-2)(N-3)} \right] - \{E(Z)\}^2$$

respectively.

If isolates are present (i.e., groups with only a single unit) and P equals the number of such isolates, it can be easily shown that

$$E(g_1) = \frac{2PD}{NP(N-1)} (for\ P > 1)$$

where, $D = \sum_{i=1}^{N-1} \sum_{j=1}^{N} d_{ij}$

and P equals the total number of groups. This is derived from the fact that

$$\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} t_{ij} = \frac{1}{P}$$

$$\sum_{k=1}^{N-1} \sum_{i,j \in i_k} t_{ij} = 1 - \frac{\overline{P}}{P}$$

$$\sum_{i \in i_k} \sum_{j \in i(k \neq 1)} t_{ij} = 1$$

and

If there are no isolates (i.e. $\overline{P} = 0$), it is clear that E(g_1)=0 (for P>1).

Similarly, values of t can be defined so that Z=w1 or Z=b1. In this case, it can be shown that

$$E(w_1) = [(1 - \overline{P}/P)D \Big/ \binom{N}{2}$$

and $E(b_1)= D/\binom{N}{2}$

If there are no isolates (i.e., $\overline{P} = 0$), then $E(w_1)=E(b_1)= D/\binom{N}{2}$

(i.e., the expected within or between-group distance is merely equal to the average pair wise distance). If there is only a single cluster (i.e., P=1 and $\overline{P} = 0$), $E(w_1)= D/\binom{N}{2}$ and $E(b_1) =0$; likewise, if each unit is grouped individually (i.e., $\overline{P} = P = N$), $E(w_1) =0$ and $E(b_1)= D/\binom{N}{2}$.

The g_2 measure

If we assume that squared Euclidean distances are used and $g_2 = w_2$ (as defined by equation 2), g_2 is equal to Z if Z is defined by equation 4, and

$$t_{ij} = \begin{cases} \frac{1}{n_k} \text{ if } i, j \varepsilon G_k (i \neq j), \\ 0 \text{ otherwise} \end{cases}$$

$E(g_2)$ and $Var(g_2)$ are then defined by equations 5 and 6 respectively. In this case,

$$\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} t_{ij} = \frac{(N - P)}{2}.$$

therefore, $E(g_2)=D(N - P)/N(N - 1)$. When all units are individually grouped (i.e., P=N), $E(g_2) = 0$; when all N units are combined into a single group (i.e., P=1), $E(g_2) = D/N$.

The g3 Measure

If $t_{ij} = \begin{cases} \dfrac{1}{\sum_{k=1}^{r}\binom{n_k}{2}} & \text{if } i\varepsilon G_k \text{ and } j\varepsilon G_k \ (i\neq j), \end{cases}$

$t_{ij} = \begin{cases} \dfrac{-1}{\binom{N}{2}-\sum\binom{n_l}{2}} & \text{if } i\varepsilon G_k \text{ and } j\varepsilon G_k \ (k\neq l). \end{cases}$

$= 0$ if i=j

then g_3, $E(g_3)$ and $Var(g_3)$ are defined by

$\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} t_{ij} = 0$ for p>1.

and $E(g_3)=0$ (if P>1); likewise, it is easily shown that the expected average within-group distance $E(w_3)$ and the expected average between-group distance $E(b_3)$ equal $D/\binom{N}{2}$.

When only one group exists,

$E(g_3) = D/\binom{N}{2}$; When N groups exist,

$E(g_3) = -D/\binom{N}{2}$.

VI. 3 CHOICE OF VARIABLES

We have principally included the economic variable for consumption estimation. However, we also have included some variable like the choice for different duration of meat consumption for the person of house holds which have some economic implication .We present the explanatory variables with the rationale used for the cluster analysis of the following table:-

Variable	**Logic**
d_3 = Manual Worker	Household head is one of the groups expenditure is insignificantly.
d_4 = Freezer or microwave ownership	Insignificantly effect in purchase.
d_5 = House ownership	Insignificantly effect in purchase.
LnME = Total meat expenditure.	Total meat expenditure causes a significant.
T = take-away	Purchase and expenditure highly significantly influenced
N_1 = Number of house hold members in the age group 5-15 years	Expenditure share is insignificantly effected.
N_2 = Number of house hold member in the age group15-64 years	Expenditure share is insignificantly effected.
σ = Standard deviation	The effect of variation in probability of buying between household

