Application of the logistic regression models for transportation data

Mokowe Rahab Makwela

Submitted in fulfilment of the requirements for the degree of **Master of Science**, in the
Department of Statistics and Operations Research
Faculty of Sciences and Agriculture
University of Limpopo

July 2007

Supervisor: Professor M.A. Lesaoana
University of Limpopo

Co-Supervisor: Professor V.S.S. Yadavalli
University of Pretoria

DECLARATION

I declare that the dissertation hereby submitted to the University of Limpopo for the degree at this or any other university is my own work in design and execution, and that all material contained therein has been duly acknowledged.

Signed:________________

Date:__________________

ACKNOWLEDGEMENTS

I would like to thank all the people who contributed their time in advising and providing assistance to me while working on this project. I am grateful to my supervisors Professor M.A. Lesaoana and Professor V.S.S. Yadavalli who helped me to make this project a success. I also portray my special thanks to Professor A. Szubarga and my colleagues in the Department of Statistics and Operations Research who gave moral support and assistance. I am grateful to SANPAD (South African-Netherlands Programme on Alternatives in Development) for funding this project. My special thanks go to Professor Del Mistro - the project leader of SANPAD in South Africa, who introduced me to this project. Statistics South Africa provided me with useful data from their Census 2001, which has added a lot of value to this project. Thank you Statistics South Africa.

My gratitude would be incomplete, if I cannot mention the name of Professor H.J. Siweya, the Director of the School of Computational and Mathematical Sciences at the University of Limpopo, who reminded me all the time that I have to complete my dissertation.

Eventually are due to my parents, brothers and sisters for their encouragement and financial assistance throughout my school life.

ABSTRACT

The multinomial model was fitted to the stated preference data of the transportation problem by commuters in Mamelodi, east of Pretoria. The data analyzed in the study was collected in 2001 among 151 less literate (highest level of education up to Standard 5 or Grade 7) and 194 literate (highest level of education from standard 6 or Grade 8 to Standard 10 or Grade 12) commuters in the CBD (Central Business District) of Pretoria. Seventeen (17) variables have been analyzed.

The objective of the study is to determine if there are differences when three types of codings (dichotomous, binary and effect) are applied to the same data. The final interest is to determine those factors that affect commuters in choosing their mode of transport to work in the CBD of Pretoria.

All the logistic regression models and multinomial logit models tested in the study were found to be statistically significant for the three different codings. Due the limitations that SAS has, the logistic regression models were fitted and used to carryout the analyses. When variables were selected by the stepwise procedure, and only those explanatory variables that were significant fitted in the model, the three models were all statistically significant for the Hosmer-Lemeshow goodness-of-fit statistics, but not for the Pearson and Deviance goodness-of-fit statistics.

TABLE OF CONTENT

Page

LIST OF TABLES

LIST OF FIGURES

KEYWORDS

CBD (Central Business Destination) of Pretoria

Deviance

Dichotomous, binary and effect codings

Goodness-of-fit statistics

Hosmer-Lemeshow statistic

Logistic regression

Multinomial logit model

Pearson Chi-square

SANPAD (South African-Netherlands Programme on Alternatives in Development)

Stated preference

Stepwise procedure

Wald Chi-square

Variance inflation factor (VIF)

Tolerance (TOL)

Multicollinearity

Likelihood ratio

Log likelihood function

Newton-Raphson algorithm

Score test

Hypothesis testing

Maximum likelihood

Diagnostics

Odds ratio

Less literate commuters

LIST OF VARIABLES

EDUC	Highest education level
PRES	Presentation method
MC	Minibus cost
MF	Minibus feeder
MSEA	Minibus seating
MSEC	Minibus security level
MTT	Minibus traveling time
TC	Train cost
TF	Train feeder
TSEA	Train seating
TSEC	Train security level
TTT	Train traveling time
BC	Bus cost
BF	Bus feeder
BSEA	Bus seating
BSEC	Bus security level
BTT	Bus traveling time

CHAPTER 1

INTRODUCTION

1.1 Background

The motivation to undertake this study emanates from the SANPAD (South African-Netherlands Programme on Alternatives in Development) project entitled "The applicability of stated preference among less-literate commuters", (Del Mistro, 2004). SANPAD is a programme initiated and sponsored by the Dutch Ministry of foreign Affairs with the aim of stimulating alternative academic research in the field of development in South Africa (www.sanpad.org.za, May 2007). SANPAD also facilitates pluralistic perspectives and practice in scientific research by establishing partnerships and strengthening collaboration between South African and Dutch academics. It is in this regard that the South African wing of SANPAD was led by Professor Del Mistro of the then University of Pretoria, and the Netherlands side was led by Professor Arentze of the Technical University of Eindehoven. A number of local scholars (including myself) also formed part of the SANPAD project team.

One of the major stumbling blocks to the development of good policies in South Africa is lack of (reliable) data. When the democratic government took power in 1994, no nation-wide census had been undertaken and a need to conduct one led to the first all-inclusive population census in 1996. As early as 1980, Morris and Vander Reis found that the diverse, cultural groups in South Africa vary in the range of qualifying adjectives used to distinguish levels of feeling in a scale of value judgments, for which they also observed a number of problems in the use of rating scales among less-literate persons, (Del Mistro, 2004). SANPAD projects are intended to develop methodologies that can aid governments to formulate policies and base their planning on community needs. In this regard democracy will be enhanced and maximum benefits will be derived from the optimal use of scarce resources.

Stated preference is one of the techniques that can be employed to measure people's perception of needs and possible solutions. In a stated preference study a set of alternative solutions is presented and respondents are asked to indicate their preference. The technique assumes that an individual makes a choice on the basis of the trade-off between alternative choices provided.

Modal choice model in the transportation planning process consists of a hierarchy of decision-making structures. A distinction is made between *revealed preference* approach and stated preference. The former seeks to measure what individuals actually do and as such often limits future planning. The latter, i.e. stated preference, which is of interest to this study, provides an opportunity for individuals to make choices about hypothetical options.

The stated preference method is based on random utility theory and attempts to estimate the probability that a person will choose a given alternative based on that person's socio-economic characteristics as well as convenience of the option. Modal choice preferences could include walking, cycling, private car, minibus (taxi), bus or train. The afore-mentioned SANPAD project (Del Mistro, 2004), was aimed at determining efficient and cost-effective methodologies that ensure the validity and reliability of socio-demographic data collection for policy development in the democratic South Africa, focusing on the mode of transportation choice by less-literate commuters living in the urban and peri-urban areas around Pretoria.

The SANPAD project is a multidisciplinary study across the psychological, anthropological and statistical fields. It is the statistical perspective for which the contribution of this dissertation is made. The statistical component of the SANPAD study plays a supportive and complementary role to both the psychological and anthropological views, and its aim is to determine the statistical aspect of conclusions on the impact of modal choice interviews and methodologies adopted in decision-making by less literate respondents.

Application of multinomial logit and logistic regression to stated preference studies is not a new phenomenon. On multinomial regression, Elango and Sambharya (2004) used the multiple logistic regression model to test 336 entry decisions from 18 countries entering the United States over the period 1989-1994. Their study examined the impact of industry structure on the foreign direct investment entry mode decisions by multinational enterprises. Street and Burgess (2004) studied stated preference choice experiments on the optimal choice sets to use when either all choice sets are to contain a common base alternative or when all choice sets contain a "none of these" option. Zandvliet et al. (2006) applied multinomial logistic regression to investigate the relationship between the space-time ecologies of different types of visitor population environment in the

Netherlands and destination choice. Loo (2007) analysed the airport choice of passengers departing from Hong-Kong International Airport to 15 destinations in different parts of the world. Varghese et al. (2007) estimated the level of quality of life and its determinants among diabetic subjects in Thiruvanthapuram, Kerala, India. In this study a response to each question had a score ranging from 1 to 5. Espino et al. (2007) used multinomial logit and mixed logit models in a stated preference study to determine the most important route connecting the Canary Islands archipelago with the Iberian Peninsula.

On the application of logistic regression to transportation choice models, Phipps (1984) carried out a centro-graphic analysis to compare the residential search and choice behaviour of 41 households who experienced either short-term or long-term displacement costs after moving out in the inner city of Saskatoon, with the behaviour of 90 households who moved as if voluntarily. Cox et al. (1999) used logistic regression and ordered probit models to assess the overall preferences for rabies-prevention policies and the importance of policy attributes and socio-economic characteristics in determining policy preferences. Young et al. (2003), published their work on the identification of factors associated with mode of transport to rural hospitals, for which 11 541 trauma patient visits that came by ground ambulance or private vehicle to the Emergency Department of one of the six rural hospitals in northwest Iowa were analysed using univariate analyses and logistic regression. Yannis et al. (2005) conducted a stated preference study by applying logistic regression to examine the behavioural parameters that influence the driver's choices in order to reduce the accident risk. Van Wezel and Potharst (2007) studied various ensemble learning methods for machine learning and statistics applied to the customer choice modelling problem, for which the logistic regression model was applied. Duerksen et al. (2007) applied logistic regression to test whether the type of restaurant a family visits most often is associated with the body mass index. Sze and Wong (2007) evaluated the injury risk of pedestrian casualties in traffic crashes in Hong Kong and explored the factors that contribute to mortality and severe injury using binary logistic regression. The authors verified the goodness-of-fit for the proposed model by means of the Hosmer-Lemeshow test and logistic regression diagnostics. Akerstedt et al. (2002) applied a multiple logistic regression model using SAS (version 6.12) to study the relationship between work and background factors on the one hand, and disturbed sleep and fatigue on the other.

In the SANPAD project the stated preference study is conducted among the less literate commuters in Mamelodi, east of Pretoria. The dissertation approaches the stated preference problem by applying statistical models by developing a SAS program to perform the analysis.

In Sections 1.2, 1.3, 1.4 and 1.5, key indicators of South Africa, Gauteng, Pretoria and Mamelodi are studied and analysed. These indicators are useful in understanding the characteristics of the commuters who responded to the SANPAD survey. The data set analysed in the dissertation is described in Section 1.6, and the introductory chapter is wrapped by providing the objectives of the study in Section 1.7 and the hypothesis in section 1.8.

1.2 Key Indicators: South Africa

Located in the southern tip of the African continent, South Africa is divided into nine provinces, and is occupied by about 45 million people according to Statistics South Africa's census 2001. South Africa has a total land of 1 219 090 km^2. The aim of this section is to give a brief overview of some of the social, demographic and economic indicators in South Africa.

The Western Cape and Gauteng are the two highly industrialized provinces in South Africa. Gauteng that hosts Pretoria, the capital city of South Africa, has the smallest area of only 1.4% (the second smallest, Mpumalanga, has an area of 6.5%). In terms of population, Gauteng carries the second largest (19.7%) to KwaZulu-Natal (21.0%) that has the largest number of South African people. The Northern Cape with the largest landscape (29.7%) is the most sparsely populated province occupied by the least number of people (only 1.8% of the population in South Africa), the second least in population being the Free State with 6.0% of the South African population.

There are eleven official languages spoken in South Africa. The predominant language is IsiZulu spoken by 23.8% of South Africans, mainly in KwaZulu-Natal, where 80.9% of the people speak isiZulu as a home language. Most people in Gauteng (21.5%) also speak isiZulu as a home language, (Statistics South Africa, Census, 2001, also www.statssa.gov.za). With Sepedi, Sesotho and Setswana fairly close, a combined 32.2% of the people of Gauteng speak these three languages at home, while 14.4% and 12.5%

speak Afrikaans and English, respectively.

1.3 Key Indicators: Gauteng

Of the nine provinces, a special interest goes to Gauteng since this study is based on data on the mode of transportation, collected in the neighbourhood of Pretoria, in Gauteng. The word "Gauteng" is derived from a "Sotho" phrase, meaning "Place of Gold". This province has traditionally been known for gold mines that attracted men, (with little education) from mostly rural areas of South Africa, and the neighbouring SADC (South African Development Community) countries.

Gauteng is further demarcated into three District Councils: Sedibeng, Metsweding and West rand, and three metros: City of Johannesburg, City of Tshwane (Pretoria) and Ekurhuleni (East Rand). Gauteng is the only province for which the proportion of male population is higher than that of the female population. Gauteng has a total population of 8 837 178 (Statistics South Africa, Census 2001), of which 53.4% are male. The largest number of people in Gauteng live in the City of Johannesburg (34.4%), followed by Ekurhuleni (East Rand) at 26.4%, and then Pretoria (City of Tshwane) at 21.1%. The same sequence applies to the number of households.

Table 1.1 shows the distribution of the population of Gauteng by gender and number of households. The data in Table 1.1 is extracted from one of the many publications by Statistics South Africa (Stats SA) on the results of their latest population census conducted in 2001.

Table 1.1: The population of Gauteng, by gender and number of households

Municipality	**Male**	**Female**	**Total**	**Total %**	**House-holds**
Johannesburg	1 607 014	1 618 799	3 225 813	34.4	1 000 932
Pretoria (Tshwane)	979 184	1 006 799	1 985 983	21.1	562 654
Ekurhuleni (East Rand)	1 258 519	1 221 758	2 480 277	26.4	744 936
Sedibeng	391 630	402 975	794 605	8.5	225 099
Metweding	83 815	76 076	159 891	1.7	44 392
West Rand	400 151	344 003	744 154	7.9	207 675
Total	**4 720 313**	**4 116 865**	**8 837 178**	**100.0**	**2 578 013**

Source: Key municipal data, Stats SA Census 2001
Pretoria (City of Tshwane) comes third in the province of Gauteng both in terms of the population and the number of households. Despite developments expected in Gauteng, in particular Pretoria, key indicators tell a different story.

1.4 Key Indicators: Pretoria (City of Tshwane)

According to key municipal data published by Statistics South Africa (Stats SA) on the basis of their Census 2001 results, the City of Tshwane (Pretoria) has 562 654 households and a total population of 1 985 983, disaggregated as in Table 1.2.

Table 1.2: Population of Pretoria, by population group and gender, and number of households

Population Group	**Pretoria Population**				**Number of Households**
	Male	**Female**	**Total Number**	**Total %**	
Black African	716 850	725 728	1 442 578	72.6	390 532
Coloured	18 400	20 321	38 721	1.9	9 871
Indian or Asian	15 084	15 047	30 131	1.5	7 432
White	228 850	245 703	174 553	8.8	154 817
Total	**979 184**	**1 006 799**	**1 985 983**	**100.0**	**562 652**

Source: Key municipal data, Stats SA Census 2001

Monthly imputed household income shows that 28.7% of the households in Pretoria earn no more than R800 (an equivalent of about US $130 by the 2001 exchange rates).

Nearly a quarter (24.5%) of the households in Pretoria lives in either informal or traditional dwelling houses (usually with poor services and facilities). About 20% of the households do not have electricity (80.6% use electricity for lighting); and less than 80% of the households have access to piped water. In fact more than 6 000 households access water from spring, rain water tank, dam/pool/stagnant water, or water vendor. Regarding sanitation, 2.6% of the households in Pretoria have no toilet facilities, and 28.0% use pit latrine, bucket latrine, or have no toilet facilities; 3.8% have no rubbish disposal. The majority 69.6% of the households in Pretoria enjoy the usage of flush toilets.

Nearly 32% of the economically active population in Pretoria are officially unemployed; and 12.0% are informally employed. For those aged 5-24 years, 28.1% are not attending school; and for those aged 20 years or more, 8.3% have no schooling; while 17.0% qualify beyond Matric (Grade 12).

On the question probing mode of travel to school or place of work, for which 1 156 990 responses were obtained, 33.4% travel on foot (walk); 20.2% drive a car; 13.2% take a taxi or minibus; 11.9% travel in someone's car; 11.8% travel by bus; 7.1% travel by train; 1.2% cycle; and the remaining 1.2% use motorcycle or other means other than those mentioned above. The category "drive a car" assumes that the driver owns the car, while "car as a passenger" assumes that a person is using a car as a mode of transport, but as a passenger, and does not necessarily own the car.

Of key interest to this study is Mamelodi in Pretoria, the key indicators of which are analysed in Section 1.5 below.

1.5 Key Indicators: Mamelodi

Mamelodi occupies a total land area of 48.7km^2. Table 1.3 shows that Mamelodi has a total population of 256 118 with more bias towards males (Stats SA's population Census, 2001). The Black (African) population constitutes 99.6% of the population of Mamelodi . Compared to the entire population of Pretoria (City of Tshwane) with 72.6% Black African, the population of Mamelodi is predominantly Black. The interest in the proportion of black population is that during the long history of apartheid in South Africa, mostly locations occupied by black people were deprived of development facilities, and this history has been inherited by the democratic South African Government, that took power in 1994.

Table 1.3: Population of Mamelodi, by population group and gender

Area	Mamelodi Population				Total % Pretoria
	Male	Female	Total Number	Total %	
Black	131 146	123 838	254 984	99.6	72.6
Coloured	488	574	1 062	0.4	1.9
Indian	4	6	10	0.0	1.5
White	32	30	62	0.0	8.8
Total	131 670	124 448	256 118	100.0	100.0

Source: Stats SA, Census 2001, and own calculated percentages

Most households (58.0%) in Mamelodi speak Sepedi and related languages (Setswana and Sesotho); followed by 21.9% who speak IsiNdebele, XiTsonga and TshiVenda; and 19.1% who speak IsiZulu, IsiXhosa or SiSwati, as their home languages. Afrikaans and English are spoken by 0.7% and 0.2% of Mamelodi households, respectively. Thus an overwhelming 98.9% of Mamelodi residents speak traditional African languages, compared with 72.0% in the Gauteng Province. This difference is acknowledged given the fact that the residents of Mamelodi are predominantly Black (African).

Mamelodi as a traditionally black location deprived of basic services for many years prior to the election of the (new) democratic government in 1994, still has a sizeable number of households living under poor conditions. Nearly 40% of the households in Mamelodi live in informal dwelling (39.8%), and these together with those who live in traditional dwelling houses constitute 41.2%.

In Table 1.4, the population of Mamelodi is reflected by specific area in terms of gender and the number of households. Mamelodi East, Mamelodi West, and Mahube Valley, in that order, have recorded the largest number of people. Mamelodi West seems to be having a smaller household size than Mamelodi East since it shows a large number of households, yet smaller in population size than Mamelodi East. Overall, Mamelodi has a total of 68 443 households. Compared with Pretoria, Mamelodi has an average of 3.74 persons per household while Pretoria has household average of 3.53 persons. Again, Mamelodi has relatively larger household size than its parent District Council, Pretoria (City of Tswane Metro).

The 68 443 households in Mamelodi are concentrated in Mamelodi West (31.8%), followed by Mamelodi East (27.3%), and then Mahube Valley (19.1%). In all the Mamelodi extensions except Mamelodi East, the proportion of males is higher than that of females as observed from Table 1.4.

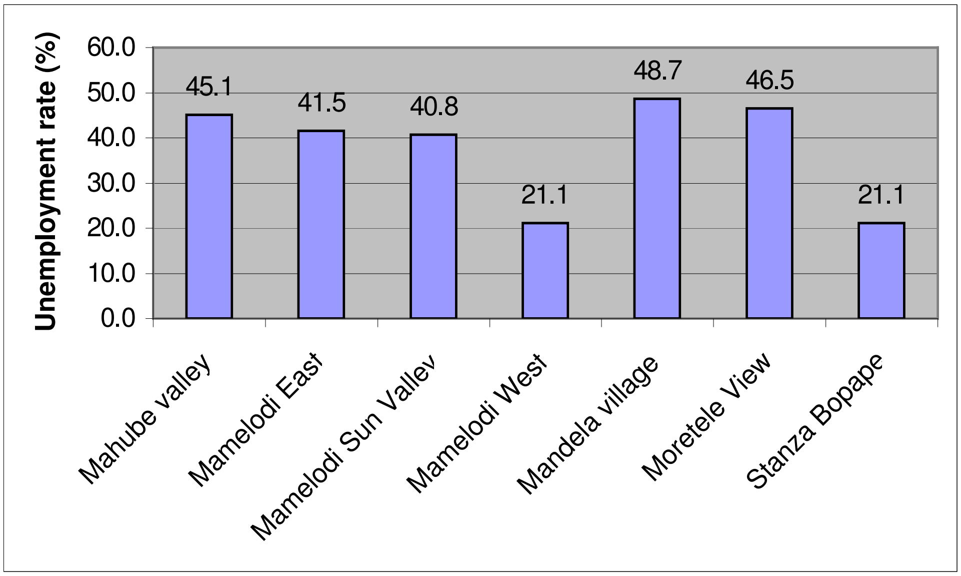
Table 1.4: Persons in the household (weighted) by Population group and Mamelodi areas

Specific Area	Male	Female	Total	Households (%)
Lusaka	936	781	1 717	0.8
Mahube Valley	23 975	22 498	46 473	19.1
Mamelodi East	40 262	41 740	82 002	27.3
Mamelodi Sun Valley	1 069	1 033	2 102	0.8
Mamelodi West	39 233	33 347	72 580	31.8
Mandela Village	10 328	9 792	20 120	8.0
Moretele View	830	822	1 652	0.5
Stanza Bopape	15 037	14 433	29 470	11.7
Total	131 670	124 446	256 116	100.0

Source: Stats SA, Census 2001, and own calculated percentages

Figure 1.1 displays the *official unemployment rates* in Mamelodi, by extension (or specific location). In 2001, Stats SA defined the officially unemployed as *those within the economically active population who did not work during the seven days prior to the interview; want to work and are available to start work within a week of the interview; and have taken active steps to look for work or to start some form of self-employment in the four weeks prior to the interview* (Stats SA, Census 2001).

Figure 1.1: A bar chart of unemployment rate by Mamelodi areas



Source: Own calculations

Most areas in Mamelodi record unemployment rates in excess of 40%, with Mandela Village, Moretele View and Mahuve Valley exceeding the 45% margin. Mamelodi in general, has an official unemployment rate of 44.0%, compared with 31.9% in Pretoria. Unemployed rate is used as a proxy to determine the extent of poverty.

Table 1.5 provides data on the highest educational level attained by the residents of Mamelodi.

Table 1.5: Mamelodi vs Pretoria residents' education level (%)

Education level	**Percentage %)**	
	Mamelodi	**Pretoria**
No schooling	10.0	7.2
Some primary	11.8	10.1
Completed primary	6.2	5.4
Some secondary	35.5	34.9
Std 10/ Grade 12	30.2	28.7
Higher than Grade 12	6.9	13.8

Source: Own calculations

The highest percentage (35.5%) of Mamelodi residents has passed some secondary education as their highest education level. This measure of education level applies to all those aged 20 years and more. The persons with Standard 10/Grade 12 as their highest education level constitute 30.2% in Mamelodi, and the percentage of people who did not go school at all is 10.0%. For those with highest qualification above Matric (Grade 12), there is only 6.9% of Mamelodi residents compared with 13.8% of Pretoria in general. In fact Mamelodi has 37.1% of their residents with Grade 12 and higher as their highest qualification, compared with 42.8% of the City of Tswane (Pretoria) residents. The residents of Mamelodi are generally less literate than those of Pretoria as a whole, and seem to be trapped between secondary education and Matric (Grade 12 or Standard 10).

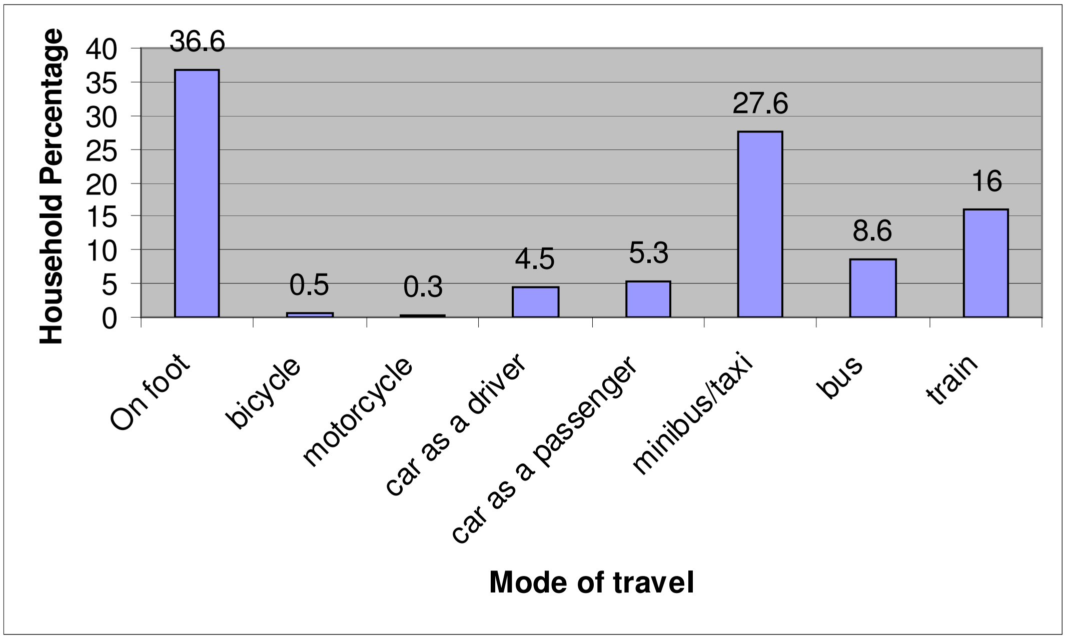
A sizeable number of households in Mamelodi earn no income at all (19.1%) and 34.4% earn only up to R800 per month. Compared with Pretoria where 28.7% earn up to R800 a month, the people in Mamelodi are relatively disadvantaged.

Most of the households in Mamelodi (72.4%) use electricity for lighting of which about 60% of the electricity users are from Mamelodi East. Despite the development that one would expect from Mamelodi households as part of the City of Tshwane (Pretoria), a relatively high percentage of the households still use candles and paraffin for lighting, (making a combined total of 27.1%). Some 30.1 of the households in Mamelodi use paraffin for cooking, and just over half a percent use either wood or animal dung for cooking (0.6%).

Compared with Pretoria in general, where 91.2% of households have their rubbish disposal removed by local authority at least once a week, 75.0% of Mamelodi residents fall in this category. Some 3.0% of Mamelodi households have no rubbish disposal, and 12.6% rely on own refuse dump. More than three-quarter (76.7%) of the households in Mamelodi use flush toilet, and 19.3% use pit and bucket latrine.

Figure 1.2 shows that most residents of Mamelodi walk (36.6%) or take a minibus/taxi (27.6%) to school or place of work. The train and bus follow in popularity at 16.0% and 8.6%, respectively.

Figure 1.2: Mode of travel to school or place of work by the residents of Mamelodi



Only 4.5% of Mamelodi residents drive a car to a place of work (or school), compared with 20.2% of all residents in Pretoria. The remaining 0.6% of the Mamelodi households use other means of transport (other than the ones shown in Figure 1.2).

The indicators analyzed in the previous sections show that the residents of Mamelodi are in general less advantaged than most parts of the entire City of Tswane (Pretoria), or Gauteng Province, in general. These findings are in terms of access to basic facilities (water, energy, education, housing), and in terms of economic and labour market indicators (employment, mode of transportation).

The remaining sections of Chapter 1 provide details of the sources of data used in the research study (Section 1.6). In Section 1.7 the objectives of the study are defined, and finally in Section 1.8 the hypothesis to be tested by the analysis of the study is formulated.

1.6 Source of Data

The data used in the study comes from secondary sources. In 2000/2001 a study on the *stated preference choice* in transportation was conducted by the University of Pretoria (South Africa) in collaboration with the Technical University of Eindhoven (Netherlands), under the auspices of SANPAD. The target area was the CBD (Central Business District) of Pretoria, in particular Mamelodi, east of Pretoria. The objective of the study was to determine if stated preference methodology could be applied among less-literate commuters. In that study less-literate commuters are defined as those having achieved as their highest qualification Standard 5 (Grade 7). In the South African context this definition is equivalent to primary education. The literate commuters are considered to have passed as their minimum highest level of education, Standard 6 (Grade 8), and they therefore have at least secondary education.

The SANPAD study focused on three types of transportation to work by commuters: train, bus, and minibus (taxi). The study define “stated preference” as a statement by an individual of his or her liking for one alternative mode of transport over another. The questionnaire was partitioned into six sections. The first part sought basic administrative questions. The second part sought place of residence, work place, transport mode selected, length of trip, transportation fare, and reasons for choice of selected mode. In the third part questions on stated preference were introduced. The fourth part comprised 16 treatments, half of which were in verbal format and the other half in pictograms. In the fifth part respondents' assessment of the questions on stated preference, and preferred method of presentation, were sought; as well as socio-economic questions such as age, gender, marital status, education level, household income, and how long respondents have been living at their place of residence.

Data was collected from 151 less-literate and 194 literate respondents. We have seen from Table 1.5 that 22.7% of Mamelodi residents can be classified as less-literate while the majority 77.3% is literate. Furthermore, the study took place in 2001 when the census was undertaken. While slightly more respondents were female in both groups of the less-literate and literate commuters, it is noted that 28.5% of the former were aged (over 25 years of age) as against 9.3% of the literate in the same age category. In South Africa the less literate are mostly in the older age group, hence the less-literate sample has been negatively biased in this regard.

1.7 Objectives of the study

The project will determine whether or not different codings of the same data will generate the same results. Three different binary codings ("0" and "1"), ("1" and "2") and ("-1" and "1") will be used for the same data and the results produced by each set will be studied, for which differences, if any, will be determined.

The secondary objective of this research project is to determine those factors that make commuters to prefer one mode of transport over another.

If one needs to test the claim that a particular model is a good fit to a given data, an appropriate test statistic is determined which would help to agree or disagree with the claim. Not only one statistic need be calculated, but several (possible test statistics) of them would be computed. Therefore, the test about the claim that a particular model is a good fit or not to a given data will be done using different measures calculated from the same data set.

1.8 Hypothesis

In this study we are testing the hypothesis that there is a difference between literate and less literate persons in the stated preference choice (with specific reference to mode choice).

Since some respondents are less literate and others literate, this study will assess whether or not there is a difference (e.g. in cost and travelling time) between literate and less literate persons in the stated preference choice.

Methodology, definitions of variables, the analysis and interpretation of data and conclusion will be discussed in Chapters 2, 3, 4 and 5 respectively. Chapter 2 presents the models to be used in the project, model fitting, different measures of goodness-of-fit and comparison of the means for two samples.

CHAPTER 2

METHODOLOGY

2.1 Introduction

When one is interested in statistical models that assess the effect of categorical variables on a dichotomous (or binary) response variable, one is often led to the formulation of the logit model for the response variable, especially if all other variables that one wishes to manipulate are categorical variables. If the response variable has more than two categories, the model is said to be the multinomial logit model.

Both logistic regression and multinomial logit models shall be used in this study. The multinomial logit model is an extension of the logistic regression model when the dependent variable has more than two categories. Since the dependent variable in this project has three categories, the multinomial logit model will be used. The logistic regression model is then used to carry out further analysis, because of the limitations that SAS version 8 has on the CATMOD procedure. All the variables included in the model (using CATMOD procedure) will be included in the multinomial logit model. Thus the binary logistic regression models shall then be used for the stepwise selection and also the diagnostics.

In this project the response variable has three categories which led to the multinomial logit model. The multinomial logit model is used in this project to model transportation mode choice among less-literate persons. The multinomial logit model uses all the data set of which it may not be easy to select the variables to be included in one model. This led to the use of the logistic regression model to select the significant variables by the stepwise selection method. The diagnostics were also done using the logistic regression model.

Of interest is to investigate if the use of different dummy codings of the transportation data change the parameter estimate, significance of the independent variables, significance of the model and the test statistics as well. This project is concerned with different dummy codings of the data and different summary measures of goodness-of-fit statistics. It will also determine whether or not binary coding and effect coding of the same data set produce different results. This project is also concerned about using different goodness-of-fit test statistics to check the in/significance of the model.

2.2 The logistic regression model

Linear regression analysis is usually applied to models in which a dependent variable is regressed on one or more independent variables. In this case the dependent variable is always continuous. This is one of the univariate techniques where the research problem involves a single dependent variable and one or more independent variables. This technique has four assumptions: linearity, constant variance (homoscedasticity) of the error terms, non-correlation of the error terms, and normality of the error terms.

The binary logistic regression is a form of regression which is used when the dependent variable is binary (or dichotomous) and the independent variables may be categorical, continuous or both. The logistic regression analysis is a technique that is based on the construction of a statistical model to describe the relationship between an outcome (dependent or response variable) and one or more independent (predictor or explanatory) variables. The goal of the analysis using this method is that of model-building technique used in statistics to find the best fitting and most parsimonious, reasonable model to describe the relationship between the response variable and a set of independent variables. These independent variables are often called covariates.

An ordinary linear regression cannot be used for dichotomous dependent variable because it violates the assumptions of normality and homoscedasticity. It is impossible for a data to follow the normal distribution with only two values. If the dependent variable Y assumes the values 0 or 1, residuals/error will be small for the regression line near $Y = 0$ and $Y = 1$, but large at the middle. Hence the error term will violate the assumption of equal variances. When the dependent variable assumes the values 0 and 1, the regression model will allow the coefficients below 0 and above 1, and multiple linear regression does not handle non-linear relationships. All these objections lead the researcher to use the logistic regression, and not an ordinary linear regression analysis.

2.2.1 Logistic model with only one independent variable

Logistic regression model has been used by Pawn (1999) to model the effect of therapaedic horse ridding. Suppose that only one independent variable, *x* and a response (dependent) variable, *Y*, are observed. If *Y* is binary, it is defined as:

$$Y = \begin{cases} 1 \text{ if a success with probability } p \\ 0 \text{ if failure with probability } 1 \text{ - } p \end{cases}$$

and is proportional where $Y = r$ successes out of *n* independent trials.

Thus, $P(Y = 1) = p$, and $P(Y = 0) = 1 - p$. The wish is to model *p*, the probability of success. A transformation of *p* that will be vital to the study of logistic regression is the logit transformation. The logistic regression model that relates *x* to *Y* is given by

$$\log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x \qquad (2.1)$$

The logit transformation (2.1) makes this function linear in its parameter, β. The natural logarithm (log) will be used throughout this project. Equivalently, the model may be written in terms of the odds of a positive response (McCullagh and Nelder, 1983), giving

$$\frac{p}{1-p} = \exp(\beta_0 + \beta_1 x) \qquad (2.2)$$

Finally the probability of a success is given by

$$\mathrm{P}(Y = 1) = p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \qquad (2.3)$$

2.2.2 Odds and Odds Ratio

Suppose that the dependent variable *Y* is defined in this way:

$$Y = \begin{cases} 1 \text{ if category A occurs, success} \\ 0 \text{ if category B occurs, failure} \end{cases}$$

Then the probability of success is p, i.e. P(Y = 1) = p, and the probability of a failure is 1 – p, i.e. P(Y = 0) = 1- p. The odds of a success (Y = 1) is defined to be the ratio of the probability of a success to the probability of a failure. Thus if p is the true success probability, the odds of a success is given by

$$\text{Odds} = \frac{p}{1 - p}$$

$$= \frac{P(Y=1)}{P(Y=0)} \tag{2.4}$$

Let Y take on two values 0 and 1, and also let x take on two values 0 and 1.

Let the logistic regression model be

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{2.5}$$

		Independent variable X	
		$x = 1$	$x = 0$
	y =1:	$p(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$p(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
Outcome variable Y			
	y = 0:	$1 - p(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1 - p(0) = \frac{1}{1+e^{\beta_0}}$

where

$$\text{P}(1) = \text{P}(Y = 1|x = 1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \text{ and } \text{P}(0) = \text{P}(Y\ = 1|x = 0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

The odds of a success (Y = 1) with x = 1 is defined as:

$$\frac{P(Y=1|x=1)}{P(Y=0|x=1)} = \frac{P(1)}{1 - P(1)} \tag{2.6}$$

The odds of a success (Y = 1) with x = 0 is defined as

$$\frac{P(Y=1|x=0)}{P(Y=0|x=0)}=\frac{P(0)}{1-P(0)} \tag{2.7}$$

The estimated odds ratio, denoted by $\overset{\wedge}{OR}$ is defined as the ratio of the odds for x =1 to the odds for x = 0, and is given by the equation

$$\overset{\wedge}{OR} \;=\; \frac{\left(\dfrac{P(1)}{1-P(1)}\right)}{\left(\dfrac{P(0)}{1-P(0)}\right)} \;=\; \frac{\left(\dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\left(\dfrac{1}{1+e^{\beta_0}}\right)}{\left(\dfrac{e^{\beta_0}}{1+e^{\beta_0}}\right)\left(\dfrac{1}{1+e^{\beta_0+\beta_1}}\right)}$$

$$= \;\frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} \;=\; e^{\beta_1} \tag{2.8}$$

That is, if x is coded as 0 and 1 and also Y is coded as 0 and 1, then the odds ratio is e^{β_1}.

Therefore, the estimate of $\overset{\wedge}{OR}$ = e^{β_1} and log($\overset{\wedge}{OR}$) = β_1. A quantity called **a relative risk** is defined as

$$\frac{P(1)}{P(0)}=\exp(\beta_0+\beta_1) \tag{2.9}$$

The log of the odds is called the logit, and these are

$$g_1=\log\left(\frac{P(1)}{1-P(1)}\right) \qquad \text{and} \qquad g_2=\log\left(\frac{P(0)}{1-P(0)}\right) \tag{2.10}$$

The log of the odds ratio (without substituting for P(1) and P(0)), called the log-odds ratio is given by

$$\overset{\wedge}{OR}= \log\left(\frac{P(1)}{1-P(1)}\right)-\log\left(\frac{P(0)}{1-P(0)}\right)$$

$$= g_1-g_2 \text{ which is the logit difference} \tag{2.11}$$

For large sample sizes, the distribution of $\overset{\wedge}{OR}$ is normal. Hence, the inferences are usually based on the sampling distribution of $\log(\overset{\wedge}{OR}) = \hat{\beta}_1$, which tends to follow a normal distribution for much smaller sample sizes. A 100(1-α)% confidence interval for the estimate of the odds ratio is obtained by first calculating the confidence limits of a confidence interval for the coefficient β_1, with a chosen significance level α. The confidence limits are then exponentiated to give a corresponding interval for the odds ratio. The confidence interval is given by

$$\exp[\hat{\beta}_1 \pm z_{1-\alpha/2} \times \overset{\wedge}{SE}(\hat{\beta}_1)].$$

That is, the confidence interval of $\log(\overset{\wedge}{OR}) = \hat{\beta}_1$ ranges from $[\hat{\beta}_1 - z_{1-\alpha/2} \times \overset{\wedge}{SE}(\hat{\beta}_1)]$ to $[\hat{\beta}_1 + z_{1-\alpha/2} \times \overset{\wedge}{SE}(\hat{\beta}_1)]$, where $z_{1-\alpha/2}$ is the upper ($100_{\alpha/2}$)% point of the standard normal distribution, $\overset{\wedge}{SE}(\hat{\beta}_1)$ is the standard error of $\hat{\beta}_1$, the parameter estimate of the coefficient β_1. This is the correct interval when the independent variable has been coded as 0 or 1.

If *x* is not coded as 0 and 1, the odds ratios are defined as follows:
(Other coding may require the calculation of the value of the logit difference for the specific coding used, and then exponentiated).

Suppose that *x* is coded as *x* = *a* and *x* = *b*. The logit difference is:

$$\begin{aligned} \mathrm{g}(x = a) - \mathrm{g}(x = b) &= [\hat{\beta}_0 + \hat{\beta}_1(a)] - [\hat{\beta}_0 + \hat{\beta}_1(b)] \\ &= \hat{\beta}_1(a-b) \end{aligned} \tag{2.12}$$

and the estimated odds ratio is

$$\overset{\wedge}{OR}(a,b) = \exp[\hat{\beta}_1(a-b)]. \tag{2.13}$$

Expression (2.13) is equal to $\exp(\hat{\beta}_1)$ only when (*a* - *b*) =1.

The odds ratio defined as the odds for $x = a$ to the odds for $x = b$ is given

$$\hat{OR}(a,b) = \frac{\left(\dfrac{P(x=a)}{[1-P(x=a)]}\right)}{\left(\dfrac{P(x=b)}{[1-P(x=b)]}\right)} \tag{2.14}$$

and when $a = 1$ and $b = 2$ we let $\hat{\text{OR}} = \hat{\text{OR}}(1,2)$.

In general the confidence interval is given by $\exp[\hat{\beta}_1(a-b) \pm z_{1-\alpha/2}|a-b| \times \hat{SE}(\hat{\beta}_1)]$.

If the independent variable has more than two categories, then design variables may be used.

The odds ratio needs to be established if the independent variables x is continuous.

Let's look at the equation $g(x) = \beta_0 + \beta_1 x$. If x is continuous, then β_1 gives the change in the log odds for an increase of 1 unit in x. That is $\beta_1 = g(x+1) - g(x)$ for any value of x. The change of c units in x can be obtained from the logit difference

$$g(x+c) - g(x) = \beta_0 + \beta_1(x+c) - \beta_0 + \beta_1(x) - c\beta_1$$

Now the odds ratio, $\hat{OR}$ is obtained by exponentiating the logit difference. That is

$$\hat{OR}(x+c, x) = \exp(c\beta_1)$$

A 100(1-α)% confidence interval for estimate for the odds ratio is

$$\exp[c\hat{\beta}_1 \pm z_{1-\alpha/2} \times c\,\hat{SE}(\hat{\beta}_1)]].$$

2.2.3 The logistic model with more than one independent variable

Let $y_1, y_2, \cdots, y_n$ be binary random variables taking values 0 or 1. Consider a collection of k independent variables denoted by the vector $\mathbf{x}_i = [1\ x_{i1}\ x_{i2} \cdots x_{ik}]'$ for individuals $i = 1, 2, \cdots, n$. Suppose that $\boldsymbol{\beta} = [\beta_1 ... \beta_k]'$ represent a vector of the coefficients. Let the probability that the outcome is 1 be denoted by $P(y_i = 1) = p_i$ and the probability that the outcome is 0 be denoted by $P(y_i = 0) = 1 - p_i$. Then the logit of the multiple logistic regression model is given by the equation

$$\begin{aligned} g(x) &= \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \\ &= \sum_{j=0}^{k} \beta_j x_{ij} \quad \text{for } i = 1, 2, \ldots, n \text{ where } x_{01} = 1 \end{aligned} \tag{2.15}$$

The explanatory variables may be categorical or continuous. The assumption of the analysis of binary data is that the observations are from a binomial distribution. That is, the distribution of the y_i is binomial with parameters (n, p_i).

Since the $\log\left(\frac{p}{1-p}\right)$ is also called the logit of p, then set $l\text{ogit}(p) = \log\left(\frac{p}{1-p}\right)$.

Hence,

$$g(x) = \log\left(\frac{p_i}{1-p_i}\right) = \text{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \tag{2.16}$$

which is now linear in the β's.

Solving the logit equation for p_i gives

$$p_i = \frac{\exp(\beta_0 x_{0i} + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik})}{1 + \exp(\beta_0 x_{0i} + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik})} \tag{2.17}$$

Dividing both the numerator and denominator by the numerator itself, produces

$$\begin{aligned} p_i &= \frac{1}{1 + \exp(-\beta_0 x_{0i} - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_k x_{ik})} \\ &= \frac{1}{1 + e^{-\boldsymbol{\beta}\mathbf{x}_i}} \end{aligned} \tag{2.18}$$

The $x_i's$, for i = 1, 2, ..., k are the independent variables and the β's are the parameters to be estimated in the model. In general, the coefficients $\hat{\beta}_j's$ in the logistic model estimate the change in the log-odds when x_i is increased by 1 unit, holding all other x's in the model fixed. The antilog of the coefficient, $e^{\hat{\beta}_i}$, then estimates the odds ratio

$$\frac{\left(\dfrac{p_{x+1}}{1-p_{x+1}}\right)}{\left(\dfrac{p_x}{1-p_x}\right)} \tag{2.19}$$

where p_x is the value of $P(y=1)$ for a fixed value of x, and $e^{\hat{\beta}_i}-1$ is an estimate of the percentage increase (or decrease) in the odds $\dfrac{p}{1-p}=\dfrac{P(y=1)}{P(y=0)}$ for every 1 unit increase in x_i, holding the other x's fixed. Since $p=P(y=1)$, then $1-p=P(y=0)$.

The ratio

$$\left(\frac{p}{1-p}\right)=\frac{P(y=1)}{P(y=0)} \tag{2.20}$$

is known as the odds of the event y = 1 occurring.

2.2.4 Multinomial logit model

Logistic regression is most frequently employed to model the relationship between a binary outcome variable and a set of covariates, but with a few modifications it may also be used when the outcome variable is polytomous (i.e. with more than two categories). The extension of the model and methods for a binary outcome variable to a polytomous outcome variable is easily illustrated when the outcome variable has three categories. This project shall focus on the extension and methods for a binary outcome variable with only three categories.

Assume that the categories of the outcome variable, Y, are coded as 0, 1, or 2. Recall that the logistic regression model for a binary outcome variable was parameterized in terms of the logit of Y = 1 versus Y = 0. In the three category model we have two logit functions:

one for $Y = 1$ versus $Y = 0$, the other for $Y = 2$ versus $Y = 0$. In the theory we could use any of the two pairwise logit comparison of outcomes, but the obvious extension from the binary case is to use the logit of $Y = 2$ versus $Y = 0$ for the second function. Thus the group coded $Y = 0$ will serve as the reference outcome value. The logit for comparing $Y = 2$ to $Y = 1$ may be obtained as the difference between the logit of $Y = 2$ versus $Y = 0$ and the logit of $Y = 1$ versus $Y = 0$.

Let $\mathbf{x}$ be the vector of covariates of length $k+1$ with $x_0 = 1$ to account for the constant term. The two logit functions are denoted as:

$$g_1(x) = \ln\left[\frac{P(Y=1\,|\,\mathbf{x})}{P(Y=0\,|\,\mathbf{x})}\right] \tag{2.21}$$

$$= \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \ldots + \beta_{1k}x_k$$

$$= (1, \mathbf{x'})\boldsymbol{\beta}_1$$

and

$$g_2(x) = \ln\left[\frac{P(Y=2\,|\,\mathbf{x})}{P(Y=0\,|\,\mathbf{x})}\right] \tag{2.22}$$

$$= \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \ldots + \beta_{2k}x_k$$

$$= (1,\ \mathbf{x'})\boldsymbol{\beta}_2$$

The intercept parameter (β_{10} or β_{20}) is the logits for success when x_i is zero and the slope parameter β_i is the logit difference indicating how much the log-odds change with a unit change on the predictor (Reise, 2000). It follows that the three conditional probabilities of each outcome category given the vector of explanatory variables are:

$$P(Y=0\,|\,\mathbf{x}) = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

$$P(Y=1\,|\,\mathbf{x})=\frac{e^{g_1(x)}}{1+e^{g_1(x)}+e^{g_2(x)}}$$

$$P(Y=2\,|\,\mathbf{x})=\frac{e^{g_2(x)}}{1+e^{g_1(x)}+e^{g_2(x)}}$$

(2.23)

Let $p_j(\mathbf{x})=P(Y=j\,|\,\mathbf{x})$ for j = 0, 1, 2; each of which is a function of the vector of 2(k + 1) parameters $\boldsymbol{\beta}'=(\boldsymbol{\beta}_1',\boldsymbol{\beta}_2')$.

A general expression for the conditional probability in the three category model is

$$P(Y=j\,|\,\mathbf{x})=\frac{e^{g_j(x)}}{\sum_{k=0}^{2}e^{g_k(x)}} \tag{2.24}$$

where the vector $\boldsymbol{\beta}_0$ = 0 and hence $g_0(\mathbf{x})$ = 0.

Wrigley (1985) states that having seen the extension of the dichotomous/binary logit and logistic models to the three-response category case it simply remains to note that the same principles of extension hold when generalizing the models from three categories to any number of categories. For example, in the case of R response categories, where R stands for any integer number (in most empirical examples R will be small), the three-category logistic model with k explanatory variables generalizes to a system of R - 1 non-linear logistic equations of the form:

$$P(Y=0\,|\mathbf{x})=\frac{1}{1+\sum_{k=0}^{R-1}e^{g_k(\mathbf{x})}}$$

$$\begin{matrix} . & & . \\ . & & . \\ . & & . \end{matrix}$$

$$P(Y=j\,|\mathbf{x})=\frac{e^{g_j(\mathbf{x})}}{1+\sum_{k=0}^{R-1}e^{g_k(\mathbf{x})}} \tag{2.25}$$

which assumes the imposition of the usual arbitrary constraints relating to the base category as

$$P(Y=j\,|\mathbf{x}) = \frac{e^{g_j(\mathbf{x})}}{1+\sum_{k=0}^{R} e^{g_k(\mathbf{x})}} \quad \text{for } j = 0, 1, 2, \ldots, R \tag{2.26}$$

Multinomial logistic regression uses the "odds-like" expressions for two comparisons.

$$(1)\ \frac{P(Y=1)}{P(Y=0)} \quad \text{and} \quad (2)\ \frac{P(Y=2)}{P(Y=0)} \tag{2.27}$$

The first expression (1) in (2.27) is the probability that the outcome is in category 1 divided by the probability that the outcome is in category 0, and the second expression (2) is the probability that the outcome is in category 2 divided by the probability that the outcome is in category 0. Since there are three categories, the total sum of the three probabilities must be equal to 1. The probabilities in the ratio of the two comparisons do not sum to 1 and therefore the two "odds-like" expressions are not the true odds. Each expression is an odds if there is a condition on the outcome being in the two categories of interest (Kleinbaum and Klein, 2002).

The two odds ratios are given by

OR_1 = category 1 versus category 0

$$= \left[\frac{\left(\dfrac{P(Y=1|x=1)}{P(Y=0|x=1)}\right)}{\left(\dfrac{P(Y=1|x=0)}{P(Y=0|x=0)}\right)}\right]$$

$$= \frac{\exp[\beta_{10}+\beta_{11}(1)]}{\exp[\beta_{10}+\beta_{11}(0)]} = \exp[\beta_{11}] \tag{2.28}$$

OR_2 = category 2 versus category 0

$$= \left[\frac{\left(\dfrac{P(Y=2|x=1)}{P(Y=0|x=1)}\right)}{\left(\dfrac{P(Y=2|x=0)}{P(Y=0|x=0)}\right)}\right]$$

$$= \frac{\exp[\beta_{20} + \beta_{21}(1)]}{\exp[\beta_{20} + \beta_{21}(0)]} = \exp[\beta_{21}] \qquad (2.29)$$

The two odds ratios are true if the independent variable *Y* is coded as 0, 1, 2 and the explanatory variable *x* is coded as 0 and 1. In general, the odds ratios where *x* has categories *a* and *b* (and *Y* = 0 is the reference category) will be estimated by:

$$\overset{\wedge}{OR}_j\ (a, b) = \left[\frac{\left(\dfrac{P(Y = j|x = a)}{P(Y = 0|x = a)}\right)}{\left(\dfrac{P(Y = j|x = b)}{P(Y = 0|x = b)}\right)}\right] \qquad (2.30)$$

where *j* denotes the value of *Y* that is compared to the reference category.

To compare any two levels ($X_1 = X_1^{**}$ versus $X_1 = X_1^{*}$) of the independent variables, the odds ratio is given by

$$\mathrm{OR}_q = \exp[\beta_{q1}(X_1^{**} - X_1^{*}] \text{ where } q = 1, 2$$

This odds ratio equation includes both categorical and continuous explanatory variables.

The 95% confidence interval for the OR is given by

$$\exp[\hat{\beta}_{\mathrm{q1}}\ (X_1^{**} - X_1^{*}) \pm 1.96\ (X_1^{**} - X_1^{*})\ \hat{SE}(\hat{\beta}_{\mathrm{q1}})]$$

In the model for nominal responses, suppose that the response variable *Y* takes possible values 1, 2, … , *p* where the numbers are labels for the categories, and neither orderings nor difference between category numbers is meaningful. Nominal responses often occur in situations where an individual faces *p* choices. The categories then refer to several alternatives. For example, in the choice of transportation mode the alternatives may be bus, train or minibus.

In probabilistic choice theory it is often assumed that an unobserved utility U_r is associated with the rth response. For the choice of transportation mode the underlying variable may be interpreted as the consumers' utility connected to the transportation mode (Fahrmeir and Tutz, 1994). *Let U_r be given by*

$$U_r = u_r + \varepsilon_r \tag{2.31}$$

where u_r is a fixed value associated with the rth response category and $\varepsilon_1, \ldots, \varepsilon_p$ are independently and identically distributed with continuous distribution function.

The multinomial logit model is then given by

$$P(Y = r) = \frac{\exp(u_r)}{\sum_{s=1}^{p} \exp(u_s)} \tag{2.32}$$

Let's consider a situation where an individual faces p choices and a set of variables characterizes the individual. Let the i^{th} individual be characterized by the vector $\mathbf{x}_i = [x_{i1}\ x_{i2} \cdots x_{ik}]'$ containing variables such as sex, age and income. Consequently, u_{ir} will denote the utility of the r^{th} category for individual i, Y_i denotes the categorical response variable. A simple linear model for the utility u_{ir} is given by

$$u_{ir} = \beta_{r0} + x_i'\boldsymbol{\beta}_r \tag{2.33}$$

where $\boldsymbol{\beta}_r = (\beta_{r1}, \beta_{r2}, \ldots, \beta_{rk})$ is a parameter vector. This means that the preference of the r^{th} alternative by the i^{th} individual is determined by x_i and a parameter β_r that depends on the category.

2.3 Fitting the logistic regression model

2.3.1 Estimating the parameters of the multiple logistic model

In linear regression the method used most often for estimating unknown parameters is least squares. In the least squares method the values of the parameters that minimize the sum of squared deviations of the observed values of *Y* from the predicted values based upon the model are chosen. Under the usual assumptions for linear regression the method of least squares yields estimators with a number of desirable properties. Unfortunately, when the method of least squares is applied to a model with a dichotomous outcome the estimators no longer have these same properties.

The general method of estimation that leads to the least squares function under the linear regression model (when the error terms are normally distributed) is called *maximum likelihood*. In a very general sense the method of maximum likelihood yields values for the unknown parameters that maximize the probability of obtaining the observed set of data. In order to apply this method a function, called the *likelihood function* must first be constructed. This function expresses the probability of the observed data as a function of these unknown parameters. The maximum likelihood estimators of these parameters are chosen to be those values that maximize this function.

If Y is coded as 0 or 1 then the expression for the equation (2.17) provides (for the value of $\boldsymbol{\beta}' = (\beta_0, \beta_1, \cdots, \beta_k)$, the vector of parameters) the conditional probability that Y is equal to 1 given **x**, denoted as $P(Y = 1 | \mathbf{x})$. It follows that the quantity $1 - p$ gives the conditional probability that Y is equal to zero given x, $P(Y = 0 | x)$.

Suppose that there are n (statistically independent) individuals ($i = 1, 2, \ldots, n$) observed. For each individual i, the data consists of y_i and x_i, where y_i is a random variable with possible values of 0 and 1; $\mathbf{x}_i = [1 \;\; x_{i1} \;\; x_{i2} \;\; \ldots \;\; x_{ik}]'$ is a vector of explanatory variables (the 1 is for the intercept); and $\boldsymbol{\beta}' = (\beta_0, \beta_1, \cdots, \beta_k)$ is the vector of parameters to be estimated. Letting p_i be the probability that $y_i = 1$, the logit model will be given as:

$$p_i = \frac{1}{1 + e^{-\boldsymbol{\beta}\mathbf{x}_i}} \tag{2.34}$$

The likelihood of observing the values of y_i for all the observations can be written as

$$\boldsymbol{L} = P(y_1, y_2, \cdots, y_n) \tag{2.35}$$

Because of the assumption that the observations are independent, the overall probability of observing all the y_i's can be written as the product of the individual probabilities:

$$\boldsymbol{L} = P(y_1)P(y_2) \ldots P(y_n) = \prod_{i=1}^{n} P(y_i) \tag{2.36}$$

where Π indicates repeated multiplication.