VII-4. ESTEMATED RESULTS AND ANALYSIS

Table -1

ESTIMATION RESULTS OF THE SYSTEM MODEL IN CLUSTER –1

LOG OF LIKELIHOOD FUNCTION= 6043.04 NUMBR OF OBSRVATION =2766

BEV: (beef and veal)

Meat type	Parameter estimate	Standard Error	t –statistic
Intercept	.1726	.013	2.122
d_3	.0161	.00813	1.996
T	000044	.00018	.235
LnME	08723	.01304	6.314
σ	.0795	.0421	1.887

R^2 = .27375

If the household head belongs to one of the groups of either farmers, fishermen, process workers, laborers, miners, tradesmen, transport works (d_3), the expenditure on beef and veal(BEF) is significantly affected. The in share is 1.6% compared to a household hear not being a manual worker. For a given level of real total meat expenditure(ME), a manual worker favors beef and veal as compared to household heads being in other profession. . The variable expenditure on T(Take-a way) has an insignificant impact on beef an veal (BEV)purchases. Total real meat expenditure (ME)influences expenditure share on BEV positively and highly significantly with an increase of 8.2% for an increase in real total meat expenditure (ME). The above results show that only the occupation of the household hear being a manual worker (d_3) and an increase in real total meat expenditure (ME) can cause arise in the share of beef and veal (BEV) purchases . The impact of variation in the probability of buying this meat category between households is trivial.

Table-2

MUL: (Mutton lamb)

	Parameter estimate	Standard Error	t-statistic
Intercept	.2566	.0705	3.636
d_3	.0116	.0070	1.655
T	-.00063	.00016	-3.921
LnME	.0395	.0113	3.496
σ	.1048	.0398	2.628

R^2 = .372965

Expenditure on mutton and (MUL) is insignificantly if he occupation of the household head is one of the groups defined by d_3 (manual workers). Both T (take-way) and total meat expenditure have the expected sign. Furthermore, T (take-away) supposedly leads to less meat purchase and greater real total meat expenditure may cause a greater share for a particular meat type. Both of these variable influence mutton purchase, highly significantly. For a $1 increase in T (take-away), expenditure share on mutton and (MUL) decreases by .06% while it is 3.9% more for an increase in real total meat expenditure(ME). The increase in availability of take-away and fast foods will cause less MUL purchase which is offset by an increase in the real total meat expenditure(ME) . For a given level of ME, more take-away (T) will cause less purchase of mutton and lamb(MUL). The impact of the selectivity variable is negative and significant. Thus, if the probability of buying mutton and lamb increases, the purchase share of it decreases.

Table-3

PIGGR: (Pork, ham and bacon)

	Parameter estimate	Standard Error	t-statistic
Intercept	.0603	.0718	.8394
d_3	.00703	.00714	.9848
T	.00053	.00016	-3.2367
LnME	-.0228	.0115	-1.9887
σ	.0582	.372	1.565

R^2 = .313017

If the household head is a manual worker (d_3), expenditure on pork, ham and bacon (PIGR) is affected insignificantly. On the other hand, an increase in T (Take- away) expenditure affect pork, ham bacon (PIGGR) purchases positively and significantly although the magnitude of the increase in PIGGR purchases is very small (only .05%). For a given level of real total meat expenditure, more take- away cause a grater share of pork, ham and bacon to be purchased. Probably household prepare more park- based foods at home than they eat outside. The expenditure on pork, ham and bacon (PIGGR) is negatively and significantly influenced by and increase in total meat expenditure. A rise in the real total meat expenditure seems to cause a rise in expenditure of some other meats at the cost of a decrease in PIGGR expenditure. The selectivity variable has an insignificant impact.

Table-4

PICS: (Processed meat)

	Parameter estimate	Standard error	t-statistic
Intercept	.1337	.0732	1.824
d_3	-.0219	.00729	-3.005
T	.00028	.000168	1.679
LnME	-.0632	.0117	-5.381
σ	.0828	.0414	2.001

R^2 = 0.33236

Occupation (d_3) has a significant adverse impact on processed meat (PRCS) purchases. Manual works reduce their expenditure on processed meat (PRCES) highly significantly. For a given level a total meat expenditure (ME), a manual worker reduces his or her purchase share of processed meat as compared to other processions. Take-ways (T) affects processed meats (PRCS) purchases insignificantly. If the real total meat expenditure increases, the expenditure share on processed meats (PRCS) decreases by 6.3% and highly significantly. This shows a shift away from PRCS to other meats as is the case with PIGGR. In this case, the variability in the probability of buying PRCS between households impacts on the amount spent, significantly and positively. A rise in the probability of buying this meat or a given household cause and 8.2% rise in the amount spent and this is statistically significant. Thus, at fixed levels of other factors, the more probable (as induced by personal characteristics) a household is to buy processed meat. More will be the share of this meat type.