By definition $P(y_i = 1) = p_{\text{i}}$ and $P(y_i = 0) = 1 - p_i$, so this can be written as:

$$P(y_i) = p_i^{y_i}(1-p_i)^{1-y_i} \tag{2.37}$$

Therefore

$$\begin{aligned} \boldsymbol{L} &= \prod_{i=1}^{n} \quad p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \prod_{i=1}^{n}\left(\frac{p_i}{1-p_i}\right)^{y_i}(1-p_i) \end{aligned} \tag{2.38}$$

Taking the logarithm on both sides of the equation result as:

$$\log L = \quad \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} \log(1-p_i) \tag{2.39}$$

In general it is easier to work with the logarithm of the likelihood function because the products are converted into sums and exponents become coefficients. Substituting the expression for the logit model (2.34) into equation (2.39) gives:

$$\text{Log}L = \sum_{i=1}^{n} \boldsymbol{\beta}\mathbf{x}_i y_i + \sum_{i=1}^{n} \log(1+e^{\boldsymbol{\beta}\mathbf{x}_i}) \tag{2.40}$$

Now the values of $\boldsymbol{\beta}$ that maximize equation (2.40) are to be estimated. The well-known approach used to find the values of $\boldsymbol{\beta}$ is to find the derivative of the function with respect to $\boldsymbol{\beta}$, set the derivative equal to 0, and then solve for $\boldsymbol{\beta}$. Taking the derivative of equation (2.40) and setting it equal to 0 results:

$$\begin{aligned} \frac{\partial \log L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \mathbf{x}_i y_i - \sum_{i=1}^{n} (1+e^{\boldsymbol{\beta}\mathbf{x}_i})^{-1} \\ &= \sum_{i=1}^{n} \mathbf{x}_i y_i - \sum_{i=1}^{n} \mathbf{x}_i \hat{\mu}_i = 0 \end{aligned} \tag{2.41}$$

where

$$\hat{\mu}_i = \frac{1}{1+e^{-\boldsymbol{\beta}\mathbf{x}_i}}$$

is the predicted probability of y for a given value of $\mathbf{x}_i$. Because $\mathbf{x}_i$ is a vector, equation (2.41) is actually a system of k + 1 equations, one for each element of $\boldsymbol{\beta}$. Since there is no explicit solution for equation (2.41), one must then rely on iterative methods, which give successive approximations to the solution until the approximations converge to the correct value. Although there are many different methods of doing this, and all give the same solution, but they differ in speed of convergence, sensitivity to starting values, and computational difficulty at each iterative. The most widely-used iterative method is the Newton-Raphson algorithm, which is described by Allison (1999) as follows:

Let $\mathbf{U}(\boldsymbol{\beta})$ *be the vector of first derivatives of logL with respect to* $\boldsymbol{\beta}$ *and let* $\mathbf{I}(\boldsymbol{\beta})$ *be the matrix of the second derivatives of logL with respect to* $\boldsymbol{\beta}$. *That is,*

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i y_i - \sum_{i=1}^{n} \mathbf{x}_i \hat{y}_i$$

$$\mathbf{I}(\boldsymbol{\beta}) = \frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \, \hat{y}_i (1-\hat{y}_i) \qquad (2.42)$$

The vector of the first derivatives $\mathbf{U}(\boldsymbol{\beta})$ *is sometimes called the gradient or score. The matrix of the second derivatives* $\mathbf{I}(\boldsymbol{\beta})$ *is called the Hessian matrix. The Newton-Raphson algorithm is then*

$$\boldsymbol{\beta}_{j+1} = \boldsymbol{\beta}_j - \mathbf{I}^{-1}(\boldsymbol{\beta}_j)\mathbf{U}(\boldsymbol{\beta}_j) \qquad (2.43)$$

where $\mathbf{I}^{-1}$ *is the inverse of* $\mathbf{I}$. *In practice a set of starting values* $\boldsymbol{\beta}_0$ *is needed. These starting values are substituted into the right-hand side of equation (2.43) which yields the results for the first iteration,* $\boldsymbol{\beta}_1$. *These values are then substituted back into the right-hand side, the first and the second derivatives are recomputed, and the result is* $\boldsymbol{\beta}_2$. *This process is repeated until the maximum change in each parameter estimate from one step to the next is less than some criteria. If the absolute value of the current parameter estimate* $\boldsymbol{\beta}_2$ *is less than or equal to 0.01, the default criterion for convergence is*

$$|\beta_{j+1} - \beta_j| < 0.0001$$

If the current parameter estimate is greater than 0.01 (in absolute value), the default criterion is

$$\left|\frac{\beta_{j+1} - \beta_j}{\beta_j}\right| < 0.0001$$

After the solution $\hat{\boldsymbol{\beta}}$ *is found, a byproduct of the Newton-Raphson algorithm is an estimate of the covariance matrix of the coefficients, which is simply* $-\boldsymbol{I}^{1}(\hat{\boldsymbol{\beta}})$*. Estimates of the standard errors of the coefficients are obtained by taking the square roots of the main diagonal elements if this matrix.*

2.3.2 Estimating the parameters of the multinomial logit model

The likelihood function is constructed by formulating three binary variables coded as 0 or 1 to indicating group membership of an observation. The variables are coded as follows: if $Y = 0$ then $Y_0 = 1$, $Y_1 = 0$ and $Y_2 = 0$; if $Y = 1$ then $Y_0 = 0$, $Y_1 = 1$ and $Y_2 = 0$ and lastly if $Y = 2$ then $Y_0 = 0$, $Y_1 = 0$ and $Y_2 = 1$. That is

$$Y_{j0} = \begin{cases} 1 & \text{if outcome} = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{j1} = \begin{cases} 1 & \text{if outcome} = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{j2} = \begin{cases} 1 & \text{if outcome} = 2 \\ 0 & \text{otherwise} \end{cases}$$

for $j = 1, 2, 3, \ldots, n$ subjects. Then $\sum Y_j = 1$, regardless of what value Y takes on.

The method of maximum likelihood yields values for the unknown parameters. In order to apply this method, a function called the log-likelihood function must first be constructed as follows

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_{1i} g_1(\mathbf{X}_i) + y_{2i} g_2(\mathbf{X}_i) - \ln(1 + e^{g_1(x_i)} + e^{g_2(x_i)}) \qquad (2.44)$$

The likelihood equations are found by taking the first partial derivatives of L($\boldsymbol{\beta}$) with respect to each of the $2(k+1)$ unknown parameters. For simplicity of the notation, let $p_{ji} = p_j(x_i)$. The general form of these equations is as follows:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{jk}} = \sum_{i=1}^{n} x_{ki}(y_{ji} - p_{ji}) \tag{2.45}$$

for $j = 1, 2$ and $k = 0, 1, 2, \ldots, p$. We recall $x_{0i} = 1$ for each subject.

The maximum likelihood estimator, $\hat{\boldsymbol{\beta}}$, of these parameters are chosen to be those values which maximize (2.45), and is obtained by setting the derivative of the log-likelihood equations to zero and solving for $\boldsymbol{\beta}$. The results of the polytomous logistic regression model and binary logistic regression models may be estimated and then compared, (Bender and Grouven, 1998).

2.4 Assessment of the logistic regression model

After fitting a multiple logistic regression model to a set of data, the model needs to be assessed. This frequently involves the formulation and testing of a statistical hypothesis to determine whether or not the model is significant, and whether or not the independent variables are significantly related to the response variable. That is, after estimating the parameters of the logistic model, it is vital to inquire about the extent to which the fitted values of the response variable under the model compare with the observed values. If the harmony between the observations and the corresponding fitted values is good, then the model may be regarded as adequate. The aspect of the adequacy of the model is usually referred to as goodness-of-fit, while an ill-fitting model is said to display lack of fit.

2.4.1 Statistical inferences

Confidence intervals and hypothesis testing are the two most common types of formal statistical inference. Both are appropriate when the aim is to estimate a population parameter. Hypothesis testing is an inference used to assess the evidence provided by the data in favour of some claim about the population. In order to formulate such a test, usually some theory has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved.

The hypothesis is a statement about the parameter(s) in a population or model. In hypothesis testing, the statement being tested is called the null hypothesis and is denoted by H_0. Hypothesis testing is designed to assess the strength of the evidence against the null hypothesis. The null hypothesis is usually a statement of “no effect” or “no difference.” The alternative hypothesis denoted by H_a is a statement that is suspected to be true instead of H_0.

The results of a test are expressed in terms of a probability that measures how well the data and the hypotheses concur. The significance level of a statistical hypothesis test is a fixed probability of wrongly rejecting the null hypothesis H_0, if it is in fact true.

A statistical test is based on the concept of proof and is composed of the five steps listed below:

- ♦ State the null hypothesis, denoted by H_0
- ♦ State the alternative hypothesis called research hypothesis, H_a, The test is designed to assess the strength of the evidence against H_0. H_a is the statement that will be accepted if the evidence enables the researcher to reject H_0. The alternative hypothesis, H_a, is a statement of what a statistical hypothesis test is set up to establish.

- ♦ The value of the test statistic on which the test will be based is calculated. Its value is used to decide whether or not the null hypothesis should be rejected in a hypothesis test. The choice of a test statistic will depend on the assumed probability model and the hypotheses under question.

- ♦ The rejection region is determined. The rejection region is a region on the sample space which leads the researcher to reject the null hypothesis H_0. So, if the observed value of the test statistic is a member of the rejection region or lies in the rejection region, the conclusion is 'reject H_0'. If it is not a member or does not lie on the rejection region then the conclusion is 'do not reject H_0. Alternatively, determine the probability value for the observed data (calculated assuming that H_0 is true) that the test statistic will weigh against H_0 at least as it does for this data.

- State a conclusion. The conclusion is a statement that summarizes what the researcher has found by using a hypothesis test. The usual way to do this is to choose a significance level α, how much evidence against H_0 one regard as decisive. If the *p*-value is less than or equal to α, one concludes that the data do not provide sufficient evidence to reject the null hypothesis.

2.4.2 Testing for the significance of the overall model

According to Sharma (1996), the null and the alternative hypotheses for assessing the overall model fit are given by

H_0: The hypothesized model fits the data
H_a: The hypothesized model does not fit the data.

H_0 is the null hypothesis and H_a is the alternative hypothesis of the test. Non-rejection of the null hypothesis is desired, as it leads to the conclusion that the model fits the data.

2.4.2.1 The deviance

Let L_c be the maximum log-likelihood function for the current model, and let L_f be the maximum log-likelihood function for the full model. Suppose that the following expression is a contribution to the likelihood function for the pair (x_i, y_i)

$$p(x_i)^{y_i}[(1-p(x_i)]^{1_i-y_i}$$

The likelihood function of the β parameters is given by:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} p^{y_i}(1-p_i)^{1_i-y_i} \qquad (2.46)$$

where $p_i = p(x_i)x_i$

hence, the log-likelihood of equation (2.46) is

$$L(\boldsymbol{\beta})\log[l(\beta)] = \sum_{i=1}^{n}\left[y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)\right] \qquad (2.47)$$

Since the deviance, *D* is defined as

$$D = -2\log\left[\frac{\text{(likelihood of the current model)}}{\text{(likelihood of the full model)}}\right], \tag{2.48}$$

equation (2.48) becomes

$$D = -2\sum_{i=1}^{n}\left[y_i \log(\left(\frac{\hat{p}_i}{y_i}\right)/) + (1-y_i)\log\left(\frac{1-\hat{p}_i}{1-y_i}\right)\right] \tag{2.49}$$

According to Collet (1991), the deviance is asymptotically distributed as chi-squared (χ^2) with *n* - *p* degrees of freedom, where *n* is the number of binomial observations, and *p* is the number of unknown parameters included in the current linear logistic model.

2.4.2.2 The Pearson chi-squared

The deviance has been considered exclusively as a summary measure of lack of fit. The other popular measure of goodness-of-fit is known as the Pearson's χ^2-statistic defined by

$$\chi^2 = \sum_{i=1}^{n}\frac{(\text{observed} - \text{fitted})^2}{\text{fitted}} \tag{2.50}$$

Both the deviance and χ^2-statistic defined by (2.50) have the same asymptotic χ^2-distribution. The chi-squared (χ^2) statistic can also be used to test for the association or correlation between variables.

2.4.2.3 The Hosmer-Lemeshow test

The Hosmer–Lemeshow test is based on the grouping of the values of the estimated probabilities. Allison (1999) describes the grouping and calculation of the Hosmer-Lemeshow statistic as follows:

Based on the estimated model, predicted probabilities are generated for all observations. These are sorted by size, and then grouped into approximately 10 intervals. Within each interval, the expected frequency is obtained by adding up the predicted probabilities. Expected frequencies are compared with observed frequencies by the conventional Pearson chi-square statistic.

Suppose that J = n, where n is the number of columns corresponding to the n values of the estimated probabilities. Two grouping strategies were proposed as follows: (1) collapse the table based on percentiles of the estimated probabilities, and (2) collapse the table based on fixed values of the estimated probabilities.

With the first method, using g = 10 groups results with the first group containing the $n_1^{'}$ *= n/10 subjects having the smallest estimated probabilities, and the last group containing* $n_{10}^{'}$ *= n/10 subjects having the largest probabilities. With the second method, use of g = 10 results in cutpoints defined at the values k/10, k = 1, 2, …,9 and the groups contain all the subjects with estimated probabilities between adjacent cutpoints. For example, the first group contains all subjects whose estimated probability is less than or equal to 0.1, while the tenth group contains those subjects whose estimated probability is greater than 0.9. For the y = 1 row, estimates of the expected values are obtained by summing the estimated probabilities over all subjects in a group. For the y = 0 row, the expected value is obtained by summing over all subjects in the group, one minus the estimated probability.*

For any grouping strategy, the Hosmer-Lemeshow goodness of fit statistic, $\hat{C}$*, is obtained by calculating the Pearson chi-squared statistic from the 2* $\times g$ *table of observed and estimated expected frequencies. A formula for defining the calculation of* $\hat{C}$ *is as follows:*

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k^{'}\overline{p}_k)^2}{n_k^{'}\overline{p}_k(1-\overline{p}_k)} \tag{2.51}$$

where $n_k^{'}$ *is the total number of subjects in the kth group,*

$$o_k = \sum_{j=1}^{c_k} y_j \tag{2.52}$$

is the number of responses among the c_k *covariate pattern, and*

$$\overline{p}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{p}_j}{n_k^{'}} \tag{2.53}$$

is the estimated probability, and c_k denotes the number of covariate patterns in the kth decile. The Hosmer-Lemeshow statistic has approximately a chi-square distribution with g – 2 (the number of intervals minus 2) degrees of freedom under the null hypothesis that the fitted model is correct.

When we model a binary outcome variable we have a single fitted value, but when the outcome variable has three categories we have two estimated logit probabilities. The proposed extensions of tests for goodness-of-fit and logistic regression diagnostics to the multinomial logit model, is to assess the fit and calculate logistic regression diagnostics using the individual logit regressions approach. For an outcome variable with three categories we would assess the fit of the two logit regression models and then integrate the results to make a statement about the fit of the multinomial logit model.

2.4.2.4 The Likelihood Ratio Test

The Likelihood Ratio Test can be used to test for the significance of the *k* coefficients of the explanatory variable in the logistic regression model. The null and the alternative hypotheses are given by:

H$_0$: All the *k* coefficients, β_i's = 0

H$_a$: At least one of the *k* coefficients, β_i's, is not zero

If *D* is set to be the deviance, then the statistic for this test is

$$G = D(\text{for the model with constant only}) - D(\text{for the model with all the variables})$$

$$= -2\log\left[\frac{\text{(likelihood with constant only)}}{\text{(likelihood with all the variables)}}\right] \tag{2.54}$$

which has a chi-squared distribution with *k* degrees of freedom, where *k* is the number of explanatory variables. The null hypothesis is rejected if the *p*-value is less than the significance level α, for which we conclude that at least one, and perhaps all the *k* coefficients, are different from zero.

2.4.3 Testing for the significance of the coefficients

The test for the significance of the independent variable (*x*) on the binary response variable, tests for the contribution of a particular independent variable to the dependent variable. For the binary logistic regression model, the null hypothesis $H_0 : \beta_j = 0$ states that the probability of success is independent of *x*. For large samples, the test statistic

$$z = \frac{\hat{\beta}_j}{ASE(\hat{\beta}_j)}, \qquad for\, j = 1, 2, \ldots, k \tag{2.55}$$

where ASE is the estimated asymptotic standard error and $\hat{\beta}_j$ is the estimated coefficient for the *j*th variable, has a standard normal distribution when $\beta_j = 0$. Equivalently, z^2 (which is approximately equal to the likelihood ratio statistic) is a Wald statistics having a large-sample chi-squared distribution with the number of degrees of freedom equal to 1. Although the Wald test works well for very large samples, the likelihood-ratio test is more powerful and reliable for sample sizes used in practice (Agresti, 1996). The null hypothesis is rejected if the value of the test statistic is more than the critical value from the chi-squared table or if the *p*-value, usually calculated by the computer software, is less than the standard significance level.

Hosmer and Lemeshow (1989) state that with any multi-degree of freedom variable, the likelihood ratio test (LRT) should be used to assess significance. The LRT is used to test for the significance of the model due to the addition of new terms. For purposes of assessing the significance of an independent variable the values of the deviance *D*, with and without the independent variable, are compared. The change in deviance due to inclusion of the independent variable in the model is obtained as follows:

$$G = D(\text{for the model without the variable}) - D(\text{for the model with the variable}) \tag{2.56}$$

The LRT statistic, G, can be expressed as

$$G = -2\log\left[\frac{\text{(likelihood without the variable)}}{\text{(likelihood with the variable)}}\right] \tag{2.57}$$

That is, the LRT is obtained by multiplying the log of the likelihood ratio by –2. The LRT can also be used to verify whether all variables, except the specific constants are zero (Ortuzar and Willumsen, 1999).

The Score test is one of the tests statistics similar to the LRT, but it is based on the distribution theory of the derivatives of the log likelihood. It is a multivariate test that requires matrix calculations.

Suppose that the model is now multinomial where the dependent variable has three categories and two explanatory variables. As with an ordinary logistic regression model, the LRT can be used to assess the significance of the explanatory variables in the model. Note that, rather than testing one β coefficient for an explanatory variable, now two coefficients are tested at a time. There is a coefficient of each comparison of the dependent variable (that is $Y = 1$ versus $Y = 0$, and $Y = 2$ versus $Y = 0$). This will always affect the number of parameters to be tested and the number of degrees of freedom. The number of parameters and also the degrees of freedom is two (2). When testing for the significance of the coefficient first fit a full model, and compare to the reduced model. The null hypothesis is that the β coefficients corresponding to the relevant variable are both set equal to zero (H_0: $\beta_{11} = \beta_{21} = 0$). The likelihood ratio statistics is given by

$$G = -2\log\left[\frac{\text{(likelihood of the reduced model)}}{\text{(likelihood of the full model)}}\right] \tag{2.58}$$

The Wald test of multinomial model also tests for the significance of the explanatory variable. The null hypothesis is that the β coefficients are equal to zero. The null hypothesis are H_0: $\beta_{11} = 0$ (for category 1 versus 0) and H_0: $\beta_{21} = 0$ (for category 2 versus 0). The Wald statistic is obtained by dividing the estimated coefficient by its standard error, and is defined as

$$z = \frac{\hat{\beta}_{g1}}{\overset{\wedge}{SE}(\hat{\beta}_{g1})} \tag{2.59}$$

where $\hat{\beta}_{g1}\ \overset{\wedge}{SE}(\hat{\beta}_{g1})$ is the estimate of the beta coefficient and $\hat{\beta}_{g1}\ \overset{\wedge}{SE}(\hat{\beta}_{g1})$ is the standard error of the estimate coefficient. As with an ordinary logistic regression model, the Wald statistic of the multinomial model is approximately normally distributed with mean zero and variance equal to one. If the *p*-value is less than the significance level, then the variable is statistically significance.

The coefficients for the multinomial logit model are obtained from the two separate logit models. The coefficients obtained from fitting separate logistic models will be close to those from the multinomial fit. Thus, the individualized logistic model fitting approach shall be used for variable selection.

2.4.4 Model-Building Strategy: Stepwise Procedure

If there are many explanatory variables, stepwise selection methods can be used to identify best subsets of variables (Dobson, 1990). The statistic used in the assessment depends on the assumptions of the model that the errors are assumed to follow the binomial distribution, and significance is assessed via the likelihood ratio chi-square test. The following statistical algorithm for the selection or deletion of the independent variables is based on Hosmer and Lemeshow, (1989, pp 106-111).

Step (0): Suppose we have available a total of p possible independent variables, all of which are judged to be of plausible "biologic" importance in studying the outcome variable. Step (0) begins with a fit of the "intercept only model" and an evaluation of its log-likelihood, L_0. This is followed by fitting each of the p possible univariate logistic regression models and comparing their respective log-likelihoods. Let the value of the log-likelihood for the model containing variable x_j at step zero be denoted by $L_j^{(0)}$. The subscript j refers to that variable which has been added to the model, and the superscript (0) refers to the step. This notation will be used throughout the discussion of stepwise logistic regression to keep track of both step number and variables in the model.

Let the value of the likelihood ratio test for the model containing x_j versus the intercept only model be denoted by $G_j^{(0)} = 2(L_j^{(0)} - L_0)$, and its p-value be denoted by $p_j^{(0)}$. Hence, this p-value is determined by the tail probability $Pr[\chi^2(v) > G_j^{(0)}] = p_j^{(0)}$, where $v = 1$ if it is continuous, and $v = k - 1$ if x_j is polytomous with k categories.

The most important variable is the one with the smallest p-value. If we denote this variable by x_{e_1}*, then* $p_{e_1}^{(0)} = \min(p_j^{(0)})$*, where "min" stands for selecting the minimum of the quantities enclosed in the brackets. The subscript "*e_1*" is used to denote that the variable is a candidate for entry at step 1. For example, if variable* x_2 *had the smallest p-value, then* $p_{2_1}^{(0)} = \min(p_j^{(0)})$*, and* e_1 *= 2. Just because* x_{e_1} *is the most important variable, there is no guarantee that it will be "statistically significant." For example, if* $p_{e_1}^{(0)}$ *= 0.83, we would probably conclude that there is little point in continuing this analysis because the "most important" variable is not related to the outcome. On the other hand, if* $p_{e_1}^{(0)}$ *= 0.003, we would like to look at the logistic regression containing this variable and then see if there are other variables which are important given that* x_{e_1} *is in the model.*

A crucial aspect of using stepwise logistic regression is the choice of an "alpha" level to judge the importance of the variables. Let p_E *denote our choice where the "E" stands for entry. The choice for* p_E *will determine how many variables will eventually be included in the model. Choosing a value for* p_E *in the range 0.15 to 0.20 is more highly recommended. Moreover, use of* p_E *in this range will provide some assurance that the stepwise procedure will select variables whose coefficients are different from zero. Sometimes the goal of the analysis may be broader, and models containing more variables are sought to provide a more complete picture of possible models. In these cases use of* p_E *= 0.25 might be a reasonable choice. Whatever the choice for* p_E*, a variable will be judged important enough to include in the model if the p-value for G is less than* p_E*. Thus, the program proceeds to Step 1 if* $p_{e_1}^{(0)} < p_E$*; otherwise, it stops.*

Step 1: *Step 1 commences with a fit of the logistic regression model containing* x_{e_1}*. Let* $L_{e_1}^{(1)}$ *denote the log-likelihood of this model. To determine whether any of the remaining p - 1 variables are important once the variable* x_{e_1} *is in the model, we fit the p - 1 logistic models containing* x_{e_1} *and* x_j*, j = 1, 2, 3, …, p and j ≠* e_1*. For the model containing* x_{e_1} *and* x_j *let the log-likelihood be denoted by* $L_{e_1 j}^{(1)}$*, and let the likelihood* *ratio chi-square statistic of this model versus the model containing only* x_{e_1} *be denoted by* $G_j^{(1)} = 2(L_{e_1 j}^{(1)} - L_{e_1}^{(1)})$*. The p-value of this statistic will be denoted by* $p_j^{(1)}$*. Let the variable with the smallest p-value at Step 1 be* x_{e_2} *where* $p_{e_2}^{(1)} = \min(p_j^{(1)})$*. If this value is less than* p_E *we proceed to Step 2; otherwise we stop.*

Step 2: *Step 2 begins with s fit of the model containing* x_{e_1} *and* x_{e_2}*. It is important that once* x_{e_2} *has been added to the model,* x_{e_1} *is no longer important. Thus Step 2 includes a check for backward elimination. In general this is accomplished by fitting models that delete one of the variables added in the previous steps and assessing the continued importance of the variable removed. At Step 2 let* $L_{-e_j}^{(2)}$ *denote the log-likelihood of the model with* x_{e_j} *removed. In similar fashion let the likelihood ratio test of this model versus the full model at Step 2 be* $G_{-e_j}^{(2)} = 2(L_{e_1 e_2}^{(2)} - L_{-e_j}^{(2)})$ *and* $p_{-e_j}^{(2)}$ *be its p-value.*

To ascertain whether a variable should be deleted from the model the program selects that variable which when removed yields the maximum p-value. Denoting this variable as x_{r_2}*, then* $p_{r_2}^{(2)} = \max(p_{-e_1}^{(2)}, p_{-e_2}^{(2)})$*. To decide whether* x_{r_2} *should be removed, the program compares* $p_{r_2}^{(2)}$ *to a pre-chosen "alpha" level,* p_R*, which will indicate some minimal level of continued contribution to the model, where "R" stands for remove. Whatever value we choose for* p_R*, it must exceed the value of* p_E *to guard against the possibility of having the program enter and remove the same variable at successive steps. If we do not wish to exclude many variables once they have entered, we might use* p_R *=0.9. A more stringent value would be used if a continued "significant" contribution were required. For example, if we used* p_E *= 0.15, then we might choose* p_R *= 0.20. If the maximum p-value to remove, i.e.* $p_{r_2}^{(2)}$*, is less than* p_R *then* x_{r_2} *remains in the model. In either case the program proceeds to the variable selection phase.*

At the forward selection phase each of the p-2 logistic regression models containing x_{e_1}*,* x_{e_2} *and* x_j *are fit, for j = 1, 2, 3,..., p, j* $\neq e_1, e_2$*. The program evaluates the log-likelihood for each model, computes the likelihood ratio test versus the model containing* *only x_{e_1} and x_{e_2}, and determines the corresponding p-value. Let x_{e_3} denote the variable with the minimum p-value, that is, $p_{e_3}^{(2)} = \min(p_j^{(2)}$. If this p-value is smaller than p_R, $p_{e_3}^{(2)} <$ p_R, then the program proceeds to Step 3; otherwise it stops.*

Step 3: *The procedure for Step 3 is identical to that of Step 2. The program performs a check for backward elimination followed by forward selection. This process continues in this manner until the last step (say, Step S).*

Step S: *This step occurs when: (a) all p variables have entered the model, or (b) all the variables not included in the model have p-values to remove which are less than p_R, and the variables not included in the model have p-values to enter which exceed p_E. The model at this step contains those variables that are important relative to the criteria of p_E and p_R. These may or may not be the variables reported in the final model. For instance, if the chosen values of p_E and p_R correspond to our belief for statistical significance, then the model at Step S may well contain the significant variables.*

There are two methods that may be used to select variables from a summary table. These are comparable to methods commonly used in stepwise linear regression. The first method is based on the p-value for entry at each step, while the second is based on a LRT of the model at the current step versus the model at the last step. Let "q" denote an arbitrary step in the procedure. In the method we compare $p_{e_q}^{(q-1)}$ to pre-chosen significance level such as α = 0.15. If the value $p_{e_q}^{(q-1)}$ is less than α, then we move to step q. We stop at the step when $p_{e_q}^{(q-1)}$ exceeds α. We consider the model at the previous step for further analysis. In this method the criteria for entry is based on a test of the significance of the coefficient for x_{e_q} conditional on $x_{e_1}, x_{e_2}, \ldots, x_{e_{q-1}}$ being in the model. The degrees of freedom for the test are 1 and k-1, depending on whether x_{e_q} is continuous or polytomous with k categories.

In the second method we compare the model at the current Step q, not to the model at the previous step, Step q - 1, but to the model at the last step, Step S. We evaluate the p-value *for the LRT of these two models and proceed in this fashion until this p-value exceeds α. This tests that the coefficients for the variables added to the model from Step q to Step S are all equal to zero. At any given step there will be more degrees of freedom than the test employed in the first method. For this reason the second method may possibly select a larger number of variables than in the first method. It is well known that the p-values calculated in stepwise selection procedures are not p-values in the traditional hypothesis testing context. Instead, they should be thought of as indicators of relative importance among variables. The variables so identified should then be subjected to the more intensive analysis.*

2.4.5 Logistic Regression Diagnostics

Logistic regression diagnostics is the stage where other measures will be examined before accepting that the model is adequate. Once a model has been fitted to the observed values of a binary or binomial response variable it is essential to check that the fitted model is actually valid, (Collet 1991). Some usual ways that a fitted model may be inadequate are:

- ✓ The model may not include explanatory variables that really should be in the model.
- ✓ The data may contain influential or outlying observations, which may have an impact on the conclusion to be drawn from the analysis.
- ✓ The assumption that the observed response data come from a particular probability distribution may not be valid.

The techniques used to examine the adequacy of a fitted model are collectively known as *diagnostics.* The techniques may be based on formal statistical tests, tables of values of certain statistics or a graphical representation of these values. There are some statistics that can provide much information about the adequacy of the fitted model.

The following notations are used in this section:

j = index for the observations

w_j = weight value of the *j*th observation; w_j is set equal to 1 if the weight statement is not used

r_j = the number of event responses out of n_j trials and n_j is the value of trials. For the actual model syntax n_j = 1 and r_j is 1 if the ordered response is 1, and 0 if the ordered response is 2.

p_j = the probability that the *j*th observation has an event response. This is given by $p_j = F(\alpha + \beta' \mathbf{x}_j)$

b = the MLE obtained by the I RLS algorithm.

$\hat{\mathbf{V}}_{\mathbf{b}}$ = the estimated covariance matrix of **b**

$\mathbf{b}_j$ = the MLE when the *j*th observation is excluded.

The matrix **H** is called the hat matrix, and is defined as

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$$

where **W** is the $n \times n$ diagonal matrix of weights used in fitting the model, with weights $w_i = n_i \hat{p}(\mathbf{x}_i)[1 - \hat{p}(\mathbf{x}_i)]$; **X** is the $n \times k$ design matrix; and *k* is the number of unknown parameters in the model.

The diagonal elements of the hat matrix are useful in detecting extreme points in the design space where they tend to have larger values. The *j*th diagonal element is

$$h_{jj} = w_j n_j \hat{p}_j (1 - \hat{p}_j)(1, \mathbf{x}'_j)\hat{\mathbf{V}}_{\mathbf{b}}(1, \mathbf{x}'_j)$$

where $\hat{p}_j$ is the estimate of p_j, and h_{ji} is the *j*th diagonal element of the $n \times n$ matrix.

2.4.5.1 Pearson residual and deviance residual

Both Pearson and deviance residuals are useful in identifying observations that are not well explained by the model. Pearson residuals are components of the Pearson chi-squared statistic, and the Pearson chi-squared statistic is the sum of squares of the Pearson residuals. The Pearson residual for the *j*th observation can be written as

$$\chi_j^2 = \frac{\mathrm{y_i} - n_i \hat{p}_i \sqrt{w_j}}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}} \tag{2.60}$$

where $\hat{p}_i$ is the fitted predicted probability of success for the model fit, y_i is the number of "successes" for n_i trials at the ith setting of the explanatory variable, and $n\,\hat{p}_i$ is the fitted number of successes. Collet (1991) states that a better procedure is to divide the raw residuals by their standard error, s.e.$(y_i - \hat{y}_i)$. He also emphasizes that this standard error is quite complicated to derive, but it is found to be

$$\text{s.e.}(y_i - \hat{y}_i) \;=\; \sqrt{\{\hat{\upsilon}_i(1-h_i)\}} \tag{2.61}$$

where $\hat{\upsilon}_i \;=\; n_i\hat{p}_i(1-\hat{p}_i)$, and h_i is the ith diagonal element of the $n \times n$ hat matrix. After dividing the Pearson residual by $\sqrt{1-h_i}$, (2.61) becomes

$$\chi_j^2 = \frac{\mathrm{y_i} - n_i\hat{p}_i\sqrt{w_j}}{\sqrt{[\hat{\upsilon}_i(1-h_i)]}} \tag{2.62}$$

which is known as standardized Pearson residuals.

Deviance residuals are components of the deviance statistic. The deviance residual for the j th observation can be written as

$$d_j = \begin{cases} -\sqrt{-2w_j n_j \log(1-\hat{p}_j)} & \mathit{i}\text{f} \quad r_j = 0 \\ \pm\sqrt{2\mathrm{w}_j(r_j\log(r_j/(n_j\hat{p}_j)) + (n_j - r_j)\log((n_j - r_j)/(n_j(1-\hat{p}_j))))} & \text{if} \quad 0 < r_j < n_j \\ \sqrt{-2w_j n_j \log(p_j)} & \text{if} \quad r_j = n_j \end{cases}$$

(2.63)

2.4.5.2 Identifying outlying observations

Outliers are defined to be those values that are far distant from other observations. Such values may occur due to recording or measurements errors. Outliers can be identified as observations that have relatively large standardized deviance or standardized Pearson residuals. They can be detected from the plot of the residuals against the corresponding observation number, or index known as the *index plot.*

2.4.5.3 Identifying influential observations

A number of observations may have much influence in fitting the model and the fit could be quite different if they were deleted. If an observation takes on an extreme value on one or more of the explanatory variables, then it is more likely to have a large influence. It may be useful if the fit of the model is reported after the deletion of the influential observation(s).

SAS produces various measures of influence such as:

- ✓ *Dfbeta* which is the change in the parameter estimate when the observation is deleted, divided by its standard error.
- ✓ The change in chi-square ($\Delta\chi_j^2$) or deviance (ΔD_j) goodness-of-fit statistics when the observation is deleted.

The larger the value, the greater the observation's influence.

2.4.6 Statistical Inference for $\mu_1 - \mu_2$:

The Central Limit Theorem (Ott,1993) implies that if independent samples of sizes n_1 and n_2 are selected from two populations 1 and 2, then, where n_1 and n_2 are large, the sampling distributions of $\bar{x}_1$ and $\bar{x}_2$ will be approximately normal, with means and variances (μ_1, σ_1^2 / n_1) and (μ_2, σ_2^2 / n_2), respectively. Since $\bar{x}_1$ and $\bar{x}_2$ are independent, normally distributed random variables, then the sampling distribution for the difference in the sample means, $\bar{x}_1 - \bar{x}_2$, will be approximately normal with mean $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$ and a variance $\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$.

Standard error is given by $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$.

There is an assumption to be made when one is making inferences about μ_1 - μ_2 based on independent samples. The assumption is that the sampling is done from two normal populations (1 and 2) with different means μ_1 and μ_2 but equal variances σ^2. Two independent random samples of size n_1 and n_2 are drawn with the sample means $\bar{x}_1$ and $\bar{x}_2$, and the corresponding sample variances are s_1^2 and s_2^2, respectively. A comparison between the population means μ_1 and μ_2 will be done using the data from two samples. The estimation and a hypothesis testing concerning the difference $\mu_1 - \mu_2$ will be done. An estimate for the difference in population means is the sample difference $\bar{x}_1 - \bar{x}_2$.

A general confidence interval for $\mu_1 - \mu_2$ is given by $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

where $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

which is the pooled sample standard deviation and degrees of freedom, $df = n_1 + n_2 - 2$. The quantity s_p in the confidence interval is an estimate of the standard deviation for the two populations and is formed by combining (pooling) the two samples. The pooled variance, s_p^2 is also defined as the weighted average of the sample variances s_1^2 and s_2^2. If the sample sizes are the same ($n_1 = n_2$), s_p^2 becomes the mean of the two sample variances [i.e. $s_p^2 = (s_1^2 + s_2^2)/2$]. The *df* for the confidence interval are a combination of the degrees of freedom for the samples; that is,
$df = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2.$

A hypothesis about the difference between two population means can also be done. As with any test procedure, begin by specifying a research hypothesis for the difference in populations' means. One might, for example, specify that the difference $\mu_1 - \mu_2$ is greater than or less than some value d_0. (Note: d_0 will often be 0).

For any test of hypothesis the researcher need to follow the five step procedure of hypothesis testing described in Section 2.4.1: the null hypothesis (H_0), the alternative hypothesis (H_a), test statistic and rejection region need to be described.

For this test, the fives steps will be given in the following way:

The null hypothesis

H_0: μ_1 - μ_2= d_0 (d_0 is specified)

The null hypothesis is tested against one of the following alternative hypothesis:

1. H_a: μ_1 - μ_2> d_0
2. H_a: μ_1 - μ_2< d_0
3. H_a: μ_1 - μ_2≠ d_0

For this project, the alternative hypothesis to be used is the third one, i.e.

H_a: μ_1 - μ_2≠ d_0

The appropriate test statistic: $t = \dfrac{\bar{x}_1 - \bar{x}_2 - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$ and $df = n_1 + n_2 - 2$,

The null hypothesis is rejected if $|t| > t_{\alpha/2}$.

Several assumptions are made in the test of hypothesis for comparing two population means.

- ✓ The first assumption is that the two samples are independent. This means that the two samples are unrelated and drawn from two different populations. If this assumption is not valid, then the *t* methods will not be appropriate.

- ✓ The second assumption is that the populations from which the samples were drawn are normal.

- ✓ The third and final assumption is that the two population variances, σ_1^2 and σ_2^2, are equal. If the sample variances (s_1^2 and s_2^2) suggest that $\sigma_1^2 \neq \sigma_2^2$, there is an approximate *t*- test using the test statistic

$$t' = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{2.64}$$

where t' has a t distribution with $df = \dfrac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)c^2 + (1 - c)^2 (n_1 - 1)'}$ and $c = \dfrac{s_1^2 / n_1}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

If the computed value of the *df* is not an integer, then it must be rounded to the nearest integer.

Most researchers use two-tailed significance tests to ascertain the statistical significance of the difference between the samples, as Bankole et al. (1999) apply this test for the comparison of years.

2.4.7 Inferences about population variances

People usually think of inferences concerning population means. The variability of a population is sometimes more important than its mean. Test hypotheses and estimation about a single population variance or comparison of two population variances can also be done.

In a hypothesis test about the population means of two samples it is assumed that the variances of the two populations are equal, i.e. $\sigma_1^2 = \sigma_2^2$. In practice they are not necessarily equal and one should first test the null hypothesis that the variances of the two populations are equal. Firstly, assume that the two populations are normally distributed and label these populations as 1 and 2, respectively. The interest is to test whether or not the variances of population 1, σ_1^2 and that of population 2, σ_2^2 are equal. When independent random samples have been drawn from the respective populations, the ratio s_1^2 / s_2^2 possesses an F distribution. The F-distribution has the following properties:

- F values are always nonnegative.
- The F-distribution is non-symmetrical.
- Critical values are found from the F-distribution table.

A statistical test of the null hypothesis $\sigma_1^2 = \sigma_2^2$ uses the test statistics s_1^2 / s_2^2 or s_2^2 / s_1^2. The *F*-distribution only gives the upper-tail values. The upper-tail *F*-values for a two-tailed test can also be obtained from the *F*-distribution table.

A statistical test for testing the equality of σ_1^2 and σ_2^2 has the procedure as follows:

- ✓ The null hypothesis is given by: H_0: $\sigma_1^2 = \sigma_2^2$ and the alternative hypothesis can either be H_a: $\sigma_1^2 > \sigma_2^2$ or H_a: $\sigma_1^2 \neq \sigma_2^2$.
- ✓ The appropriate alternative hypothesis for this project is H_a: $\sigma_1^2 \neq \sigma_2^2$.
- ✓ According to Underhill and Bradfield (1994), the test statistic is

  $F = \dfrac{s_1^2}{s_2^2}$ for $s_1 > s_2$ and $F = \dfrac{s_2^2}{s_1^2}$ for $s_1 > s_2$

  where $\dfrac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1}$ and $\dfrac{s_2^2}{s_1^2} \sim F_{n_2-1,n_1-1}$
- ✓ Rejection region: The null hypothesis is rejected if the value of the test statistic exceeds tabulated value of F_{n_1-1,n_2-1} or F_{n_2-1,n_1-1} for $\alpha/2$ depending on the test statistic used.

2.4.8 Test of association in a two-way contingency table

There might be no significant correlation or association between the variables as tested by Banerjee et al. (2000). The chi-squared, (χ^2) statistic can also be used to test for the association between the two variables at a time.

In this section all outcomes are recorded in a two-way table.

For example,

		Total
O_{11}	O_{12}	$n_{1.}$
O_{21}	O_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	N

The entries of the table are the observed counts or frequencies. The rows and columns of a two-way table represent values of two categorical variables. Each combination of values for these two variables defines a cell where $n_{1.}$ and $n_{2.}$ are the row totals, $n_{.1}$ and $n_{.2}$ are the column totals, and N is the grand total. A two-way table with r rows and c columns contains $r \times c$ cells. The table with four rows and two columns is a 4×2 table with 8 cells. The interest is to test whether or not a relationship exists between the row and column variables.

The null and alternative hypotheses for the test of association in contingency tables are:

- ✓ The null hypothesis H_0 is: there is no association between the row variable and the column variable.
- ✓ The alternative hypothesis H_a is: there is no association between the row and the column variables. The alternative hypothesis H_a cannot be described as either one-sided or two-sided, because it includes all kinds of association that are possible.

When testing the null hypothesis in $r \times c$ tables, one has to compare the observed cell counts with expected cell counts calculated under the assumption that the null hypothesis is true. The product of the row and column totals divided by the grand total gives the expected cell count:

$$\text{Expected cell count} = \frac{\text{row total} \times \text{column total}}{N} \qquad (2.65)$$

where N is the total number of observations in the sample. A statistic that compares the entire set of observed counts with the set of expected counts is used. First, take the difference between each observed count and its corresponding expected count, and square it so that it is positive. Divide each squared difference by the expected count, a kind of standardization. Finally, sum all cells. The result is called the chi-squared statistic, χ^2 and is written as:

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \qquad (2.66)$$

This statistic is a measure of how much the observed cell counts in a two-way table depart from the expected cell counts. A larger value of χ^2 provides evidence against the null hypothesis. The sampling distribution of χ^2 is needed to obtain the *p*-value for the test, under the assumption that H_0 is true. The resulting distribution is the chi-squared distribution with the number of degrees of freedom given by $df = (r - 1)(c\text{-}1)$, where c is the number of columns and r is the number of rows. The chi-squared test always uses the upper tail of the χ^2distribution, because the distribution is non-symmetrical and its values are non-negative.

CHAPTER 3

DEFINITION OF VARIABLES

This chapter gives a brief description of the data format, data sets and how they differ. It presents how the data was collected and all the SAS (Version 8) codes used in the data analysis. It also gives all different codings of the independent variables used in the analysis of the data. The dependent and independent variables are also described.

The multinomial logit model is estimated in two ways; Firstly, by running PROC CATMOD procedure in SAS Version 8, and secondly by running PROC LOGISTIC for the three models separately.

Stated preference data has been collected through questionnaire interviews from persons (males and females) being residents of formal and informal settlements within Mamelodi and its extensions, who commute to work in the CBD of Pretoria and are perceived to be able to choose a public transport mode from at least two alternatives. Stated preference is a statement by an individual of his or her liking for one alternative over another. Respondents were asked to make a choice from each of 16 choice sets (with 8 pictorial and 8 verbal presentations). As Deshazo and Fermo, 2002 analyzed the stated choices from pre-specified choice sets, eight questions were presented in pictorial and another eight in verbal presentation method with choice set that contains alternatives that vary along several attributes.

3.1 Variables

The dependent variable is CHOICE='Mode choice of transport' and the independent variables are given below with their different codings.

Table 3.1: Explanatory variables used in the analysis together with their different odings.

Variable/Coding	Dichotomous Coding	Binary Coding	Effect Coding
EDUC='Education level'	1='less literate' 2='literate'	0='less literate' 1='literate'	-1='less literate' 1='literate'
Pres='Presentation method'	1 = verbal and 2 = pictorial	0 = verbal and 1 = pictorial	-1 = verbal and 1 = pictorial
TF='Train feeder: changing from minibus to train'	1 = no feeder and 2 = with feeder	0 = no feeder and 1 = with feeder	-1 = no feeder and 1 = with feeder
TC='Train cost' (in Rands)			
TSEC='Train security level'	1= not improved, 2 = improved	0= not improved, 1 = improved	-1= not improved, 1 = improved
TTT='Train traveling time' (in minutes)			
TSEA='Train seating'	1 = often and 2 = seldom	0 = often and 1 = seldom	-1 = often and 1 = seldom
BF='Bus feeder: changing from minibus to bus’	1 = no feeder and 2 = with feeder	0 = no feeder and 1 = with feeder	-1 = no feeder and 1 = with feeder
BC='Bus cost' (in Rands)			
BSEC='Bus security level'	1= not improved, 2 = improved	0= not improved, 1 = improved	-1= not improved, 1 = improved
BTT='Bus traveling time' (in minutes)			
BSEA='Bus seating'	1 = often and 2 = seldom	0 = often and 1 = seldom	-1 = often and 1 = seldom
BSEA='Bus seating'			
MF='Minibus feeder: changing from minibus to another minibus'	1 = no feeder and 2 = with feeder	0 = no feeder and 1 = with feeder	-1 = no feeder and 1 = with feeder
MC='Minibus cost' (in Rands)			
MSEC='Minibus security level'	1= not improved, 2 = improved	0= not improved, 1 = improved	-1= not improved, 1 = improved
MTT='Minibus traveling time' (in minutes)			
MSEA='Minibus seating'	1 = often and 2 = seldom	0 = often and 1 = seldom	-1 = often and 1 = seldom

3.2 SAS program

SAS was be used for the analysis. LOGISTIC and CATMOD procedures are mainly used in the analyses. A brief description of CATMOD and LOGISTIC procedures in SAS are given below.

CATMOD procedure is the one appropriate in multinomial logit model, that is the models in which the response variable has three categories. The maximum likelihood method is the default estimation method. Although a single data set is presented, different data sets were used for the PROC CATMOD procedure. The SAS Version 8 program used to analyze the data is given below, but since the data set is too large only eight observations of the data set is shown.

```
Title1'* Determine the influence of literacy (using different dummy
coding for the variables)
  and testing alternative measures of goodness-of-fit on the
applicability of the multinomial
logit model to model choice of transport*';

Title2'*Using binary coding for the variables*';

Proc Format;
Value CHOICEfmt 1='Train' 2='Bus' 3= 'Minibus';
Value EDUC 0='less literate' 1='literate';
Value PRESfmt 0='verbal' 1='pictorial';
Value TFfmt 0='no feeder' 1='feeder';
Value TSECfmt 0='not improved' 1='improved';
Value TSEAfmt 0='often' 1='seldom';
Value BFfmt 0='no feeder' 1='feeder';
Value BSECfmt 0='not improved' 1='improved';
Value BSEAfmt 0='often' 1='seldom';
Value MFfmt 0='no feeder' 1='feeder';
Value MSECfmt 0='not improved' 1='improved';
Value MSEAfmt 0='often' 1='seldom';

Data transpot;
Input RESP EDUC PRES DESCRIP QNO TF TC TSEC TTT TSEA BF BC BSEC BTT
BSEA MF MC MSEC MTT MSEA CHOICE;
```

```
Label EDUC='Highest standard passed'
      PRES='Presentation method'
        TF='Train feeder: changing from minibus to train' TC='Train
cost'
      TSEC='Train ecurity level' TTT='Train travelling time'
TSEA='Train seating'
      BF='Bus feeder: changing from minibus to bus' BC='Bus cost'
      BSEC='Bus security level' BTT='Bus travelling time' BSEA='Bus
seating'
      BF='Bus feeder: changing from minibus to bus' BC='Bus cost'
      BSEC='Bus security level' BTT='Bus travelling time' BSEA='Bus
seating'
      MF='Minibus feeder: changing from minibus to another minibus'
MC='Minibus cost'
      MSEC='Minibus security level' MTT='Minibus travelling time'
MSEA='Minibus seating';

Format TF TFfmt. TSEC TSEcfmt. TSEA TSEAfmt. BF BFfmt.
BSEC BSECfmt. BSEA BSEAfmt. MF MFfmt. MSEC MSECfmt. MSEA MSEAfmt. PRES
PRESfmt.
;

cards;

2 0 0 254 8 0 1.10 0 65 1 0 2.80 0 55 1 M M M M M 1

2 0 0 254 2 1 4.60 0 65 0 0 2.80 1 55 0 M M M M M 2

2 0 0 254 3 0 2.10 0 65 0 1 5.30 1 105 0 M M M M M 1

2 0 0 254 5 0 2.10 1 65 0 0 5.30 1 105 1 M M M M M 1

11 0 0 254 2 1 4.60 0 65 0 0 2.80 1 55 0 M M M M M 1

11 0 0 254 8 0 1.10 0 65 1 0 2.80 0 55 0 M M M M M 1

11 0 0 254 3 0 2.10 0 65 0 1 5.30 1 105 0 M M M M M 1

11 0 0 254 5 0 2.10 1 65 0 0 5.30 1 105 1 M M M M M 1

12 0 1 352 3 1 3.60 1 65 0 0 2.80 1 55 0 1 9.70 0 40 1 2

12 0 1 352 2 1 4.60 0 65 1 1 7.80 0 55 1 1 6.30 1 40 1 3


;
```

```
Proc CATMOD data=transpot;
DIRECT EDUC TC BC MC  BTT MTT;
model choice = EDUC PRES TF TC TSEC TTT TSEA BF BC BSEC BTT BSEA
 MF MC MSEC MTT MSEA / NOITER NOPROFILE;
run;

Proc logistic data=transpot DESCENDING;
where choice NE 2;
model choice = EDUC PRES TF TC TSEC TTT TSEA
               MF MC MSEC MTT MSEA / selection=stepwise slentry=0.30
sls=0.05;
run;

Proc logistic data=transpot DESCENDING;
where choice NE 1;
model choice = EDUC PRES BF BC BSEC BTT BSEA
               MF MC MSEC MTT MSEA / selection=stepwise slentry=0.30
sls=0.05;
run;

Proc logistic data=transpot DESCENDING;
where choice NE 3;
model choice = EDUC PRES TF TC TSEC TTT TSEA BF BC BSEC BTT BSEA
                / selection=stepwise slentry=0.30 sls=0.05;
run;

/*Proc logistic data=transpot;
where choice NE 2;
model choice = TF TSEC TTT MF MC MSEC MTT MSEA / influence iplots;
run;

Proc logistic data=transpot;
where choice NE 1;
model choice = BF BC BTT BSEA
              MC MTT MSEA / influence iplots;
run;

Proc logistic data=transpot;
where choice NE 3;
model choice = PRES TC TSEC TTT TSEA BC BSEC BTT
              BSEA / influence iplots;
```

```
run;

Proc logistic data=transpot DESCENDING;
where choice NE 2;
model choice = TF TSEC TTT MF MC MSEC MTT MSEA / maxiter=25 ;
output out=a pred=phat;
data b;
set a;
w = phat*(1-phat);
proc reg data=b;
weight w;
model choice = TF TSEC TTT MF MC MSEC MTT MSEA / TOL VIF;
run;

Proc logistic data=transpot DESCENDING;
where choice NE 1;
model choice = BF BC BTT BSEA MC MTT MSEA / maxiter=25 ;
output out=c pred=phat;
data d;
set c;
w = phat*(1-phat);
proc reg data=d;
weight w;
model choice = BF BC BTT BSEA MC MTT MSEA / TOL VIF;
run;

Proc logistic data=transpot DESCENDING;
where choice NE 3;
model choice = PRES TC TSEC TTT TSEA BC BSEC BTT BSEA  / maxiter=25 ;
output out=e pred=phat;
data f;
set e;
w = phat*(1-phat);
proc reg data=f;
weight w;
model choice = PRES TC TSEC TTT TSEA BC BSEC BTT
              BSEA  / TOL VIF;
run;

Proc freq DATA=transpot;
tables EDUC*PRES/chisq;
tables EDUC*CHOICE/chisq;
tables PRES*CHOICE/chisq;
run;

Proc ttest;
class EDUC;
var TC BC MC;
run;

Proc logistic data=transpot;
where choice NE 2;
model choice = TC TSEC TTT BC MF MC MTT MSEA / influence iplots;
run;

Proc logistic data=transpot;
where choice NE 1;
model choice = TF TSEC BF BSEC BTT
             BSEA MC MSEC MTT / influence iplots;
run;

Proc logistic data=transpot;
where choice NE 3;
model choice = TSEC TTT BF BC BSEC BTT
             BSEA MF MSEC / influence iplots;
run;

Proc logistic data=transpot DESCENDING;
where choice NE 2;
model choice = TF TSEC TTT MF MC MSEC MTT MSEA;
output out=a(keep= pred up lo chi dev)P=pred U=up L=lo RESCHI=chi
RESDEV=dev;
output out=b(keep= choice dTF dTSEC dTTT dMF dMC dMSEC dMTT dMSEA dev
chi pred)
DFBETAS=int dTF dTSEC dTTT dMF dMC dMSEC dMTT dMSEA dev P=pred
DIFDEV=dev DIFCHISQ=chi;
run;



Proc GPLOT data=b;
where choice NE 2;
```

```
Plot dev*pred chi*pred / vaxis=axis1 haxis=axis1;
symbol v=dot height=0.35;
axis1 minor=none width=2 major=(width=2)
run;

Proc logistic DES data=transpot DESCENDING;
where choice NE 1;
model choice = BF BC BTT BSEA MC MTT MSEA;
output out=a(keep= pred up lo chi dev)P=pred U=up L=lo RESCHI=chi
RESDEV=dev;
output out=b(keep= choice dBF dBC dBTT dBSEA dMC dMTT dMSEA dev chi
pred)
DFBETAS=int dBF dBC dBTT dBSEA dMC dMTT dMSEA dev P=pred
DIFDEV=dev DIFCHISQ=chi;
run;

Proc GPLOT data=b;
where choice NE 1;
Plot dev*pred chi*pred / vaxis=axis1 haxis=axis1;
symbol v=dot height=0.35;
axis1 minor=none width=2 major=(width=2)
run;

Proc logistic data=transpot DESCENDING;
where choice NE 3;
model choice = PRES TC TSEC TTT TSEA BC BSEC BTT BSEA ;
output out=a(keep= pred up lo chi dev)P=pred U=up L=lo RESCHI=chi
RESDEV=dev;
output out=b(keep= choice dPRES dTC dTSEC dTTT dTSEA dBC dBSEC dBTT
dBSEA dev chi pred)
DFBETAS=int dPRES dTC dTSEC dTTT dTSEA dBC dBSEC dBTT dBSEA dev P=pred
DIFDEV=dev DIFCHISQ=chi;
run;

Proc GPLOT data=b;
where choice NE 3;
Plot dev*pred chi*pred / vaxis=axis1 haxis=axis1;
symbol v=dot height=0.35;
axis1 minor=none width=2 major=(width=2);
run;
```

CHAPTER 4

ANALYSIS AND INTERPRETATION OF DATA

The aim of this chapter is to present the results generated by the SAS program developed in Chapter 3. We recall that three different codings of the same data set are studied and compared. The independent variables used in this study are: dichotomous (1 and 2), binary (0 and 1), and effect (-1 and 1) codings. Multinomial logit model is applied, but due to the procedures that SAS has, this model was estimated using logistic binary models where the diagnostics was also done.

4.1. Dichotomous coding of the explanatory variables

In this section the results of a dichotomous coding (1 and 2) are reported, analysed and interpreted.

4.1.1 Dichotomous coding: general test

The multinomial logit model was performed and used to estimate two logit models using PROC CATMOD procedure in SAS, and the results are given in Table 4.1. Each Wald chi-square is a test of the null hypothesis that the explanatory variable has no effect on the outcome variable, CHOICE. There are 2 degrees of freedom for each chi-squared because each variable has two coefficients. So the null hypothesis is that both coefficients are zero.

The variables that are statistically significant at the 0.05 significance level are the following:

(a) Bus traveling time (BTT) with p-value =0.0001
(b) Minibus cost (MC) with p-value =0.0001
(c) Minibus seating (MSEA) with p-value =0.0001
(d) Train security level (TSEC) with p-value =0.0002
(e) Bus security level (BSEC) with p-value =0.0015
(f) Minibus feeder (MF) with p-value =0.0025
(g) Train traveling time (TTT) with p-value =0.0032
(h) Minibus traveling time (MTT) with p-value =0.0053
(i) Bus cost (BC) with p-value =0.0062
(j) Bus cost (BC) with p-value =0.0062
(k) Bus seating (BS) with p-value = 0.0090

Education level (EDUC); train traveling cost (TC); train feeder (TF); presentation method (PRES), train seating (TSEA); and bus feeder (BF) are not statistically significant at the 0.05 level of significance.