Table-5

POUL: (Poultry)

Parameter estimate		Standard error	t-statistic
Intercept	.2134	.0619	3.448
d_3	.00251	.00616	.408
T	-.00013	.000142	-.927
LnME	.0301	.0009930	3.032
σ	.0805	.032115	2.506

$R^2 = .26645$

Poultry (POUL) purchases are unaffected by d_3 (manual worker). Take-away (T) has and insignificant impact on poultry (POUL) purchases as well. For a given level of real total meat expenditure (ME), on socio-demographic factor affects the purchase share of poultry. While at given levels of other factors, an increase in real total meat expenditure affects poultry purchases positively and highly significantly. For an increase in real total meat expenditure. There is a 3.01% increase in the share of poultry (POUL) purchases. Selectivity appears to be positive and highly significant for this seat category. If the probability of buying poultry (POUL) increases, amount of the POUL purchase share increases significantly by 8.05%.

Table-6

FISH :

Parameter estimate		Standard error	t-statistic
Intercept	-.0314	.070379	-.447
d_3	-.00820	.007007	-1.170
T	.000152	.000162	.941
LnME	-.0684	.01128	-6.067
σ	.1006	.3650	2.757

R^2 = .26547

The occupation of the household head (d_3) does not significantly affect the expenditure share of fish (FISH). Take-away expenditure (T) had and insignificant effect as well. For a given level of real total meat expenditure, no socio-demographic variable affects the purchase share on fish. Forever, if real total meat expenditure increases, it highly significantly affects the fish share and decreases expenditure on fish (FISH) causing a shift to other meats. However, there is a non-trivial effect of variation in probability of buying fish between households. An increase in the probability that a given household will buy fish increases the share (10%) spent on it. This effect is statistically significant as with processed meat and poultry.

Now, we produce the parameter estimates of the equation for the meat type all other (ELSGR) which has been residually estimate keeping track of parametric restrictions imposed in the structural model as highlighted in Chapter III.

Table-7

ELSGR : (All others)

	Parameter estimate	Standard error	t-statistic
Intercept	-.01050	.0668	-.152
d_3	.04379	.0089	4.885
T	-.00035	.00013	-2.46
LnME	.03701	.0192	1.982
σ	.01236	0.0570	2.168

$R^2 = 0.4985$

Manual workers (d_3) do significantly favor meat type all others (ELSGR). If the household head is a manual worker, the consumption expenditure on this meat category increases by 4.3% as compared to not being a manual worker.

Expenditure on take-away foods has significant negative impact on this meat type although the decrease is slight, only .05%. An increase in real total meat expenditure causes a significant rise in expenditure share o meat type all other (ESGR). We note that ELSGR includes canned meat, Game, Offal and meat undefined. It may be that when household's purchasing capacity increases, more inducement is gained to consume this meat type. This same type of interpretation holds for the selectivity variable. We also notice that the response of the selectivity variable to consume meat type ELSGR is about three and half times more than that of real total meat expenditure.

We present a comparative analysis of different meat categories in terms of the significant impacts of different variables in cluster 1 in Table 6 below.

Table-8

IMPACT OF VARIABLES IN PERCENTAGES IN CLUSTER 1:

Meat types	Significant variables		
	d_3	T	In ME
BEV	+1.6	0	+ 8.2
MUL	0	- .06	+ 3.9
PIGR	0	+ .05	- 2.3
PRCS	- 2.2	0	- 6.3
POUL	0	0	+ 3.0
FISH	0	0	- 6.8
ELSGR	+ 4.3	- .03	+ 3.8

Out of 7 meat categories, processed meat (PRCS), is negatively and significantly affected while beef and veal (BEV) and all others (ELSGR) is positively and significantly affected if the household head is a manual worker. We can say that as compared to not being a manual worker, for a given level of real total meat expenditure (ME), a manual worker will favor bee and veal (BEV) and all others (ELSGR) and disfavor processed meat. An increases in expenditure on take-away (T) causes a significant reduction in the purchase of mutton and lamb (MUL) and all others (ELSGR), leading to a significant increase in pork, ham and bacon (PLGGR) consumption.