Table 4.1: Results of the multinomial logit model using dichotomous coding

Source	DF	Chi-Square	P-value
Intercept	2	5.65	0.0592
Education level (EDUC)	2	0.19	0.9115
Presentation method (PRES)	2	2.24	0.3257
Train feeder (TF)	2	1.18	0.5550
Train traveling cost (TC)	2	1.17	0.5572
Train security level (TSEC)	2	16.66	0.0002
Train traveling time (TTT)	2	11.47	0.0032
Train seating (TSEA)	2	3.53	0.1711
Bus feeder (BF)	2	4.69	0.0957
Bus cost (BC)	2	10.17	0.0062
Bus security level (BSEC)	2	13.06	0.0015
Bus traveling time (BTT)	2	26.85	<.0001
Bus seating (BSEA)	2	9.42	0.0090
Minibus feeder (MF)	2	11.97	0.0025
Minibus cost (MC)	2	18.50	<.0001
Minibus security level (MSEC)	2	9.11	0.0105
Minibus traveling time (MTT)	2	10.47	0.0053
Minibus seating (MSEA)	2	22.15	<.0001
Likelihood Ratio	**198**	**202.87**	**0.3913**

The likelihood ratio statistic is equivalent to the deviance statistic which is twice the positive difference between the log-likelihoods for the fitted model and the saturated or full model (a model with all the variables). This statistic tests the null hypothesis that the hypothesized model fits the data against the alternative that the hypothesized model does not fit the data. The null hypothesis will be rejected if the p-value is less than the significance level of 0.05. Since the p-value (0.39) is not less than 0.05, then the null hypothesis cannot be rejected. Therefore the null hypothesis that the hypothesized model fits the data is not rejected because the p-value is greater than the significance level of 0.05, suggesting a good fit for the model.

Table 4.2 gives the results of the multinomial logit model which estimates two logit models. Each Wald chi-square is a test of the null hypothesis that the independent variable has no effect on the outcome variable, CHOICE.

Table 4.2: Results of the multinomial model with parameter estimates using dichotomous coding

Variable Name	Function number	Parameter Estimate	Standard Error	Wald Chi-square Statistics	P-value
Intercept	1	-0.9102	0.7737	1.38	0.2394
	2	-0.7025	0.9743	0.52	0.4709
EDUC	1	-0.00264	0.0284	0.01	0.9261
	2	0.0159	0.0371	0.19	0.6671
PRES	1	-0.0315	0.0756	0.17	0.6772
	2	-0.1453	0.0975	2.22	0.1360
TF	1	0.1640	0.2057	0.64	0.4253
	2	0.2610	0.2699	0.94	0.3335
TC	1	-0.1549	0.1521	1.04	0.3085
	2	-0.0185	0.1981	0.01	0.9257
TSEC	1	-0.2091	0.0765	7.46	0.0063
	2	0.1769	0.1003	3.11	0.0779
TTT	1	-0.00800	0.00278	8.28	0.0040
	2	0.00205	0.00368	0.31	0.5775
TSEA	1	0.00930	0.0758	0.02	0.9023
	2	0.1720	0.0968	3.16	0.0754
BF	1	0.0906	0.1097	0.68	0.4091
	2	0.3173	0.1465	4.69	0.0303
BC	1	0.1284	0.0596	4.64	0.0313
	2	-0.1044	0.0779	1.80	0.1802
BSEC	1	0.0761	0.0766	0.99	0.3205
	2	-0.2921	0.1009	8.37	0.0038
BTT	1	-0.00082	0.00311	0.07	0.7932
	2	-0.0207	0.00420	24.22	<.0001
BSEA	1	-0.0562	0.0766	0.54	0.4629
	2	-0.3313	0.1080	9.41	0.0022
MF	1	0.2537	0.0919	7.63	0.0057
	2	-0.1222	0.1325	0.85	0.3560
MC	1	0.1927	0.0446	18.63	<.0001
	2	0.1221	0.0594	4.23	0.0398
MSEC	1	0.0394	0.0770	0.26	0.6085
	2	0.3344	0.1097	9.28	0.0023
MTT	1	0.00904	0.00441	4.21	0.0403
	2	0.0187	0.00613	9.34	0.0022
MSEA	1	0.4789	0.1023	21.92	<.0001
	2	0.2674	0.1363	3.85	0.0498

Under the function number there is 1 and 2, where 1 denotes the model for train vs. minibus, and 2 for bus vs. minibus. There are two parameter estimates, two Wald chi-square statistics, two standard errors, and two p-values. The ones under the functional number 1 are for the model of train vs. minibus, and those under number 2 are for the model of bus vs. minibus. The interpretation of the parameter estimates and the odds ratio are provided.

Table 4.3 gives the estimate and the odds ratios of each variable for both two equations. The outcome/ dependent variable CHOICE is coded as:

1 = train, 2 = bus, 3 = minibus.

The odds ratios are calculated as $\exp(\hat{\beta})$ for each coefficient denoted by β.

4.1.2 Dichotomous coding: Train vs. minibus

Starting with the results of the first model for train vs. minibus, the following variables are *not statistically significant* at 0.05 significant level:

(a) Education level (EDUC), with p-value = 0.9261
(b) Train seating (TSEA), with p-value = 0.9023
(c) Bus traveling time (BTT), with p-value = 0.7932
(d) Presentation method (PRES), with p-value = 0.6772
(e) Minibus security level (MSEC), with p-value = 0.6085
(f) Bus seating (BSEA), with p-value = 0.4629
(g) Train feeder (TF), with p-value = 0.4253
(h) Bus feeder (BF), with p-value = 0.4091
(i) Bus security level (BSEC), with p-value = 0.3205
(j) Train cost (TC), with p-value = 0.3085

Minibus cost (MC); minibus seating (MSEA); train traveling time (TTT); minibus feeder (MF); train security level (TSEC); bus cost (BC); and minibus traveling time (MTT), are all *statistically significant* at 0.05 significant level. Parsons et al. (1999) established that not all the coefficients of the explanatory variables have the expected signs. Train security level (TSEC), bus cost (BC), minibus feeder (MF), minibus cost (MC) and minibus traveling time (MTT) now have the opposite sign.

The coefficients and odds ratios for train traveling time (TTT) and minibus seating (MSEA) give a highly significant negative effect on traveling time by train, and a highly significant positive effect on the availability of seats in a minibus. Each additional minute of traveling time reduces the odds of choosing a taxi by 1%, (100×(1-0.99)). The odds that a seat is often available in a train than in a minibus, is about 1.6 times the odds for seldom availability of seats.

Table 4.3: Results of the multinomial model using dichotomous coding with parameter estimates (Table 4.2 rearranged)

Variable Name	Train vs. Minibus			Bus vs. Minibus		
	Estimate	**P-value**	**Odds Ratio**	**Estimate**	**P-value**	**Odds Ratio**
Intercept	-0.9102	0.2394		-0.7025	0.4709	
EDUC	-0.00264	0.9261	1.00	0.0159	0.6671	1.02
PRES	-0.0315	0.6772	0.97	-0.1453	0.1360	0.86
TF	0.1640	0.4253	1.18	0.2610	0.3335	1.30
TC	-0.1549	0.3085	0.86	-0.0185	0.9257	0.98
TSEC	-0.2091	0.0063	0.81	0.1769	0.0779	1.19
TTT	-0.00800	0.0040	0.99	0.00205	0.5775	1.00
TSEA	0.00930	0.9023	1.01	0.1720	0.0754	1.19
BF	0.0906	0.4091	1.09	0.3173	0.0303	1.37
BC	0.1284	0.0313	1.14	-0.1044	0.1802	0.90
BSEC	0.0761	0.3205	1.08	-0.2921	0.0038	0.75
BTT	-0.00082	0.7932	1.00	-0.0207	<.0001	0.98
BSEA	-0.0562	0.4629	0.95	-0.3313	0.0022	0.72
MF	0.2537	0.0057	1.29	-0.1222	0.3560	0.88
MC	0.1927	<.0001	1.21	0.1221	0.0398	1.13
MSEC	0.0394	0.6085	1.04	0.3344	0.0023	1.40
MTT	0.00904	0.0403	1.01	0.0187	0.0022	1.02
MSEA	0.4789	<.0001	1.61	0.2674	0.0498	1.31

4.1.3 Dichotomous coding: Bus vs. minibus

The model for bus vs. minibus (Table 4.3) has the following variables which are *not statistically significant,* at 0.05 significant level:

(a) Train cost (TC), with p-value = 0.9357

(b) Education level (EDUC), with p-value = 0.96671

(c) Train traveling time (TTT), with p-value = 0.5775

(d) Minibus feeder (MF), with p-value = 0.3560

(e) Train feeder (TF), with p-value = 0.3335

(f) Bus cost (BC), with p-value = 0.1802

(g) Presentation method (PRES), with p-value = 0.1360
(h) Train security level (TSEC), with p-value = 0.0779
(i) Train seating (TSEA), with p-value = 0.0754

The variables which are *statistically significant* at the significance level of 0.05 are bus traveling time (BTT); bus seating (BSEA); minibus security level (MSEC); bus security level (BSEC); bus feeder (BF); minibus cost (MC); and minibus seating (MSEA). Amongst these statistically significant variables bus security level (BSEC), minibus cost (MC), and minibus traveling time (MTT) have opposite signs.

The coefficients of bus feeder (BF), bus traveling time (BTT) and minibus security level (MSEC) show highly positive effect on minibus security, highly negative effect on bus traveling time, and highly positive effect on availability of seats in a minibus, respectively. The odds ratio for bus feeder (BF) implies that any use of double transport (minibus and bus), decreases the odds of traveling by bus. Each additional minute of traveling time reduces the odds of using that mode by (100×(1-0.98)) = 2%. The odds that security is not improved in a train than in a minibus are about 1.4 times the odds for improved security.

4.2 Binary coding of the explanatory variables

The objective of Section 4.2 is to report on the analysis of a binary coding (0 and 1).

4.2.1 Binary coding: general test

Table 4.4 gives the results of the multinomial logit model which estimates two logit models using the binary coding (0 and 1) of the explanatory variables. Each Wald chi-square and the likelihood ratio statistics test the same null hypothesis as in Table 4.1.

The following variables are all *statistically significant* at 0.05 significance level, with their corresponding p-values as indicated
(a) Bus traveling time (BTT), with p-value less than 0.0001
(b) minibus cost (MC), with p-value less than 0.0001
(c) minibus seating (MSEA), with p-value less than 0.0001
(d) Train security level (TSEC), with p-value of 0.0002
(e) bus security level (BSEC) with p-value of 0.0015

(f) minibus feeder (MF) with p-value of 0.0025

(g) train traveling time (TTT), with p-value of 0.0032

(h) minibus traveling time (MTT) with p-value of 0.0053

(i) bus cost (BC) with p-value of 0.0062

(j) bus seating (BSEA) with p-value of 0.0090

(k) minibus security level (MSEC) with p-value of 0.0105

Bus traveling time (BTT), minibus cost (MC), and minibus seating (MSEA), all with a p-value less than 0.0001, are *highly significant* at 0.05 significance level. Education level (EDUC), train cost (TC), train feeder (TF), presentation method (PRES), train seating (TSEA), and bus feeder (BF) are *not statistically significant* at the 0.05 significance level.

The null hypothesis that the hypothesized model fits the data is accepted as shown in Table 4.4. A high p-value suggests a good fit. The likelihood ratio statistic has a p-value of 0.40 that supports a good fit

Table 4.4 Results of the multinomial logit model using binary coding

Variable Name	**Degrees of freedom**	**Wald Chi-Square**	**P-value**
Intercept	2	1.54	0.4632
EDUC	2	0.19	0.9115
PRES	2	2.24	0.3257
TF	2	1.18	0.5550
TC	2	1.17	0.5572
TSEC	2	16.66	0.0002
TTT	2	11.47	0.0032
TSEA	2	3.53	0.1711
BF	2	4.69	0.0957
BC	2	10.17	0.0062
BSEC	2	13.06	0.0015
BTT	2	26.85	<.0001
BSEA	2	9.42	0.0090
MF	2	11.97	0.0025
MC	2	18.50	<.0001
MSEC	2	9.11	0.0105
MTT	2	10.47	0.0053
MSEA	2	22.15	<.0001
Likelihood Ratio	**198**	**202.87**	**0.3913**

The interpretation of the parameter estimates and the odds ratio are provided in Section 4.2.2, below.

Table 4.5: Results of the multinomial logit model with parameter estimates using binary coding

Variable	Function Number	Parameter Estimate	Standard Error	Wald Chi-Square	P-value
Intercept	1	-0.9232	0.7514	1.51	0.2192
	2	-0.5853	0.9333	0.39	0.5305
EDUC	1	-0.0329	0.0765	0.19	0.6668
	2	-0.0171	0.0988	0.03	0.8624
PRES	1	-0.0348	0.0757	0.21	0.6459
	2	-0.1457	0.0975	2.23	0.1353
TF	1	0.1631	0.2058	0.63	0.4280
	2	0.2642	0.2700	0.96	0.3278
TC	1	-0.1567	0.1522	1.06	0.3031
	2	-0.0146	0.1981	0.01	0.9413
TSEC	1	-0.2101	0.0766	7.52	0.0061
	2	0.1787	0.1004	3.17	0.0751
TTT	1	-0.00805	0.00278	8.38	0.0038
	2	0.00204	0.00368	0.31	0.5782
TSEA	1	0.00836	0.0758	0.01	0.9122
	2	0.1719	0.0967	3.16	0.0756
BF	1	0.0912	0.1097	0.69	0.4059
	2	0.3174	0.1465	4.69	0.0303
BC	1	0.1295	0.0597	4.70	0.0301
	2	-0.1041	0.0780	1.78	0.1821
BSEC	1	0.0800	0.0767	1.09	0.2967
	2	-0.2869	0.1011	8.06	0.0045
BTT	1	-0.00078	0.00311	0.06	0.8016
	2	-0.0207	0.00420	24.25	<.0001
BSEA	1	-0.0577	0.0766	0.57	0.4517
	2	-0.3294	0.1079	9.31	0.0023
MF	1	0.2557	0.0919	7.73	0.0054
	2	-0.1278	0.1328	0.93	0.3359
MC	1	0.1910	0.0447	18.30	<.0001
	2	0.1188	0.0596	3.98	0.0461
MSEC	1	0.0366	0.0771	0.23	0.6346
	2	0.3261	0.1100	8.78	0.0030
MTT	1	0.00918	0.00441	4.33	0.0374
	2	0.0187	0.00613	9.31	0.0023
MSEA	1	0.4812	0.1023	22.12	<.0001
	2	0.2734	0.1360	4.04	0.0444

4.2.2 Binary coding: Train vs. minibus

Table 4.6 gives the results when estimating the same multinomial model, but using the binary coding (0 and 1) for the independent variables.

Table 4.6**:** Results of the multinomial logit model with parameter estimates using binary coding (Table 4.5 rearranged)

Variable Name	**Train vs. Minibus**			**Bus vs. Minibus**		
	Estimate	**P-value**	**Odds Ratio**	**Estimate**	**P-value**	**Odds Ratio**
Intercept	-0.9232	0.2192		-0.5853	0.5305	
EDUC	-0.00264	0.6668	1.00	-0.0171	0.8624	0.98
PRES	-0.0348	0.6459	0.97	-0.1457	0.1353	0.86
TF	0.1631	0.4280	1.18	0.2642	0.3278	1.30
TC	-0.1567	0.3031	0.85	-0.0146	0.9413	0.99
TSEC	-0.2101	0.0061	0.81	0.1787	0.0751	1.20
TTT	-0.00805	0.0038	0.99	0.00204	0.5782	1.00
TSEA	0.00863	0.9122	1.01	0.1719	0.0756	1.19
BF	0.0912	0.4059	1.10	0.3174	0.0303	1.37
BC	0.1295	0.0301	1.14	-0.1041	0.1821	0.90
BSEC	0.0800	0.2967	1.08	-0.2869	0.0045	0.75
BTT	-0.00078	0.8016	1.00	-0.0207	<.0001	0.98
BSEA	-0.0577	0.4517	0.94	-0.3294	0.0023	0.72
MF	0.2557	0.0054	1.29	-0.1278	0.3359	0.88
MC	0.1910	<.0001	1.21	0.1188	0.0461	1.13
MSEC	0.0366	0.6346	1.04	0.3261	0.0030	1.39
MTT	0.00918	0.0374	1.01	0.0187	0.0023	1.02
MSEA	0.4812	<.0001	1.62	0.2734	0.0444	1.31

For the results of the first model for train vs. minibus education level, the following variables are *not statistically significant* at the 0.05 level of significance

(a) Train seating (TSEA), with a p-value of 0.9122

(b) Bus traveling time (BTT), with a p-value of 0.8016

(c) Education level (EDUC), with a p-value of 0.6668

(d) Presentation method (PRES), with a p-value of 0.6459

(e) Minibus security level (MSEC), with a p-value of 0.6346

(f) Bus seating (BSEA), with a p-value of 0.4517

(g) Train feeder (TF), with a p-value of 0.4280

(h) Bus feeder (BF), with a p-value of 0.4059

(i) Train cost (TC), with a p-value of 0.3031

(j) Bus security level (BSEC), with a p-value of 0.2967

On the other hand, minibus cost (MC); minibus seating (MSEA); train traveling time (TTT); minibus feeder (MF); train security level (TSEC); bus cost (BC); and minibus traveling time (MTT); are statistically significant. Train security level, minibus traveling time, bus cost, minibus feeder and minibus cost, have opposite signs.

The coefficients of train traveling time and minibus seating have the following interpretations: A highly significant negative effect on traveling time and a highly significant positive effect on availability of seats, respectively. Each additional minute of train traveling time reduces the odds of commuting using that mode by 1% (i.e. 100 × (1-0.99)). The odds that a seat is often available in a train than in a minibus, is about 1.6 times the odds for seldom availability of seat.

4.2.3 Binary coding: Bus vs. minibus

The model for bus vs. minibus has the following variables which are *not statistically significant*

(a) Train cost (TC), with p-value=0.9413
(b) Education level (EDUC), with p-value=0.8624
(c) Train traveling time (TTT), with p-value=0.5782
(d) Minibus feeder (MF), with p-value=0.3359
(e) Train feeder (TF), with p-value=0.3278
(f) Bus cost (BC), with p-value=0.1821
(g) Presentation method (PRES), with p-value=0.1353
(h) Train seating (TSEA), with p-value=0.0756
(i) Train security level (TSEC), with p-value=0.0751

The *significant variables* at 0.05 level of significance, are bus traveling time (BTT); minibus security level (MSEC); bus seating (BSEA); minibus traveling time (MTT); bus security level (BSEC); bus feeder (BF); minibus seating (MSEA); and minibus cost (MC). Of these non-significant variables, bus feeder (BF), bus security level (BSEC), bus seating (BSEA) and minibus traveling time (MTT), bear opposite signs.

The results on bus traveling time (BTT), minibus security level (MSEC), and minibus seating (MSEA), show highly positive effect on security guards in minibus, highly negative effect on bus traveling time, and positive effect on availability of seats in minibuses. Each additional minute of traveling time reduces the odds of choosing that mode by (100×(1-0.98)) = 2%. The odds that security is not improved in a train than in a minibus are about 1.4 times the odds for improved security. The odds that a seat is often available in a train than in a minibus, is about 1.3 times the odds for seldom availability of seats.

4.3 Effect coding of the explanatory variables

Table 4.7 shows the results of the multinomial logit model which estimates two logit models using the effect coding (-1 and 1) of the explanatory variables. The results given in Table 4.7 (effect coding) are similar to those in Table 4.1(dichotomous coding) and those in Table 4.4 (the binary coding) except for those associated with the intercepts.

Table 4.7**:** Results of the multinomial logit model using effect coding

Source	Degrees of freedom	Wald Chi-Square	Pr > Chi-Square
Intercept	2	1.54	0.4632
EDUC	2	0.19	0.9115
PRES	2	2.24	0.3257
TF	2	1.18	0.5550
TC	2	1.17	0.5572
TSEC	2	16.66	0.0002
TTT	2	11.47	0.0032
TSEA	2	3.53	0.1711
BF	2	4.69	0.0957
BC	2	10.17	0.0062
BSEC	2	13.06	0.0015
BTT	2	26.85	<.0001
BSEA	2	9.42	0.0090
MF	2	11.97	0.0025
MC	2	18.50	<.0001
MSEC	2	9.11	0.0105
MTT	2	10.47	0.0053
MSEA	2	22.15	<.0001
Likelihood Ratio	**198**	**202.87**	**0.3913**

Table 4.8: Results of the multinomial logit model with parameter estimates using effect coding

Variable Name	Function Number	Parameter Estimate	Standard Error	Chi-Square	P-value
Intercept	1	-0.9232	0.7514	1.51	0.2192
	2	-0.5853	0.9333	0.39	0.5305
EDUC	1	-0.0329	0.0765	0.19	0.6668
	2	-0.0171	0.0988	0.03	0.8624
PRES	1	-0.0348	0.0757	0.21	0.6459
	2	-0.1457	0.0975	2.23	0.1353
TF	1	0.1631	0.2058	0.63	0.4280
	2	0.2642	0.2700	0.96	0.3278
TC	1	-0.1567	0.1522	1.06	0.3031
	2	-0.0146	0.1981	0.01	0.9413
TSEC	1	-0.2101	0.0766	7.52	0.0061
	2	0.1787	0.1004	3.17	0.0751
TTT	1	-0.00805	0.00278	8.38	0.0038
	2	0.00204	0.00368	0.31	0.5782
TSEA	1	0.00836	0.0758	0.01	0.9122
	2	0.1719	0.0967	3.16	0.0756
BF	1	0.0912	0.1097	0.69	0.4059
	2	0.3174	0.1465	4.69	0.0303
BC	1	0.1295	0.0597	4.70	0.0301
	2	-0.1041	0.0780	1.78	0.1821
BSEC	1	0.0800	0.0767	1.09	0.2967
	2	-0.2869	0.1011	8.06	0.0045
BTT	1	-0.00078	0.00311	0.06	0.8016
	2	-0.0207	0.00420	24.25	<.0001
BSEA	1	-0.0577	0.0766	0.57	0.4517
	2	-0.3294	0.1079	9.31	0.0023
MF	1	0.2557	0.0919	7.73	0.0054
	2	-0.1278	0.1328	0.93	0.3359
MC	1	0.1910	0.0447	18.30	<.0001
	2	0.1188	0.0596	3.98	0.0461
MSEC	1	0.0366	0.0771	0.23	0.6346
	2	0.3261	0.1100	8.78	0.0030
MTT	1	0.00918	0.00441	4.33	0.0374
	2	0.0187	0.00613	9.31	0.0023
MSEA	1	0.4812	0.1023	22.12	<.0001
	2	0.2734	0.1360	4.04	0.0444

Table 4.9 gives the results when estimating the same multinomial model, but using the effect coding (-1 and 1) for the independent variables. The results are the same as in Tables 4.3 and 4.6. This means that their interpretations will also be the same as those in Table 4.1, Table 4.2 and Table 4.3.

Table 4.9: Results of the multinomial logit model with parameter estimates using effect coding (Table 4.8 rearranged)

Variable Name	Train vs. Minibus			Bus vs. Minibus		
	Estimate	P-value	Odds Ratio	Estimate	P-value	Odds Ratio
Intercept	-0.9232	0.2192		-0.5853	0.5305	
EDUC	-0.00264	0.6668	1.00	-0.0171	0.8624	0.98
PRES	-0.0348	0.6459	0.97	-0.1457	0.1353	0.86
TF	0.1631	0.4280	1.18	0.2642	0.3278	1.30
TC	-0.1567	0.3031	0.85	-0.0146	0.9413	0.99
TSEC	-0.2101	0.0061	0.81	0.1787	0.0751	1.20
TTT	-0.00805	0.0038	0.99	0.00204	0.5782	1.00
TSEA	0.00863	0.9122	1.01	0.1719	0.0756	1.19
BF	0.0912	0.4059	1.10	0.3174	0.0303	1.37
BC	0.1295	0.0301	1.14	-0.1041	0.1821	0.90
BSEC	0.0800	0.2967	1.08	-0.2869	0.0045	0.75
BTT	-0.00078	0.8016	1.00	-0.0207	<.0001	0.98
BSEA	-0.0577	0.4517	0.94	-0.3294	0.0023	0.72
MF	0.2557	0.0054	1.29	-0.1278	0.3359	0.88
MC	0.1910	<.0001	1.21	0.1188	0.0461	1.13
MSEC	0.0366	0.6346	1.04	0.3261	0.0030	1.39
MTT	0.00918	0.0374	1.01	0.0187	0.0023	1.02
MSEA	0.4812	<.0001	1.62	0.2734	0.0444	1.31

4.4 The stepwise selection results

The SAS procedure used for the analysis in this section is PROC LOGISTIC which estimates the three models separately. It differs with PROC CATMOD in this way: PROC CATMOD uses all the observations at a time, and estimates two logit models while PROC LOGISTIC uses the observations of the two categories of the outcome variable which are specified in the model, and estimates the particular model.

In this section three models are specified separately. The models are, train versus minibus, bus versus minibus, and train versus bus. The PROC CATMOD procedure was discontinued because it has some limitations on some parts of the logistic regression. One cannot run the stepwise and diagnostics with PROC CATMOD of the multinomial logit model.

Using PROC LOGISTIC procedure, the stepwise procedure was performed and the probability defined for the independent variables to enter into the models was 0.3, and the probability defined for the variables to stay in the model was 0.05. Thus, Tables 4.10, 4.11 and 4.12 give the analysis of maximum likelihood estimates after the explanatory variables were selected by the stepwise procedure, where all the explanatory variables are significant at 0.05 significance level. The second part of the output in Tables 4.10, 4.11 and 4.12 are estimates of the odds ratios and their corresponding 95% confidence intervals.

The interpretation of Table 4.10 about the parameter estimates and the odds ratio is given as follows. The parameter estimates of train feeder (TF), train traveling time (TTT) and minibus feeder (MF), have opposite signs, whereas the estimates of minibus cost (MC), train security level (TSEC), minibus security level (MSEC), minibus traveling time (MTT) and minibus seating (MSEA), have the following interpretation respectively: a negative effect on minibus cost, significant positive effect on availability of guards at train stations, significant positive effect on availability of guards at minibus stations, significant negative effect on minibus train traveling time, and highly significant positive effect on availability of seats in a minibus.

Each additional cost of one rand for the use of a minibus reduces the odds of commuting by that mode by (100 × (1 – 0.83)) = 17%. The odds that security is not improved in train and minibus stations are respectively 0.7 and 1.4 times the odds for improved security. Each additional minute of traveling to work by minibus reduces the odds of choosing that mode by (100 × (1 – 0.99)) = 1%. The odds that a seat is often available in a minibus, is about 1.8 times the odds for seldom availability of seat.

Table 4.10: Parameter estimates, *Model: Train versus minibus*

Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Odds Ratio**	**P-value**
Intercept	1	0.1688	0.3508	0.2316	-	0.6303
TF	1	0.6087	0.1071	32.2759	1.838	<.0001
TSEC	1	-0.2936	0.1113	6.9565	0.746	0.0084
TTT	1	0.0103	0.00196	27.4412	1.010	<.0001
MF	1	0.2917	0.1274	5.2444	1.339	0.0220
MC	1	-0.1893	0.0313	36.4842	0.828	<.0001
MSEC	1	0.3550	0.1090	10.6117	1.426	0.0011
MTT	1	-0.0161	0.00305	27.6806	0.984	<.0001
MSEA	1	0.5694	0.1361	17.5058	1.767	<.0001

Table 4.10: Parameter estimates, *Model: Train versus minibus* (Continued)

Odds ratio estimates and their 95% confidence intervals

	Point	**95% Wald**	
Effect	**Estimate**	**Confidence Limits**	
TF	1.838	1.490	2.268
TSEC	0.746	0.599	0.927
TTT	1.010	1.006	1.014
MF	1.339	1.043	1.718
MC	0.828	0.778	0.880
MSEC	1.426	1.152	1.766
MTT	0.984	0.978	0.990
MSEA	1.767	1.353	2.307

The results for the model of bus versus minibus as given in Table 4.11 gives the parameter estimates of bus feeder (BF), bus cost (BC), bus traveling time (BTT), and bus seating (BSEA) with the opposite signs. The parameter estimates of minibus cost (MC), minibus traveling time (MTT) and minibus seating (MSEA), indicate a significant negative effect on minibus traveling time, significant positive effects on availability of security guards at minibus and train stations, and a negative effect on minibus cost, respectively. Each additional minute of minibus traveling time reduces the odds of choosing the mode by (100 × (1 – 0.98)) = 2%. Each additional rand in a minibus reduces the odds of choosing that mode by 20% (100 × (1 – 0.80)). The odds that a seat is often available in a minibus, is about 1.8 times the odds for seldom availability of seats.