For a given level of all other factors, an increase in real total meat expenditure (ME) causes a significant rise in the purchase of beef and veal (BEV), mutton and lamb (MUL). Poultry (POUL) and all others (ELSGR) and a significant decrease in other meat types. The highest increase (8.2%) due to an increase in real total meat expenditure (ME) is for beef and veal (BEV) and the lowest (3%) is for poultry (POUL). The highest decrease (6.8%) for a rise in real total meat expenditure (ME) is for the fish (FISH) and the lowest decrease (2.3%) is

for pork, ham and bacon (PIGGR). The decrease in processed expenditure (ME).

For a quick comprehension of significant variables, we present the number of times a variable turns out to be significant at 5% level among meat categories in Table -7 below.

Table-9

FREQUNCY TABLE FOR SIGNIFICANT VARIABLES IN CLUSTER 1:

Variables	**Mat types**							**Frequency**	
	BEV	MUL	PIGGR	PRCS	POUL	FISH	ELSGR	+	-
d_3	*+			*-			*+	2	1
T		*-	*+				*-	1	2
InME:	*+	*+	*-	*-	*+	*-	*+	4	3

* Indicates significant at 5% level, + indicates positively and – indicates negatively.

We notice Table 5.7 real total meat expenditure (ME) is the most influential variable for meat purchase decisions in cluster 1.It influences all the meat categories significantly, either positively or negatively which is not he case with any other variable. Expenditure on take-away (T) affects 3meat types significantly and so does the variable occupation of household head (manual worker) but only one meat type is common to these two variables in terms of significant impact.

We now present the estimation results for cluster 2:

Table-10

ESTIMATION RESULTS OF THE STRUCTURAL SYSTEM MODEL IN CLUSTR 2:

LOG OF LIKELIHOOD FUNCTION =3315.81

NUMBER OF OBSERVATIONS = 1552

BEV : (Beef and veal)

	Parameter estimate	Standard error	t-statistic
Intercept	- .0645	.0303	-2.125
D_4	-.0357	.0285	1.253
D_5	-.0104	.0134	0.789
T	-.0006	.0003	2.106
N_1	-.0183	.0171	-1.071
N_2	.00261	.0157	.166
LnME	.0432	.0174	2.475
σ	0.457	.056274	.813

R^2 = .268789

The variable d_4 (freezer or microwave ownership) impacts on beef and veal (BEV) purchase insignificantly. d_5(house ownership) also has an insignificant effect. Expenditure on T (take-away) has the expected sign and its impact is minor but significant. If at a given level of household members in the age group 5<15 years (N_1) increases by one ,the expenditure share on beef and veal (BEV) is insignificantly affected and so is the case with the increase of a member of the age group 15-64 years(N_2) . If total real meat expenditure increases, the expenditure share on beef and veal (BEV) increases by 4.3% and significantly. The effect of variation in probability of buying beef and veal (BEV) between households is trivial. For a given level of real total meat expenditure no socio-demographic factors cause a significant rises in the purchase share of beef and veal.

Table-11

MUL: (Mutton and lamb)

	Parameter estimate	Standard error	t-statistics
Intercept	-.1204	.0678	-1.775
D_4	-.0611	.0260	-2.347
D_5	-.0099	.0120	-.825
T	-.000902	.000261	-3.456
N_1	-.00527	.0156	-.337
N_2	-.007008	.014	-.487
LnME	.0783	.0159	4.909
σ	0.1660	.0514	3.230

$R^2 = .32198$

If the household owns a freezer or microwave (d_4), the expenditure share on mutton and lamb (MUL) decreases by 6.1% as compared to not owning and is statistically significant. Probably the storing tendency is for other meats and so is the case with easy cooking methods. Ownership of a house (d_5) impacts on mutton and lamb (MUL) purchase insignificantly. Expenditure on T (Take-away) has the expected sign and a significant impact on the purchase of mutton and (MUL). The greater T the less is the MUL expenditure. Both N_1(members between 5<15 years) and N_2 (members between 15-64 years) have an insignificant impact on mutton and lamb (MUL) purchases. An incase in real total meat expenditure affects mutton and lamb (MUL) purchases positively and highly significantly. For an increase in real total meat expenditure, the expenditure share on MUL increases by 7.8%. The selectivity variable appears to be positive and significant. This implies that the impact of variation in the probability of buying mutton and lamb (MUL) between households is non-trivial and an increase in probability of buying increases consumption of MUL substantially and significantly which is also true with the rise in real total meat expenditure. For a given level of real total meat expenditure, on socio-

demographic factors cause a significant rises in the purchase share of mutton and lamb.