Table 4.11: Parameter estimates, *Model: Bus versus minibus*

Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Odds Ratio**	**P-value**
Intercept	1	0.1864	0.4415	0.1782	-	0.6730
BF	1	0.5687	0.1806	9.9198	1.766	0.0016
BC	1	0.1732	0.0499	12.0280	1.189	0.0005
BTT	1	0.0218	0.00261	69.7392	1.022	<.0001
BSEA	1	-0.4514	0.1292	12.2062	0.637	0.0005
MC	1	-0.2300	0.0306	56.2947	0.795	<.0001
MTT	1	-0.0215	0.00367	34.4463	0.979	<.0001
MSEA	1	0.7003	0.1585	19.5231	2.014	<.0001

Odds ratio estimates and their 95% confidence intervals

	Point	**95% Wald**	
Effect	**Estimate**	**Confidence Limits**	
BF	1.766	1.240	2.516
BC	1.189	1.078	1.311
BTT	1.022	1.017	1.027
BSEA	0.637	0.494	0.820
MC	0.795	0.748	0.844
MTT	0.979	0.972	0.986
MSEA	2.014	1.476	2.748

It is noticed from Table 4.12 that the coefficients of the explanatory variables, train cost (TC) and train traveling time (TTT) have opposite signs. The coefficient of bus security level (BSEC) shows a high significant positive effect on the availability of security guards at train stations. The coefficients of bus security level (BSEC), bus cost (BC) and bus traveling time (BTT) show a high significant positive effect on availability of security guards at bus stations, a negative effect on bus cost, and a significant negative effect on traveling time, respectively. The coefficients of bus seating (BSEA) and train seating (TSEA) show a significant positive effect on availability of seats in buses and trains.

The odds that a seat is often available in a bus than in a train, is about 0.4 times the odds for seldom availability of seats. The odds that a seat is often available in a train than in a minibus is about 0.7 times the odds for seldom availability of seat. The odds for train security level (TSEC), bus security level (BSEC) and bus cost (BC) are interpreted respectively as follows: The odds that security is not improved in train stations than in minibus are about 0.5 times the odds for improved security. The odds that security is not improved in bus stations than in minibus are about 2.0 times the odds for improved security. Each additional cost of one rand in a bus reduces the odds of choosing a bus by 26%. Each additional minute of traveling to work by bus reduces the odds of choosing that mode by (100 × (1 – 0.98)) = 2%. Presentation method (PRES) is also significant at 0.05 level of significance, and its coefficient shows a positive effect on the presentation method. The odds for verbal presentation is about 1.4 times the odds for pictorial presentation.

Table 4.12: Parameter estimates, *Model: Train versus bus*

Variable	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Odds Ratio**	**P-value**
Intercept	1	0.2053	0.4205	0.2384	-	0.6253
PRES	1	0.3453	0.1381	6.2539	1.412	0.0124
TC	1	0.1867	0.0505	13.6527	1.205	0.0002
TSEC	1	-0.5935	0.1432	17.1762	0.552	<.0001
TTT	1	0.0150	0.00254	34.8623	1.015	<.0001
TSEA	1	-0.3107	0.1426	4.7443	0.733	0.0294
BC	1	-0.2992	0.0371	65.1880	0.741	<.0001
BSEC	1	0.6703	0.1437	21.7728	1.955	<.0001
BTT	1	-0.0194	0.00289	45.0206	0.981	<.0001
BSEA	1	0.3571	0.1400	6.5044	1.429	0.0108

Table 4.12: Parameter estimates, *Model: Train versus bus* (Continued)

Odds ratio estimates and their 95% confidence intervals

Effect	Point Estimate	95% Wald Confidence Limits	
PRES	1.412	1.078	1.851
TC	1.205	1.092	1.331
TSEC	0.552	0.417	0.731
TTT	1.015	1.010	1.020
TSEA	0.733	0.554	0.969
BC	0.741	0.689	0.797
BSEC	1.955	1.475	2.591
BTT	0.981	0.975	0.986
BSEA	1.429	1.086	1.880

The deviance statistic in Table 4.13 is contrasting the fitted model with the saturated model. It tests the null hypothesis that all the main effects and all the interaction terms among the independent variables are 0. Also the chi-squared statistic tests the same hypothesis with the deviance. The p-values of 0.03 and 0.04 suggest a poor goodness-of-fit for the model of train vs. minibus. The model of bus vs. minibus has the p-values of the deviance and Pearson Chi-square respectively as 0.0001 and <.0001 which shows a poor fit. The p-values 0.04 and 0.03 of the model of bus vs. train also indicate a poor fit of the Pearson chi-square, but low fit of the deviance.

Table 4.13: The deviance and Pearson goodness-of-fit statistics

	Deviance				**Pearson**			
Model	**Value**	**DF**	**Value/DF**	**P-value**	**Value**	**DF**	**Value/DF**	**P-value**
Train vs minibus	110.8483	84	1.3196	0.0265	106.9148	84	1.2728	0.0466
Bus vs. minibus	135.3290	81	1.6707	0.0001	155.2461	81	1.9166	<.0001
Bus vs. Train	103.7555	80	1.2969	0.0383	104.9955	80	1.3124	0.0320

The Hosmer and Lemeshow goodness-of-fit test whose results are reflected in Table 4.14 tests the null hypothesis that the model fits the data. This test also shows a good fit with p-values of 0.7659, 0.4271 and 0.5443 for the three models, respectively.

Table 4.14: Hosmer and Lemeshow goodness-of-fit test

Train vs. minibus			**Bus vs. minibus**			**Bus vs. Train**		
Chi-Square	**DF**	**P-value**	**Chi-Square**	**DF**	**P-value**	**Chi-Square**	**DF**	**P-value**
4.9222	8	0.7659	8.0655	8	0.4271	6.9287	8	0.5443

Table 4.15 gives the interpretation (for goodness-of-fit, testing global null hypothesis: beta=0) of the results after estimating the multinomial logit model using PROC LOGISTIC procedure in SAS. Three binary logit models were estimated separately, namely, the model of train versus minibus, bus versus minibus, and train versus bus. The null hypothesis of testing for the goodness-of-fit in the model is that the model fits the data, against the alternative hypothesis that the model does not fit the data. The deviance statistic is twice the positive difference between the log-likelihood for the fitted model and the saturated model (a model with all the variables).

When testing global null hypothesis: BETA=0, each of the three tables for each model gives the three statistics, namely Likelihood Ratio, Score and the Wald, which all are chi-squared statistics as provided in Table 4.15 below They all test the null hypothesis that all coefficients of the independent variables are zero against the alternative hypothesis that at least one of the coefficients of the independent variables is not zero. The null hypothesis is rejected at 0.05 level of significance. In this case we reject the null hypothesis that all coefficients of the independent variables are zero if p-value is less than 0.05 and are recorded as p<0.0001. This shows that all the coefficients are non-zero and implies that all the independent variables have an effect on the dependent variable CHOICE.

Table 4.15: Testing global null hypothesis: BETA=0

	Likelihood			**Score**			**Wald**		
Model	**Chi-Square**	**DF**	**P-value**	**Chi-Square**	**DF**	**P-value**	**Chi-Square**	**DF**	**P-value**
Train vs minibus	161.2928	8	<.0001	154.0120	8	<.0001	141.2437	8	<.0001
Bus vs. minibus	287.6845	7	<.0001	265.3668	7	<.0001	223.0293	7	<.0001
Bus vs. Train	236.6475	9	<.0001	222.3834	9	<.0001	191.2220	9	<.0001

4.5 Logistic regression diagnostics

After the model has been estimated, one needs to check if there are any unusual observations. The residuals are used to identify the observations not well explained by the model. They are also used to identify influential and outlying observations. This was done using graphical presentations.

LOGISTIC procedure in SAS uses the residuals' abbreviations as follows:

- RESDEV: The Deviance residual
- DIFDEV: Change in deviance (ΔD) with deletion of the observation
- RESCHI: The Chi-squared residual
- DIFCHISQ: Change in Pearson chi-square ($\Delta\chi^2$) with deletion of an observation.

The graphs (Figure 4.1– Figure 4.6), show the plots of the residuals (Deviance and Pearson) against case number (the observations). The main aim of these plots is to detect the outlying and influential observations. Cases with extremely large or small (say, more than 3 standard deviations from 0 in absolute value) residuals are declared to be influential. Large residuals, regardless of the sign, correspond to poor fit points.

4.5.1: Plots of the residuals against case number (observation)

Model: Bus versus train

Figure 4.1: The deviance residual against case number (observation)

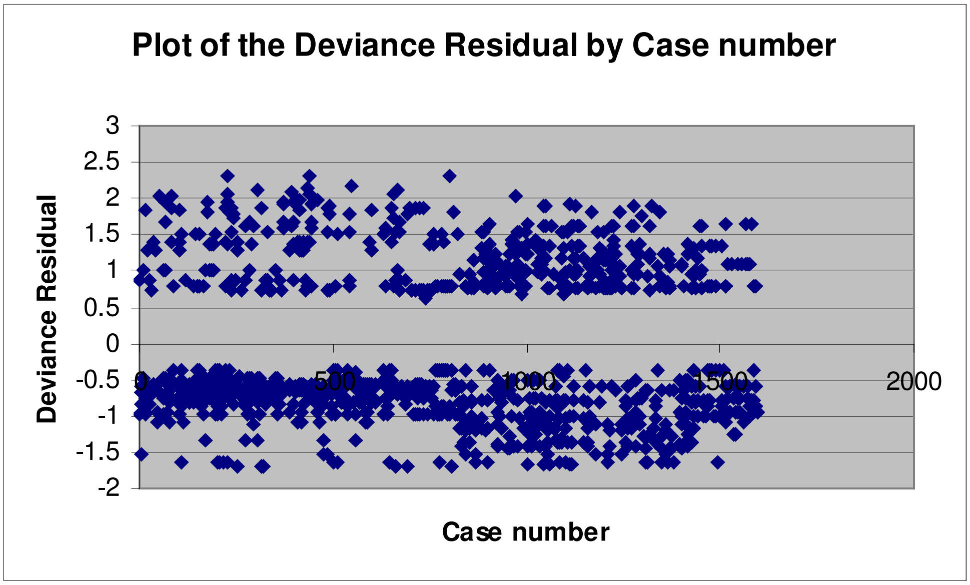
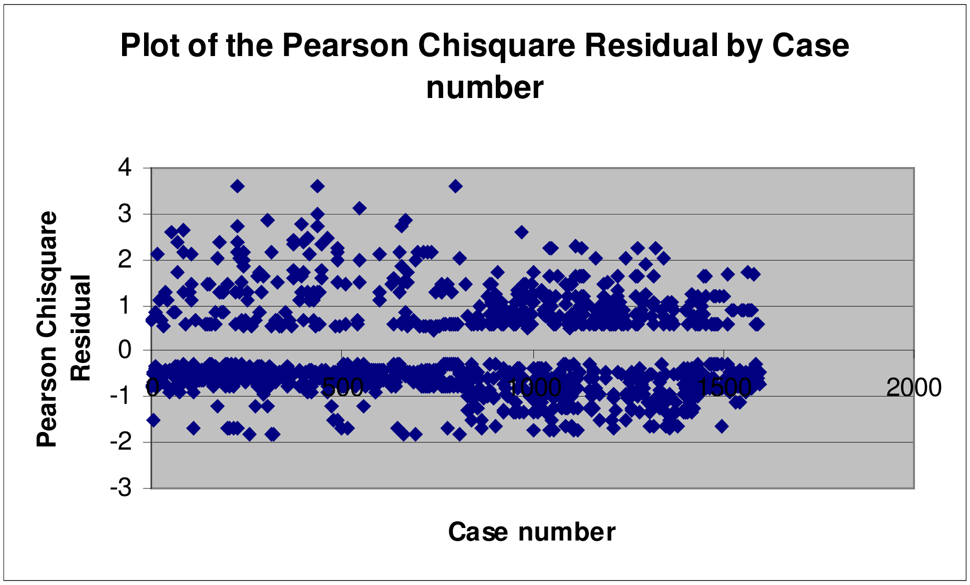


Figure 4.2: Pearson residual against case number



The plot of the model for **bus versus train** shows neither the influential nor the outlying observation on the deviance plot against case number since all the residuals fall within 3 standard deviations from 0. From the plot of the Pearson residual against each observation, *it has been noticed five observations* of which their residuals are greater than 3 namely, observations 226 (3.61934), 437 (3.61934), 800 (3.61934), 546 (3.1339) and 435 (3.00458) with residual values in brackets. The residual themselves are not extremely large, and since 95% of the residuals fall within 3 standard deviations of 0, then the model fits reasonably well.

Model: Train versus minibus

The plot of the Pearson Residual and deviance were also plotted against the case number to detect the influential and outlying observations for the model of minibus versus train as shown in Figure 4.3 and Figure 4.4.

Figure 4.3 and Figure 4.4 show no influential and outlying observations from both the deviance and Pearson residuals since all the residuals fall within 3 standard deviations from 0.

Figure 4.3: Pearson residual against case number

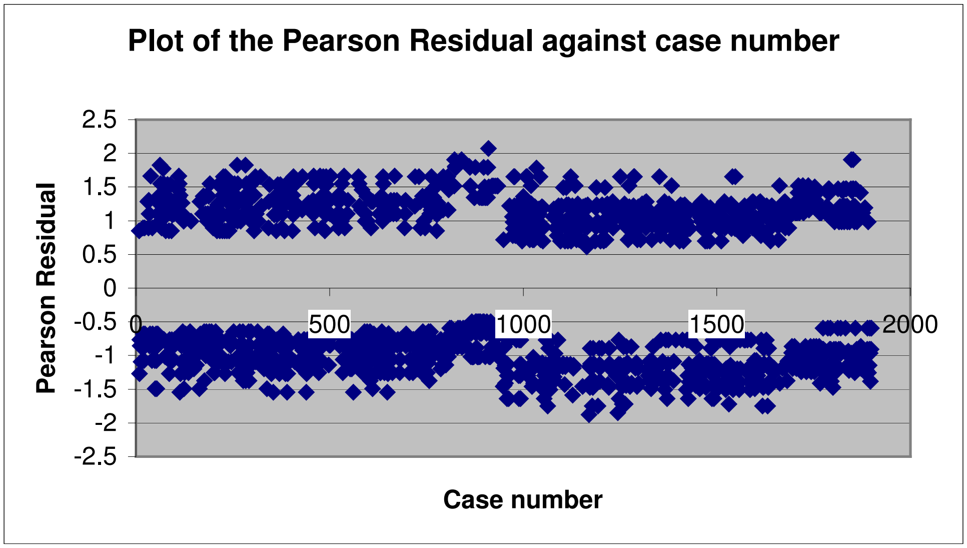
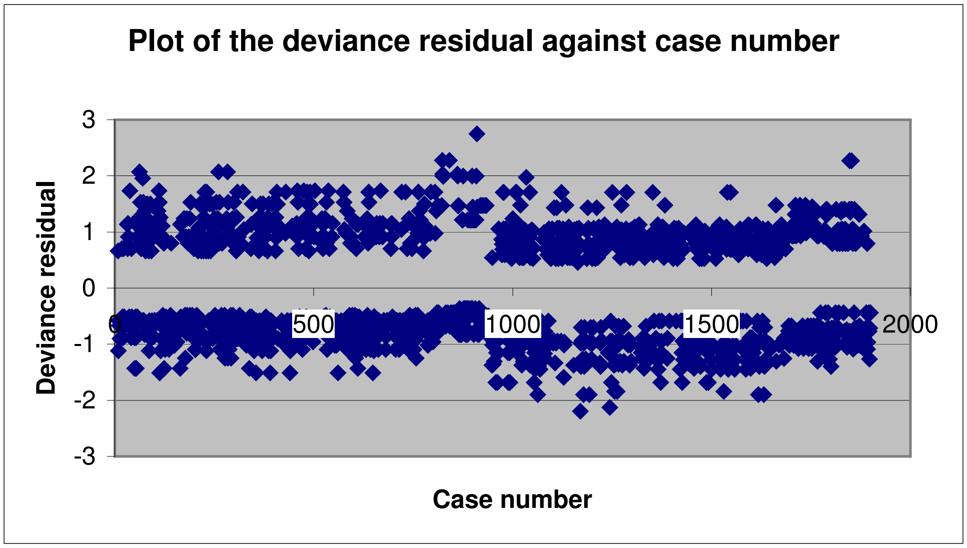


Figure 4.4: The deviance residual against case number



Model: Bus versus minibus

The plot of the Pearson residual and deviance against the case number follows from Figure 4.5 and Figure 4.6 to detect any influential and outlying observations for the model of bus versus minibus. Figure 4.5 and Figure 4.6 also shows no detection of influential and outlying observations from both the deviance and Pearson residuals since all the residuals fall within 3 standard deviations from 0.

Figure 4.5: The deviance residual against case number

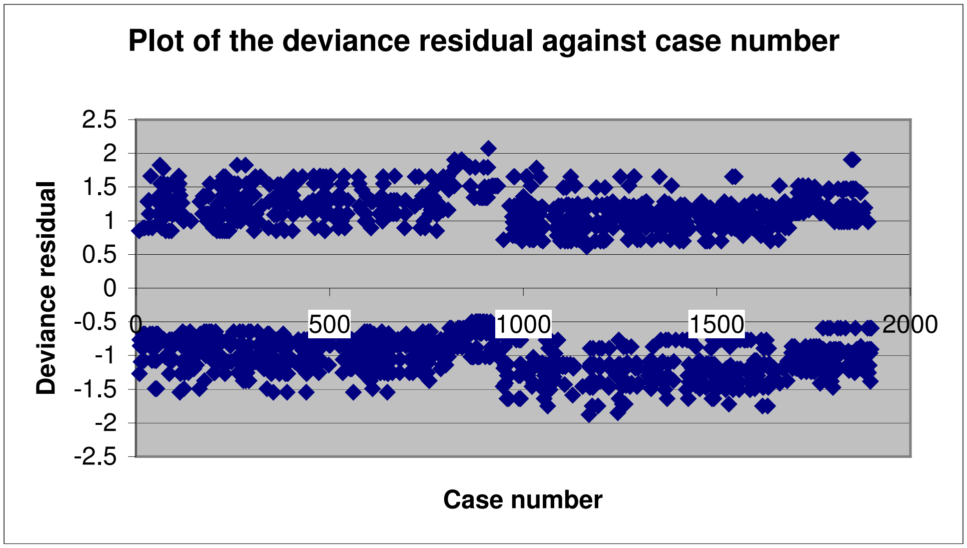
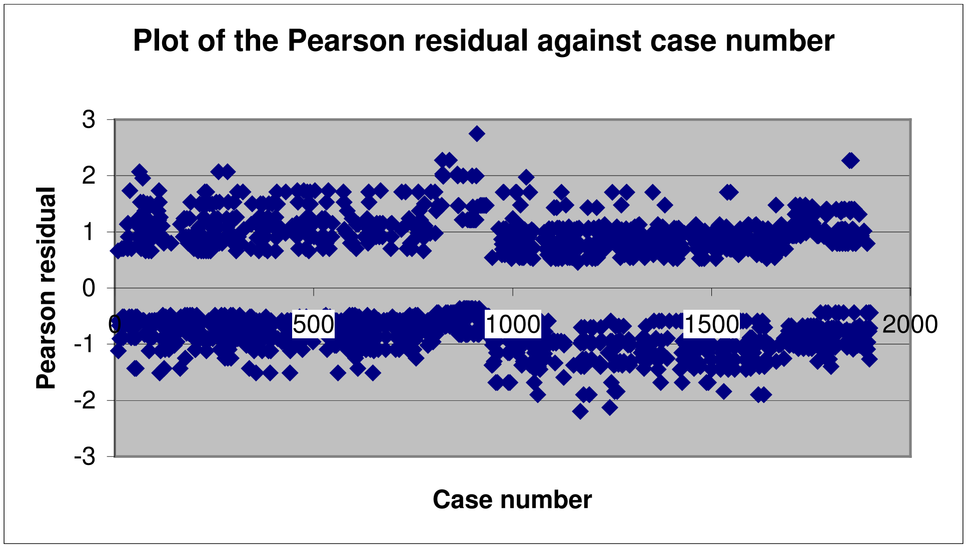


Figure 4.6: Plot of the Pearson residual against case number



4.5.2 Plots of the change in deviance and Pearson residuals against the predicted probabilities

Model: Bus versus train

The model of bus versus train seems to fit quite well. However, there is one point observed at the top left corner in Figure 4.8. This is the plot of Change in Pearson chi-square ($\Delta\chi^2$) with deletion of the observation against the estimated probability. Most of the values of $\Delta\chi^2$ and ΔD are less than 4, except this one point in Figure 4.8. According to Hosmer and Lemeshow (1989), 4 is used as a crude approximation to the upper ninetyfifth percentile of the distribution of $\Delta\chi^2$ and ΔD, since under m-asymptotics these quantiles would be distributed approximately as $\chi^2(1)$ with $\chi^2_{0.95}(1) = 3.84$.

Figure 4.7: DIFDEV (ΔD) statistics versus predicted probability

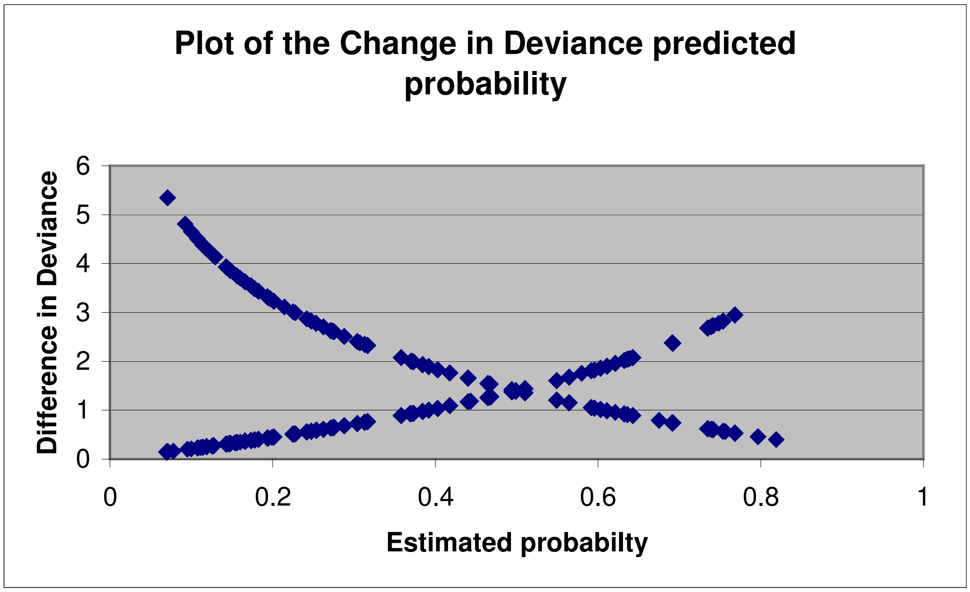
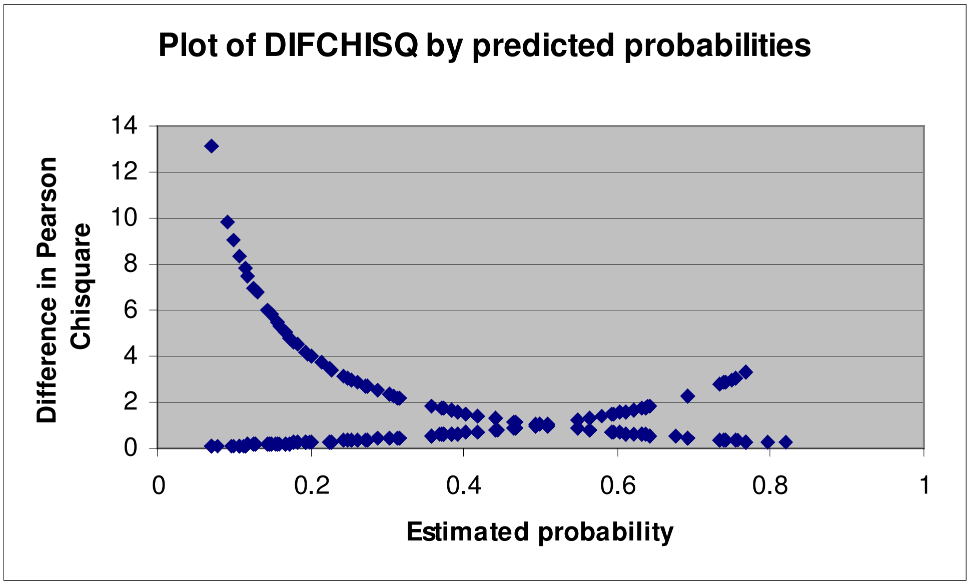


Figure 4.8: DIFCHISQ ($\Delta\chi^2$) statistics versus predicted probabilities



Model: Bus versus minibus

The diagnostics for the model of bus versus minibus are done in Figures 4.9 and 4.10 Most of the values for the diagnostics statistics, $\Delta\chi^2$ and ΔD, are less than 4. Figure 4.9 shows no observation that is suspected to be poorly fitted. The plot shows that the model fits the data reasonably well. In Figure 4.10, there is an observation at the top right corner far from the rest. Its value is however, not that large to conclude that the model does not fit the data well.

Figure 4.9: DIFDEV (ΔD) statistics versus predicted probabilities

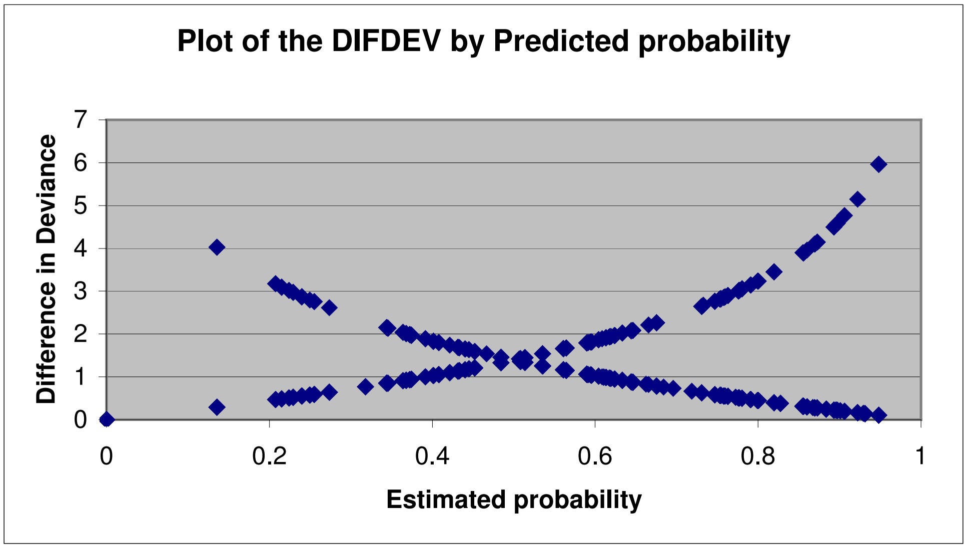
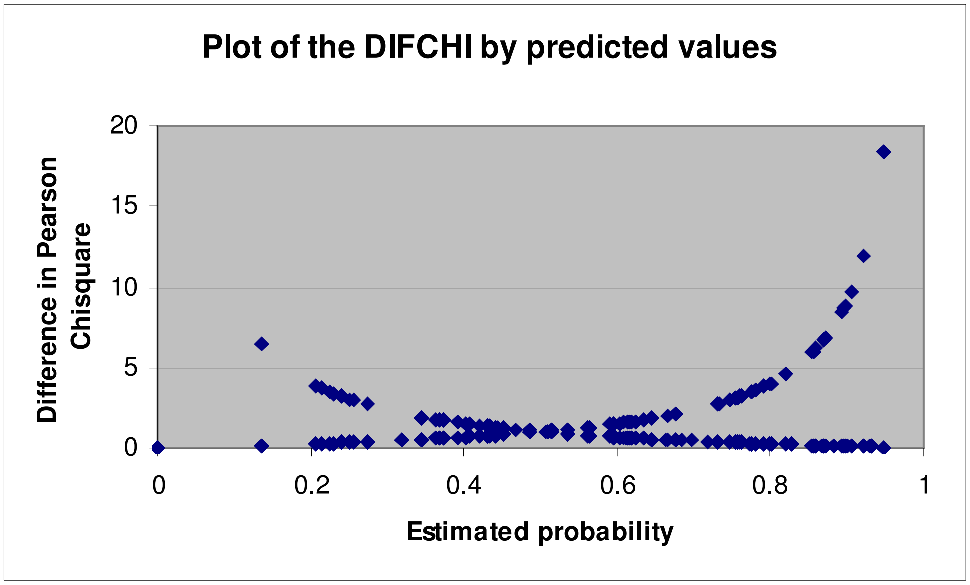


Figure 4.10: DIFCHISQ ($\Delta\chi^2$) statistics versus predicted probabilities



Model: Train versus minibus

Examining Figures 4.11 and 4.12 there is one point at the top left corner of Figure 4.12. Numerically the value is not that large in terms of the distance between that point and the before it. In addition, the corresponding predicted probability is small. Most of the values of $\Delta\chi^2$ are less than 4 and the predicted probability of the distant point is also small (about 0.1) whereas its value of $\Delta\chi^2$ is about 7.5 of which is not that large. Therefore the plot shows that a model fits reasonably well. In Figure 4.11, most of the values of ΔD seem to have a good fit.

Figure 4.11: DIFDEV (ΔD) statistics versus predicted probabilities

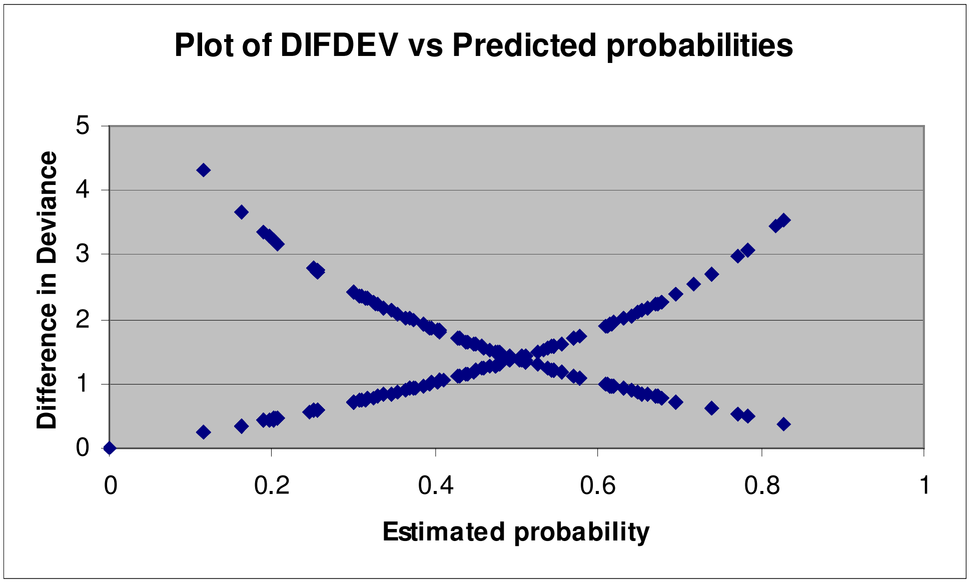
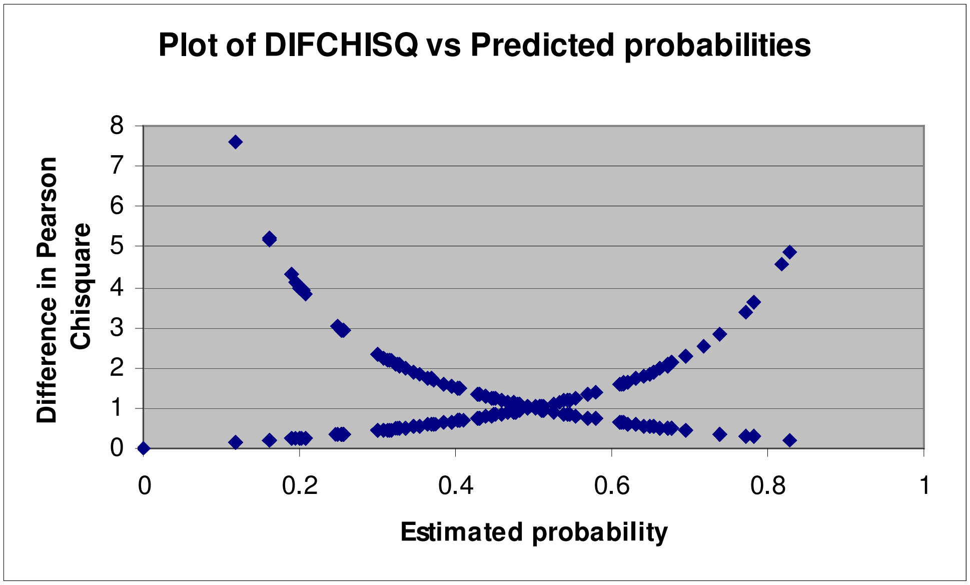


Figure 4.12: DIFCHISQ ($\Delta\chi^2$) statistics versus predicted probabilities



4.6 Assessing the problem of multicollinearity

In multiple linear regression: The purpose of a regression model is to find out to what extent the outcome (dependent variable) can be predicted by the independent variables. The strength of the prediction is indicated by R^2, also known as *variance explained* or *coefficient of determination*. It is important to notice that the value of R^2 alone cannot specify how well the model is explained

The absence of multi-collinearity is also essential to a multinomial logit model. In regression when several predictors (regressors) are highly correlated, this problem is called multi-collinearity or collinearity. When things are related, they are *linearly dependent* on each other because one can nicely fit a straight regression line to pass through many data points of those variables. Collinearity simply means *co-dependence*. Including too many regressors (explanatory variables) in a regression model often causes the problem of multi-collinearity. It is a common misconception that *stepwise regression* enables a researcher to select a subset of variables based upon their relative "importance." Indeed if variables are correlated, the "importance" of the variables are tied to the selection order.

Model: Train vs. Minibus

Tables 4.16, 4.17 and 4.18 test for multi-collinearity (by stepwise procedure), amongst the selected independent variables. Two common measures for assessing collinearity are tolerance (denoted by TOL) and its inverse, the variance inflation factor (denoted by VIF). These measures tell us the degree to which each independent variable is explained by other independent variables. Small tolerance values or large VIF values denote a high collinearity since $VIF = 1/TOL = 1/(1 - R^2)$ and $TOL = 1 - R^2$. A threshold for VIF is 10 which corresponding to a tolerance of 0.10.

All the three tables show no multi-collinearity. All variance inflation factors are less than 10, or all the tolerances are greater than 0.10 which indicates that there is no problem of multi-collinearity with the independent variables.

Table 4.16: Results of multicollinearity for model of train versus minibus

Variable	DF	Tolerance	Inflation
Intercept	1	.	0
TF	1	0.98316	1.01713
TSEC	1	0.99550	1.00452
TTT	1	0.97175	1.02907
MF	1	0.69513	1.43858
MC	1	0.69292	1.44317
MSEC	1	0.98866	1.01147
MTT	1	0.98773	1.01243
MSEA	1	0.98186	1.01848

Table 4.17: Results of multicollinearity for model of bus versus minibus

Variable	DF	Tolerance	Variance Inflation
Intercept	1	.	0
BF	1	0.48967	2.04217
BC	1	0.49335	2.02698
BTT	1	0.97075	1.03013
BSEA	1	0.94824	1.05458
MC	1	0.97219	1.02860
MTT	1	0.95725	1.04466
MSEA	1	0.97941	1.02102

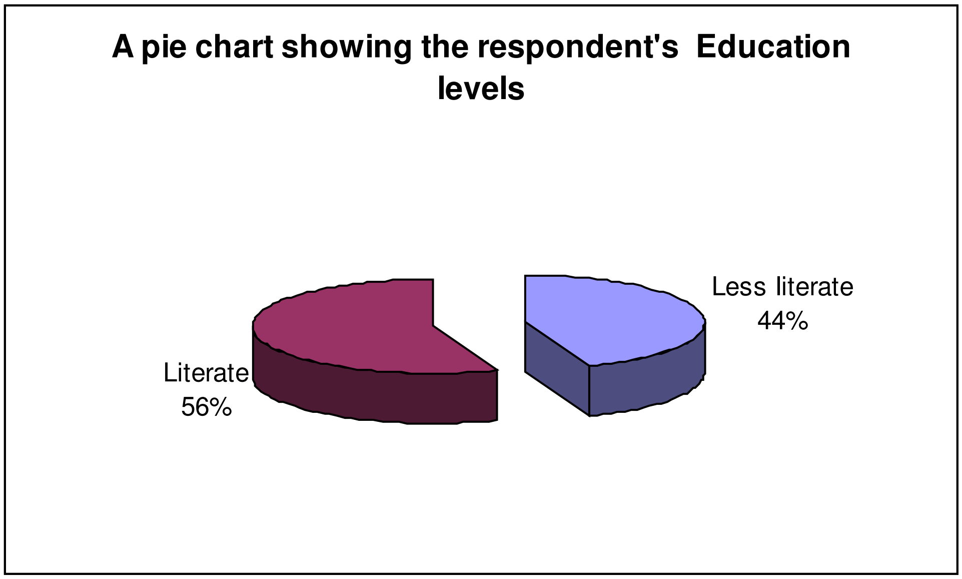
Table 4.18: Results of multicollinearity for model of bus versus train

Variable	DF	Tolerance	Variance Inflation
Intercept	1	.	0
PRES	1	0.98248	1.01783
TC	1	0.97629	1.02428
TSEC	1	0.98772	1.01243
TTT	1	0.95883	1.04294
TSEA	1	0.91568	1.09208
BC	1	0.98637	1.01382
BSEC	1	0.97690	1.02364
BTT	1	0.90103	1.10984

4.7 Tests for the difference between the mean costs of less literate and literate commuters

All commuters who passed up to Standard five (Grade 7) are declared to be less literate and those who passed Standard six or higher are literate. Figure 4.13 shows that the sample has more literate commuters than the less literate ones.

Figure 4.13: A bar graph of the number of less literate and literate commuters



The t-test is used to test the difference between the means of the two populations. In general before carrying out a test for the difference between the two populations means, we first need to test for the equality of variances for the very same two populations. From Table 4.19, the test for the difference between the means of train cost (TC), bus cost (BC) and minibus cost (MC) is performed between less literate and literate commuters. From Table 4.20, the p-values for all three variables assure equality of variances. Looking at the results in Table 4.20, under "Equality of Variances" shows an insignificant difference between the mean costs for all variables. This implies that there is no difference in the mean costs of train, bus and minibus between less literate and literate commuters.

Table 4.19: Results of the means for commuters' costs

Var	**EDUC**	**N**	**Mean**	**Lower CL Mean**	**Upper CL Mean**	**Std Dev**	**Lower CL Std Dev**	**Upper CL Std Dev**	**Std Error**
TC	less literate	1883	2.7974	2.8581	2.9188	1.3014	1.3429	1.3872	0.0309
TC	literate	2272	2.7966	2.8522	2.9078	1.3125	1.3507	1.3911	0.0283
TC	Diff (1-2)		-0.076	0.0059	0.0882	1.3188	1.3472	1.3768	0.042
BC	less literate	1888	5.258	5.3397	5.4214	1.7544	1.8104	1.87	0.0417
BC	literate	2240	5.2127	5.2877	5.3627	1.758	1.8095	1.8641	0.0382
BC	Diff (1-2)		-0.059	0.052	0.1629	1.7719	1.8101	1.85	0.0566
MC	less literate	1969	6.6035	6.6971	6.7906	2.0535	2.1176	2.1859	0.0477
MC	literate	2463	6.7056	6.7881	6.8706	2.0317	2.0884	2.1484	0.0421
MC	Diff (1-2)		-0.216	-0.091	0.0333	2.0588	2.1017	2.1464	0.0635

Table 4.20: Results of the variances for commuters' costs

T-Tests (for equality of means)

Variable	Method	Variances	DF	t Value	P-value
TC	Pooled	Equal	4153	0.14	0.8883
TC	Satterthwaite	Unequal	4019	0.14	0.8882
BC	Pooled	Equal	4126	0.92	0.3578
BC	Satterthwaite	Unequal	4008	0.92	0.3578
MC	Pooled	Equal	4430	-1.43	0.1519
MC	Satterthwaite	Unequal	4192	-1.43	0.1525

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	P-value
TC	Folded F	2271	1882	1.01	0.7952
BC	Folded F	1887	2239	1.00	0.9811
MC	Folded F	1968	2462	1.03	0.5143

CHAPTER 5

CONCLUSIONS

The multinomial model, using three different codings, was fitted to the stated preference data for commuters in Mamelodi, east of Pretoria, and the overall models were statistically significant.

After selecting the significant variables by the stepwise selection procedure, and fitting only the significant explanatory variables in the logistic regression models, the three models were all statistically significant for the Hosmer and Lemeshow goodness-of-fit statistics. The three models were all not statistically significant for the Pearson and deviance goodness-of-fit statistics. It has been found that there is no significant difference in the mean costs (train cost, bus cost and minibus cost) of literate and less literate commuters.

Education level (EDUC) was not statistically significant in the multinomial logit model. Even when using different codings, education level fails to reach statistical significance. Thus education level does not have any effect on the stated preference choice of transport.

Presentation method (PRES) was not statistically significant in the multinomial logit model. When using different codings presentation method also fails to reach statistical significance. After the stepwise logistic regression model, PRES was found to be statistically significant for the model of train versus bus only. This is an indication of knowledge of pictures and verbal communication amongst commuters with standard ten (Grade 12 or matric) or lower, as their highest education level.

Although logistic regression diagnostics reveals some of the suspicious observations that are not well fitted, the models fit moderately well with the Pearson residuals. From the deviance statistic, all the three models were found to have no outlying observations, whereas with the Pearson statistic only the model of bus versus train has some outlying observations. The deviance statistics shows that all the three models fit quite well.

Since this analysis did not include some of the factors that may be of importance to the choice of transport, namely, monthly income and some biographical characteristics (e.g., gender, area), it is recommended that other research include these indicators in the analysis and not apply the stated preference. It is also recommended that the respondents should be extended to matric plus certificate and higher; and also extended to private car as one of the mode of transport. For all the three models, it has been found that different factors affect commuters in choosing their mode of transport. These factors permit government to design relevant strategies and policies to improve the needs of commuters in the CBD of Pretoria.

References

Agresti, A. (1996) *An introduction to Categorical Data Analysis,* John Wiley, New York.

Akerstedt, T., Fredlund, P., Gillberg, M. and Jansson, B. (2002) Work load and work hours in relation to disturbed sleep and fatigue in a large representative sample. *Journal of Psychosomatic Research*, **53**, 585-588.

Allison, P.D. (1999) *Logistic Regression Using the SAS System: Theory and Application,* NC: SAS Institude Inc, Cary.

Banerjee, S., Weston, A. P., Zoubine, M. N., Campbell, D. R. and Cherian, R. (2000) Expression of Cdc2 and Cyclin B1 in Helicobacter pylori-Associated Gastric MALT and MALT Lymphoma. *The American Journal of Pathology*, **156**, 217-225.

Bankole, A., Darroch, J. E. and Singh, S. (1999) Determinants of Trends in Condom Use in the United States, 1988-1995. *Family Planning Perspectives*, **31**, 264-271.

Bender, R. and Grouven, U. (1998) Using Logistic Regression Models for Ordinal Data with Non-Proportional Odds. *Journal of Clinical Epidermiology*, **51**, 809-816.

Cox, M., Barbier, E.B., White, P.C.L., Newton-Cross, G.A., Kinsella, L. and Kennedy, H.J. (1999) Public preferences regarding rabies-prevention policies in the UK. *Preventive Veterinary Medicine*, **41**, 4, 257-270.

Collet, D. (1991) *Modelling Binary Data,* Chapman and Hall, London.

Del Mistro R. (2004) *The applicability of Stated Preference Among Less-Literate Commuters*.

DeShazo, J.R., and Fermo, G. (2002) Designing Choice Sets for Stated Preference methods: The effects of Complexity on Choice Consistency. *Journal of Environmental Economics and Management*, **44**, 123-143.

Dobson, A. J. (1990) *An Introduction to Generalized Linear Models*, Chapman and Hall, London.

Duerksen, S.C., Elder, J.P., Arredondo, E.M., Ayala, G.X., Slymen, D.J., Campbell, N.R. and Baquero, B. (2007) Family restaurant Choices are associated with Child and Adult Overweight Status in Mexican-American Families. *Journal of the American Dietetic Association*, **107**, 849-853.

Elango, B. and Sambharya, R.B. (2004) The influence of industry structure on the entry mode choice of overseas entrants in manufacturing industries. *Journal of International Management*, **10**, 107-124.

[i] Espino, R., et al., Analyzing the effect of preference heterogeneity on willingness to pay for improving service quality in an airline choice context. Transport. Res. Part E (2007), doi:10.1016/j.tre.2007.05.007.

Fahrmeir, L. and Tutz, G. (1994) *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer, New York.

Hosmer, D. W. and Lemeshow, S. (1989) *Applied Logistic Regression,* John Wiley, New York.

Kleinbaum, D. W. and Klein, M. (2002) *Logistic Regression*: *A Self-Learning Text*, 2nd edn, Springer – Verlag, New York.

Labour force survey, September 2001, Statistics South Africa, P0210.

[ii] Loo, B.P.Y., Passengers' airport choice within multi-airport regions (MARs): some insights from a stated preference survey at Hong Kong International Airport, J. Transp. Geogr. (2007), doi:10.1016/j.jtrangeo.2007.05.003.

Moore, D. S. and McCabe, G. P. (1998) *Introduction to the Practice of Statistics*, W.H. Freeman and Company, New York.

Ortuzar, J. D. and Willumsen, L. G. (1999) *Modelling Transport*, 2nd edn, John Wiley, Chichester.

Ott, R. L. (1993) *An Introduction to Statistical Methods and Data Analysis, Fourth Edition*, Duxbury Press, California.

Parsons, G. R., Jakus, P. M. and Tomasi, T. A. (1999) Comparison Of Welfare Estimates From Four Models For Linking Seasonal Recreational Trips To Multinomial Logit Models Of Site Choice. *Journal of Environmental Economics and Management*, **38**, 143-157.

Pauw, J. (1999) *A Simulation Study of the Effect of Therapaedic Horse Riding: A Logistic Regression Approach*. MSc. Thesis.

Phipps, A.G. (1984) Residential search and choice of displaced households. *Socio-Economic Planning Sciences*, **18**, 25-35.

Reise, S. P. (2000) Using Multilevel Logistic Regression to Evaluate Person-Fit in IRT Models. *Multivariate Behavioral Research*, **35**, 543-568.

SAS Program Version 8.

Sharma, S. (1996) *Applied Multivariate Techniques,* John Wiley, New York.

Statistics South Africa, Census 2001.

Street, D. and Burgess, L. (2004) Optimal Stated preference Choice experiments when all choice sets contain a specific option, *Office Journal of the International Indian Statistical Association*, **1**, 37-45.

[iii] Sze, N.N. and Wong, S.C., Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes, Accid. Anal. Prev. (2007), doi:10.1016/j.aap.2007.03.017.

Underhill, L. and Bradfield, D. (1994) *IntroSTAT 5.0*, Juta & Co, Ltd, University of Cape Town.

Van Wezel, M., and Potharst, R. (2007) Improved customer choice predictions using ensemble methods. *European Journal of Operational Research*, **181**, 436-452.

[iv] Varghese RT, et al., Determinants of the quality of life among diabetic subjects in Kerala, India. Diab Met Syndr Clin Res Rev (2007), doi:10.1016/j.dsx.2007.05.005.

Wrigley, N. (1985) *Categorical Data Analysis for Geographers and Environmental Scientists*, Longman, New York.

www.sanpad.org.za, May 2007.

Yannis, G., Kanellopoulou, A., Aggeloussi, K. and Tsamboulas, D. (2005) Modelling driver choices towards accident risk reduction. *Safety Science*, **43**, 173-186.

Young, T., Torner, J.C., Sihler, K.C., Hansen, A.R., Peek-Asa, C. and Zwerling, C. (2003) Factors associated with mode of transport to acute care hospital in rural communities. *The journal of Emergency Medicine*, **24**, 189-198.

Zandvliet, R., Dijst, M. and Bertolini, L. (2006), Destination choice and the identity of places: A disaggregated analysis for different types of visitor population environment in the Netherlands. *Journal of Transport Geography*, **14**, 451-462.

[i] **Espino, R., et al. (2007) is referenced the way the authors wanted.**

[ii] **Loo, B.P.Y (2007) is referenced the way the authors wanted.**

[iii] **Sze, N.N. and Wong, S.C. (2007) is referenced the way the authors wanted.**

[iv] **Varghese RT, et al. (2007) is referenced the way the authors wanted.**

APPENDIX A

QUESTIONNAIRE

Interviewer

Lit- | Respondent

South African - Netherlands
Research Programme on
Alternatives in Development

APPLICABILITY OF STATED PREFERENCE
TECHNIQUE AMONG URBAN COMMUTERS
MAKING MODE CHOICES

University of Pretoria

WELCOMING THE RESPONDENT

Thank you for coming here today and taking part in our project. We appreciate it, and hope that you will enjoy it. We would like to assure you at the outset that there are no right or wrong answers to the questions we are going to ask you; you simply have to tell us what you think. There is nothing to be afraid of, and we will not reveal your name to anyone outside the project team. I would like to start by telling you a bit about the purpose of the survey and then the programme which we will follow.

When the City Council plans for transport services for Pretoria it needs to know what passengers will do if the conditions of the services change. For instance if the fares change, or the time it takes to get to work, or if security is improved or if a passenger will be provided with a seat more frequently or not.

Surveys are used to ask the passenger what they would do if the conditions change. I want to stress that the purpose of this survey to is to find out what is the best way to ask questions when doing surveys.

The survey is a research project being done by students from the Universities of Pretoria and the North and is funded by the Netherlands Government.

The interview this morning / afternoon has five sections:

- *In the first section; I will ask you questions about how you usually travel to work.*
- *In the second section we will work through an example to see what kind of information is used to make decisions*
- *In the third section; I will ask you to choose between different ways of getting from Mamelodi to the city. I will do this in two ways: using pictures and words and using words only*
- *In the fourth section; I will ask you some questions about how you found the interview procedure*
- *Then in the fifth section; I will ask you some questions about how you make decisions about other things; such as decorating your house, or buying furniture*
- *Then after all this work we can all have a cool drink and something to eat*

Interview started at ....

Interview was finished at...................

For further information on this questionnaire or study please contact:
Prof.Romano Del Mistro, Department of Civil Engineering, University of Pretoria, Pretoria, 0002
Phone: (012) 420 2184; Fax (012) 362 5218; E-mail rmistro@eng.up.ac.za

Interviewer	□	□

Lit-	Respondent	□	□	□

SECTION 1 REVEALED PREFERENCE SURVEY

In this section we are talking about the trip that you usually make to work.

1.1 *What is your home address*

1.2 *What is you work address:*..

(If this is not the CBD or Sunnyside or Arcadia stop the interview and let Esau tell the interviewee that the interview is over)

1.3 *How do you usually travel to work?*

Train only	0	1
Bus only	0	2
Minibus only	0	3
Double transport (Specify...		

1.4 *What time do you usually leave home to* ***get to work?*** □□:□□

1.5 *What time do you usually get to work?.* □□:□□

1.6 *How much does it cost you to get to* work?
(Ask the respondent to give you the amount either per trip, per day; per week or per month and note to which period the amount refers)

R	□	□	□	.	□	□

per trip	1	per day	2	per week	3	per month	4

1.7 *Why did you choose this way to travel to work rather than any other?*

□□

□□

□□

Interviewer □ Lit- Respondent □□

1.8 *Could you have travelled to work in another way?* YES | NO

(If the respondent says no at first, try to get the respondent to think about it and give an answer to this question. If this is not possible go to the next page)

1.9 *How could you have travelled to work?*

Train only	0	1
Bus only	0	2
Minibus only	0	3
Double transport (Specify……………………………………………………		

At what time would you have to leave home to get to work using this way? □□:□□

At what time would you get to work using this way? □□:□□

How much would it cost you to travel to work using this way?

(Ask the respondent to give you the amount either per trip, per day; per week or per month and note to which period the amount refers)

R □□□□.□□ /trip 1 /day 2 /week 3 /month 4

How many times have you used this way to travel to work this month? □□

Why do you not travel to work more often **(Make sure that you get a reply form the rspondent)***?*

□□

□□

□□

Interviewer			Lit-	Respondent		

SECTION 2: STATED PREFERENCE EXAMPLE

In this section we are going to do an example of the questions that you will be asked in the next section.

In these questions I will describe two or three ways that can be used to get to work

will describe these ways by telling you

- *Whether a train a bus or a minibus is used*
- *Whether it needs double transport*
- *Whether the security has been improved from what exists today to having guards at the stations, bus terminals and ranks and even on the stations as well policing along the bus and minibus routes*
- *How long it will take you to get from home to work*
- *And if you can get a seat often or if you usually have to stand.*

So here are the two ways. Please choose the one that you think is the best for you.

Would you choose

1		2
to use a minibus to the bus ; as double transport *that costs you R5.30 per trip* *that takes you 55 minutes* *and on which you usually have a seat*	*OR*	*to use only the train* *that costs you 2.10 per trip* *that takes you 65 minutes* *and on which you seldom have a seat*

Which alternative do you prefer? **1** | **2** ***(Mark which is chosen)***

This is how we will give you information about ways of getting from home to work and ask you to choose the one that is the best for you.

INSERT SP QUESTIONNAIRES

SECTION 3: STATED PREFERENCE QUESTIONS (3alt/5att)

In this section I will be giving you information about three alternative ways of getting to work and I will then ask you to choose between them. I will do this eight times.

Interviewer	Lit- 3/5 verb2	Respondent		

Here are three alternatives which one would you choose?

	Minibus to	Minibus to	Only
		Bus	Minibus
Cost	4.60		7.20
Security	Guards on stations and on trains		Guards at ranks and on route
Travel time	120		40
Seating	Seldom		Always

Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTt	TSea	BF	BC	BSec	BTt	BSea	MF	MC	MSec	MTt	TSea	Choice
	-	V	35		2	4.60	2	120	1	2	5.30	1	55	2	1	7.20	2	40	2	

Interviewer		

Lit- 3/5 verb2	Respondent			

Here are another three alternatives which 1 would you choose? | train | bus | minibus |

	Minibus to	Only	Minibus to
	Train	Bus	Minibus
Cost	4.60	2.80	9.70
Security	As is	Guards at terminals and on route	As is
Travel time	120	105	75
Seating	Seldom	Always	always

Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTt	TSea	BF	BC	BSec	BTt	BSea	MF	MC	MSec	MTt	TSea	Choice
	-	V	35		2	4.60	1	120	1	1	2.80	2	105	2	2	970	1	75	2	

Interviewer		

Lit- 3/5 verb2	Respondent	

Here are another three alternatives which 1 would you choose?

	Minibus to	Minibus to	Minibus to
	Train	Bus	Minibus
Cost	3.60		6.30
Security	Guards on stations and on trains		Guards at ranks and on route
Travel time	120		75
Seating			always

Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTt	TSea	BF	BC	BSec	BTt	BSea	MF	MC	MSec	MTt	TSea	Choice
	-	V	35		2	3.60	2	120	2	2	780	1	105	1	2	6.30	2	75	2	

Interviewer [ ] Lit- 3/5 verb2 Respondent □

Here are another three alternatives which 1 would you choose? train | bus | minibus

	Minibus to	Only	Only
	Train	Bus	Minibus
Cost	3.60	5.30	3.80
Security	As is	Guards at terminals and on route	As is
Travel time	120	55	40
Seating	Always	Seldom	Always

Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTt	TSea	BF	BC	BSec	BTt	BSea	MF	MC	MSec	MTt	TSea	Choice
	-	V	35		2	3.60	1	120	2	1	5.30	2	55	1	1	3.80	1	40	2	

a In the last question what was the travel time of the train?. □

Interviewer		

Lit+2/3 verb3	Respondent		

Here are another two alternatives which of the two would you choose? | 1 | 2 |

	1	2
	Minibus to	Minibus to
	Bus	Minibus
Cost	7.80	9.70
Security	Guards at terminals and on route	Guards at ranks and on route

Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTt	TSea	BF	BC	BSec	BTt	BSea	MF	MC	MSec	MTt	MSea	Choice
	-	V	23		0	0	0	0	0	2	7.80	2	0	0	2	9.70	2	0	0	

Interviewer		

Lit+2/3 verb3	Respondent			

Here are another two alternatives which of the two would you choose?