Table-12

PIGGR: (Pork, ham and bacon)

s	Parameter estimate	Standard error	t-statistics
Intercept	-.1801	.1073	-1.677
D_4	.0381	.0247	1.540
D_5	.0208	.0114	1.816
T	.000528	.000248	2.125
N_1	.0482	.01487	3.243
N_2	.0353	.01367	2.587
LnME	.0805	.02907	2.768
σ	0.1934	.063766	3.03431

R^2 = .411011

Both d_4 (ownership of freezer or microwave) and d_5 (ownership of a house) have an insignificant impact on pork, ham and bacon (PIGGR) purchase shares. The variable take-away (T) has an unexpected sign but a significant influence. It may be that and increase in take-away (T) for non-PIGGR leads to increase in PIGR purchase. An increase in the number of household members of both the age groups (5<15 years and 15-64 years) affect pork, ham and bacon (PIGGR) purchases positively and significantly. For a given amount of real total meat expenditure (ME), an increase in the number of members of age group 5<15 years causes a 4.8% rise in the purchase share o pork, ham and bacon(PIGGR) while such rise is 3.5% for and extra member o age group 15-64. If real total meat expenditure increases, pork, ham and bacon (PIGGR) expenditure shares increase by 08.05% and significantly.

CHAPTER VII
CONCLUTION

Main thrust of our endeavor in this thesis was to demonstrate the construction of a Multivariate Binary Choice Model. We have also highlighted how such a model is different from Multinomial Choice Model.

(A) In this thesis our main interest is in constructing a multivariate binomial choice model. In doing so we attempt to circumvent the shortcomings with the traditional argument that the multivariate choice problem can be coSSnverted to multinomial one and then to apply all statistical properties of multinomial analysis.

In this thesis we have attempted to demonstrate the construction of a multivariate binomial choice model keeping in view that although univariate choice model is frequently observed, multivariate framework is rare. We have also discussed few variants of resolving estimation issues for such a complex model. We do hope that such a model construction would be very useful to practitioners to resolve practical problems. Application of such a model to practical situation would to a large extent ease decision making process, We have a focused an intnicaeies associated with estimation testing problems. Several attempts have been made in the post in this regard. We believe that McFaddens MSM is a pioneering attempt in this regard although computationally it is a very hard task. However in modern age of computing facilities such a task does not need to be a great burden. The main concern should be the construction and application of such a lucid model for analysing decision making behavior.

(B) However, although elegant in formulation as well as from practical viewpoint, choice models pose severe estimation problems in addition to complex testing issues. There have been various attempts to circumvent such

problematic issues. In this thesis we have attempted to provide a picturesque idea on such attempts along with their comparative assessment.

In line with the foregoing assertion it is believed that such a modeling strategy will open up frontiers for applied researchers to focuses and analyse practical situations more elegantly and prominently.

Further horizon of research :

It is an accepted truth that research works do not have any end and same is the case with the present work. Construction of multivariate multinomial model is an interesting avenue. It becoures more interesting when one considers estimation and testing issues associated with such complex types of models. Such models can definitely reflect real world scenarios in a befitting manner. Another interesting feature of such types of models is their practical applications and adaptability construction and adoption need apt, tact, wit and promptness of mind. An interesting example can be cited in the field of selection of peoples representatives. A voter necessarily considers a couse bination of characteristics of a representative for voting. Now several characteristics of a voter influence him to make a choice of such combinations. An elegant analysis of such scenarios can be performed in the framework of the present type of multivariate structure on more interesting avenue is to explore that a choice maker accounts for attributes of choices but choices also depend on attributes of choice makers themselves. However, we keep such tasks as future research and look forward to apply more and more in real world situations.

References:

1. Mallar, C.D. 1977. The Estimation of simultaneous Probability Modeles, Econometrica, Vol.45 no.7,1721-27.
2. McFadden,.D. 1989. A Method of simultated Moments for Estimation of Discrete Response Modeles, Econometrica, Vol.57 no.5, 5995-1026.
3. Larman, s,And C.Manski(1981): "On the use of simulated Frequencies to Approximate Choice probabilities" in Structual Analysis of discrete Data with Econometric Applications by C. Manski and D. McFadden, Cambridge: MIT Press.
4. Stern, S.1987. A Method for smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models ,Mimeo,University of Verginia.
5. Ball, R.J. (1968), "Econometric Model Building", in Mathematical Model Building in Economics and Industry, London: Charles Griffin & Co., Ltd.
6. Beach, E.F.(1957), Economic Models: An Exposition, New York: John Wiley & Sons, Inc.
7. Borsch-Supan , A.and Hajivassiliou, V.A. 1990. Smooth unbased multivariate probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models, Cowles Foundation Discussion Paper No.960,Yale University.
8. Borsch-SupanA.andHajivassiliou,V.A.1990.Smooth unbased MultivariateP analysis of discrete choice with applications on the demand for housing in the U.S. and West Germany,Springer-Verlang,N.Y.
9. Clark,C.(1961): "The Greatest of a Finite Set of Random Variables."
10. Daganzo,C.(1980): Multinomial Probit. New York:Academic Press.
11. Heckman,J.J.(1976a):The common structure of statistical models of truncation, sample selection of limited dependent variables and asimple

estimation for such models, Annalysis of Economic and Social Measurement 5.

12. Cain, G.G., and H.W. Watts, Eds. (1973), *Income Maintenance and Labor Supply.* New York: Academic Press, Inc.

13. Christ, C. (1966), *Econometric Models and Methods.* New York: John Wiley & Sons, Inc.

14. Christ, C.F. (1966), Econometric Models and Methods. New York: John Wiley & Sons, Inc.

15. De Groot, M. (1970), Optimal Statistical Decisions. New York: McGraw-Hill Book Company.

16. Dhrymes, P. (1970), *Econometrics.* New York: Harper & Row, Publishers.

17. Doman, E.D. (1957), Essays in the theory of Economic Growth. New York: Oxford University Press.

18. Enke, S. (1951), "Equilibrium among Spatially Separated Markets: Solution by Electric Analogue. "Econometrica, 19: 40-47.

19. Fisher W.D. (1961), "A Note on Curve Fitting with Minimum Deviations by Linear Programming," Journal of the American Statistical Association, 56: 359-62.

20. Fishman, G.S. (1969), *Spectral Methods in Econometrics.* Cambridge: Harvard University Press.

21. Granger, C. W. J., and M. Hatanaka (1964), *Spectral Analyses of Economic Time Series.* Princeton: Princeton University Press.

22. Graybill, F.A. (1961), An introduction to Linear Statistical Models. New York: McGraw-Hill Book Company.

23. Hamilton, H. R., S. E. Goldstone, J.W. Milliman, A.L. Pugh, E.R. Ronerts, and A. Zellner (1969), *Systems Simulation for Regional Analysis: An Application to River Basin Planning.* Cambridge: MIT Press.

24. Horowitz, I. (1963), "An Econometric Analysis of Supply and Demand in the Synthetic Rubber Industry." *International Economic Review,* 4: 325-45.

25. Intriligator, M.D. (1971b), *"Econometrics and Economic Forecasting."* in J.M. English, Ed., *The Economics of Engineering and Social Systems.* New york: John Wiley & Sons, Inc.

26. Kendall, M.G. (1968), "Introduction to Model Building and its Problems", in *Mathematical Model Building in Economics and Industry.* London: Charles Griffin & Co., Ltd.

27. Kogiku, K. C. (1967), "A Model of the Raw Materials Market. "International Economic Review, 8: 116-20.

28. Nerlove, M. (1964), "Spectral Analysis of Seasonal Adjustment Procedures," *Econometrica*, 32: 241-86.

29. Phelps, E.S., Ed., (1970). *Microeconomic Foundations of Employment and Inflation Theory.* New York: W.WE. Norton & Company, Inc.

30. sPhillips, A.W. (1958), "The relation between Unemployment and the Rate of Change of Money Wages in the United Kingdom, 1861-1957." Economica, 25: 283-99.

31. Rao, P., and R.L. Miller (1971), Applied Econometrics. Belmont, Calif.: Wadsworth Publishing Co., Inc.

32. Samuelson, Paul. (1947), Foundations of Economic Analysis. Cambridge: Harvard University Press.

33. Sewell, W.P. (1969), "Least Squares, Conditional Predictions, and Estimator properties," Econometrica, 37: 39-43.

34. Smithies, A. (1957), "Economic Fluctuations and Growth." Econometrica, 25: 1-52.

35. Whittle, P. (1963), Prediction and Regulation by Linear Least Square Methods. New York: Van Nostrand-Reinhold.

36. Williams, E. J. (1959), Regression Analysis. New York: John Wiley & Sons, Inc.

37. Zellner, A., Ed. (1968), *Readings in Economic Statistics and Econometrics.* Boston: Little, Brown and Company.

38. Labor force Participation: Winkkelmann (2000)

39. Choice of Jobs: Greene (2002)