1	2

	1	2
	Minibus to	Only
	Bus	Minibus
Cost	5.30	3.80
Security	Guards at ranks and on route	As is

Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTt	TSea	BF	BC	BSec	BTt	BSea	MF	MC	MSec	MTt	MSea	Choice
	-	V	23		0	0	0	0	0	2	5.30	2	0	0	1	3.80	1	0	0	

Interviewer		

Lit+2/3 verb3	Respondent			

Here are another two alternatives which of the two would you choose? □ 1 □ 2

	1	2
	Minibus to	Only
	Bus	Minibus
Cost	5.30	7.20
Security	As is	Guards at ranks and on route

Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTt	TSea	BF	BC	BSec	BTt	BSea	MF	MC	MSec	MTt	MSea	Choice
	-	V	23		0	0	0	0	0	2	5,30	1	0	0	1	7.20	2	0	0	

Interviewer		

Lit+2/3 verb3	Respondent		

Here are another two alternatives which of the two would you choose? | 1 | 2 |

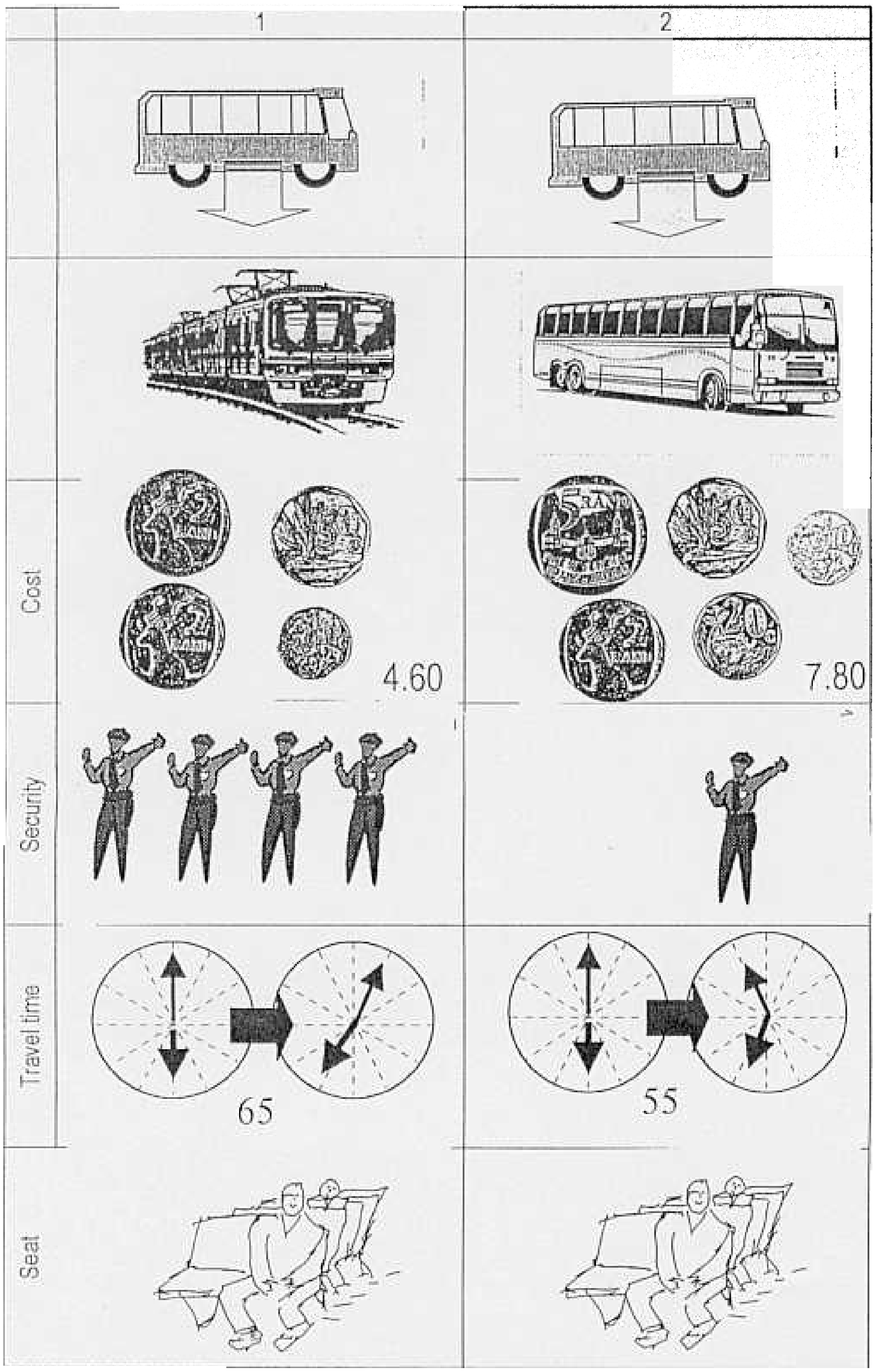
	1	2
	Only	Only
	Bus	Minibus
Cost	2.80	7.20
Security	As is	Guards at ranks and on route

Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTt	TSea	BF	BC	BSec	BTt	BSea	MF	MC	MSec	MTt	MSea	Choice
	-	V	23		0	0	0	0	0	1	2.80	1	0	0	1	7.20	2	0	0	

a In the last question what was the cost of the trip for the second alternative? □

Interviewer		Lit-2/5 pict4	Respondent			

Here are two alternatives which of the two would you choos 1 | 2

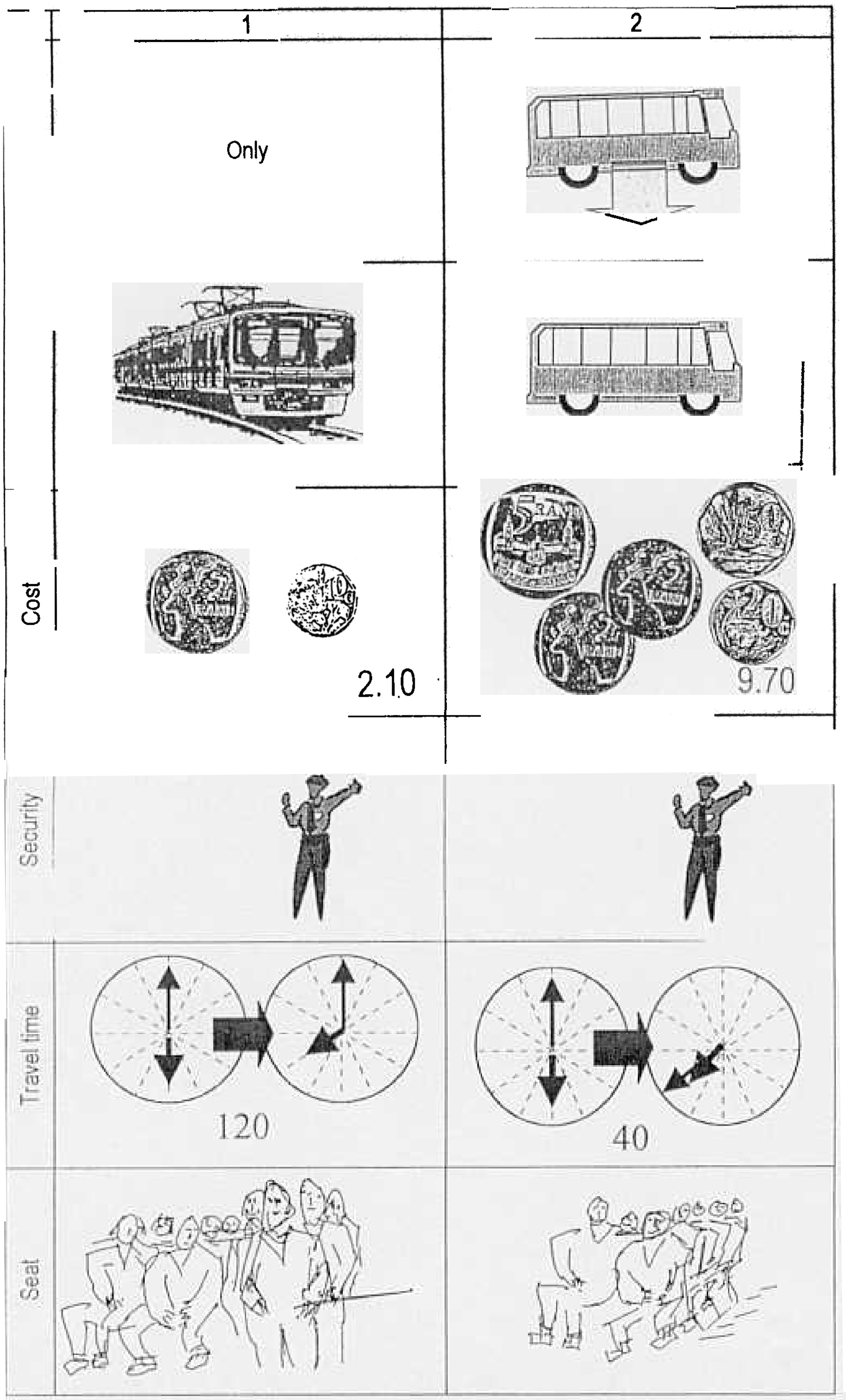


Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTI	TSea	BF	BC	BSec	BTI	BSea	MF	MC	MSec	MTI	TSea	Choice
	1	P2	25		2	4.60	2	65	2	2	7.80	1	55	2	0	0	0	0	0	

Interviewer

Lit-2/5 pict5 Respondent

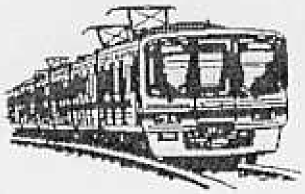
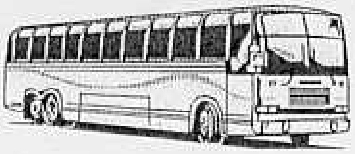
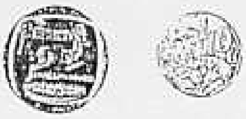
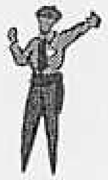
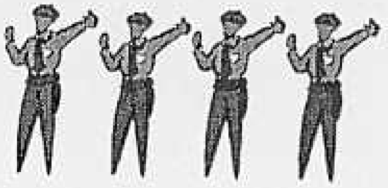
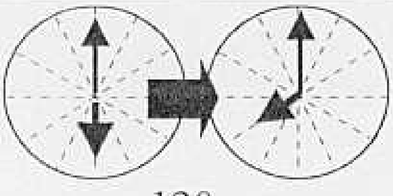
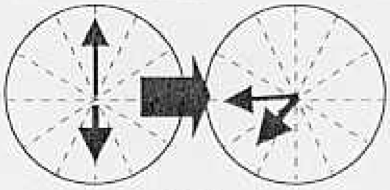
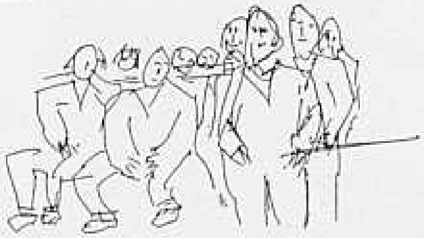
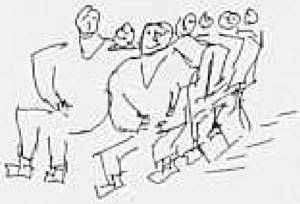
... Here are another two alternatives which of the two would you choose? | 1 | 2 |



Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTI	TSea	BF	BC	BSec	BTI	BSea	MF	MC	MSec	MTI	TSea	Choice
	1	FZ	25		1	2.10	1	120	1	0	0	0	0	0	2	9.70	1	40	1	

Interviewer	□	Lit-2/5 pict4	Respondent	□	□

Here are another two alternatives which of the two would you choose? | 1 | 2 |

	1	2
	Only	Only
		
Cost	 1.10	 5.30
Security		
Travel time	 120	 105
Seat		

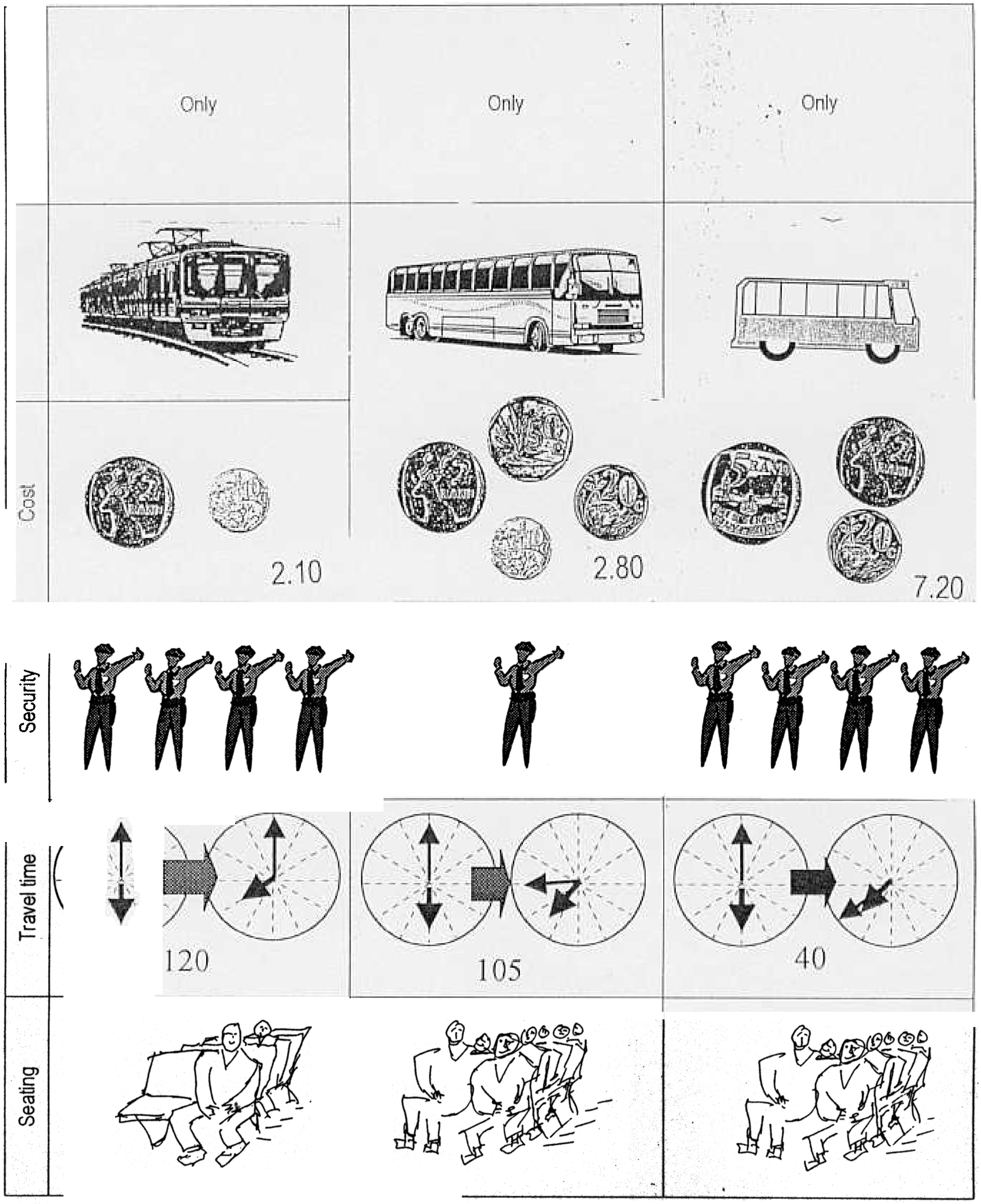
Int #	Educ	Pros	Des	Q #	TF	TC	TSec	TTt	TSea	BF	BC	BSec	BTt	BSea	MF	MC	MSec	MTt	TSea	Choice
	1	P2	25		1	1.10	1	120	1	1	5.30	2	105	1	0	0	0	0	0	

Interviewer		

Lit- 3/5 pict2	Respondent	

Here are another three alternatives which 1 would you choose?

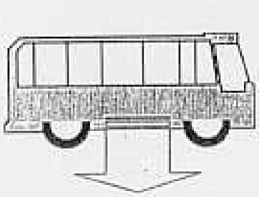
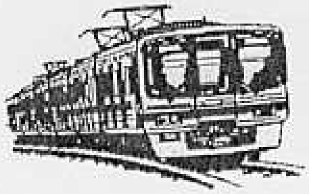
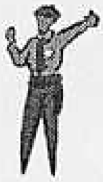
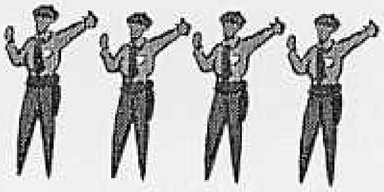
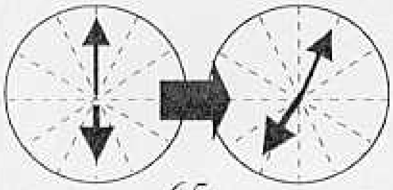
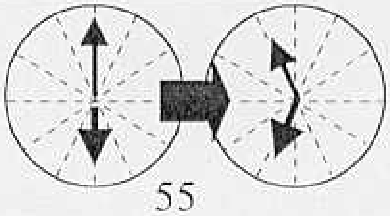
train	bus	minibus



Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTI	TSea	BF	BC	BSec	BTI	BSea	MF	MC	MSec	MTI	MSea	Choice
	-/	B	35		1	2.10	2	120	2	1	2.80	1	105	1	1	7.20	2	40	1	

Interviewer	□	Lit-2/5 pict4	Respondent	□

Here are another two alternatives which of the two would you choose? | 1 | 2 |

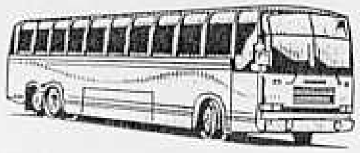
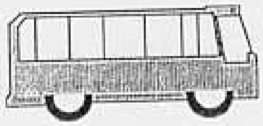
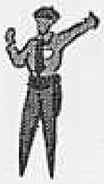
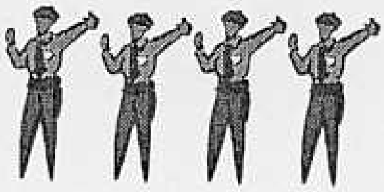
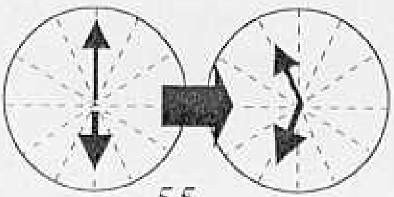
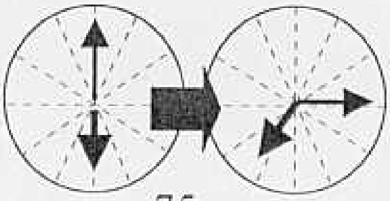
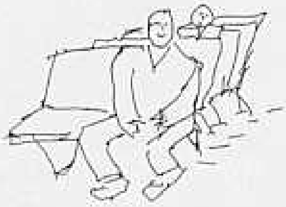
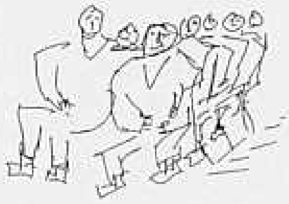
	1	2
		Only
		
Cost	4.60	2.80
Security		
Travel time	 65	 55
Seat		

Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTI	TSea	BF	BC	BSec	BTI	BSea	MF	MC	MSec	MTI	TSea	Choice
	1	2	25		2	4.60	1	65	2	1	2.80	2	55	2	0	0	0	0	0	

Interviewer

Lit-2/5 pict6	Respondent

Here are another two alternatives which of the two would you choose? □ 1 □ 2

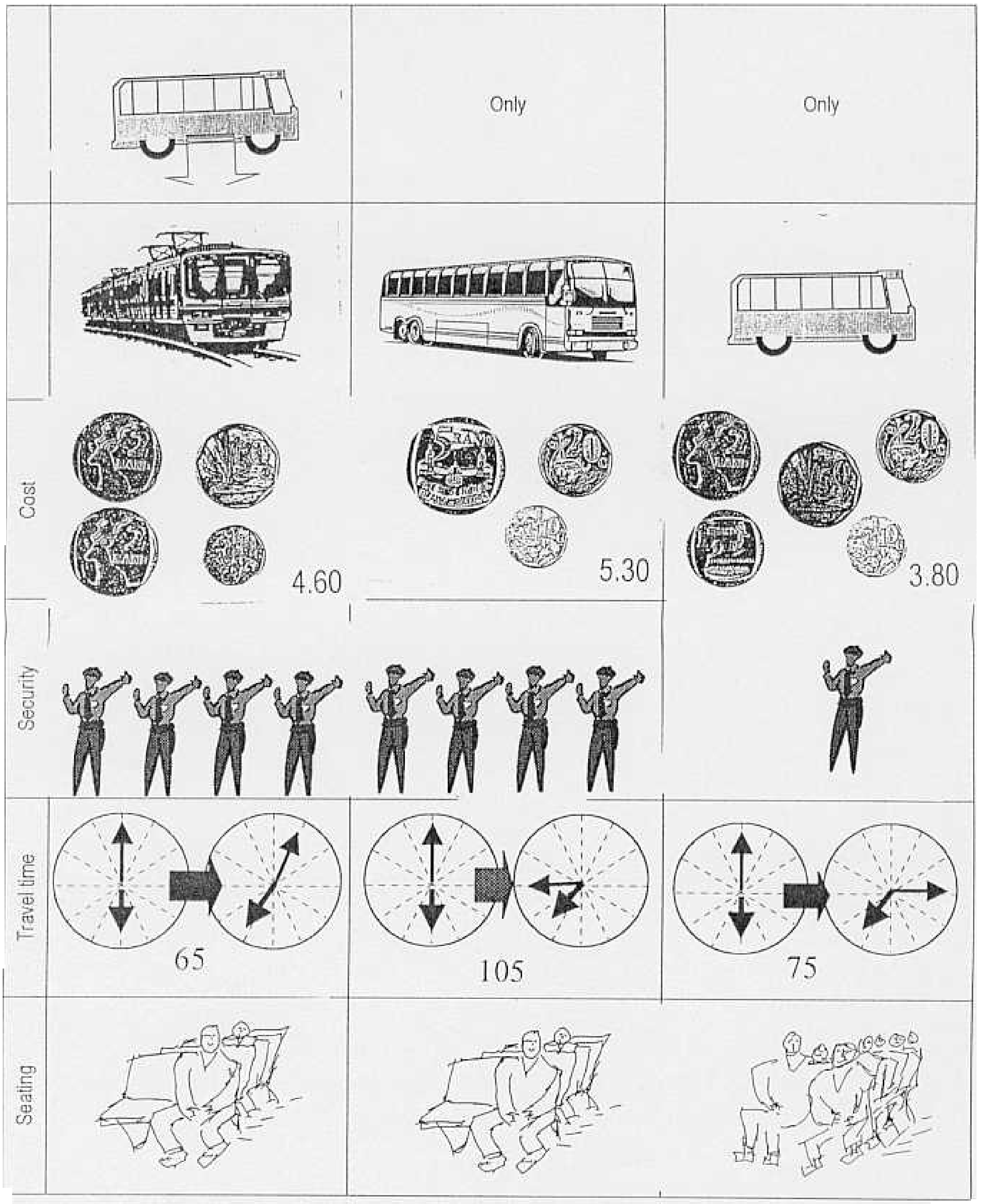
	1	2
	Only	Only
		
Cost	2.80	7.20
Security		
Travel time	 55	 75
Seat		

Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTt	TSea	BF	BC	BSec	BTt	BSea	MF	MC	MSec	MTt	TSea	Choice
	1	2	25		0	0	0	0	0	1	2.80	1	55	2	1	7.20	2	75	1	

a In the last question what was the travel time of the first alternative? □

Interviewer		Lit- 3/5 pict2	Respondent

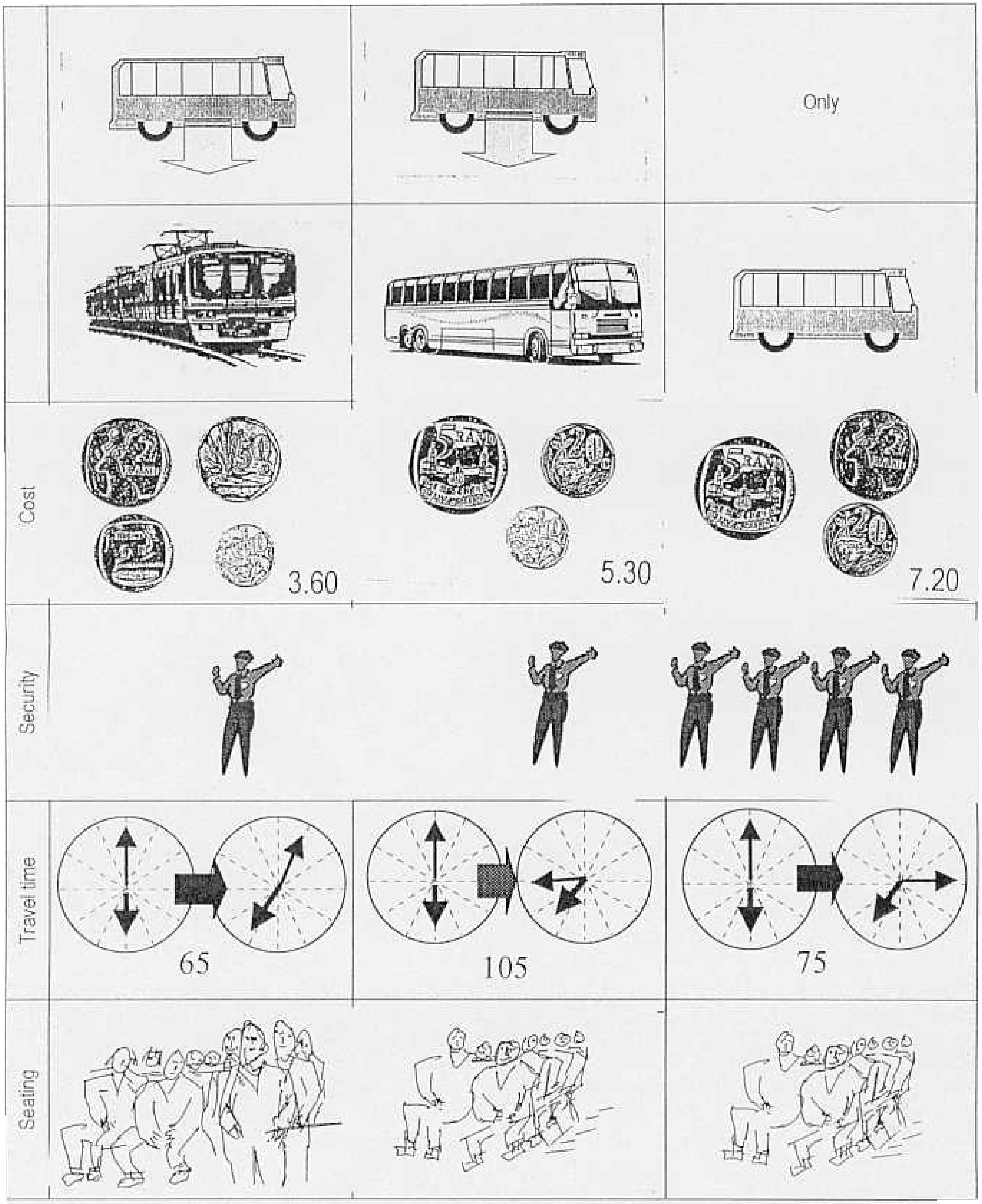
Here are three alternatives which one would you choose? | train | bus | minibus |



Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTI	TSea	BF	BC	BSec	BTI	BSea	MF	MC	MSec	MTI	MSea	Choice
	1	P2	35		2	4.60	2	65	2	1	5.30	2	105	2	1	3.80	1	75	1	

Interviewer		Lit- 3/5 pict2	Respondent			

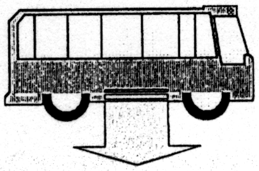
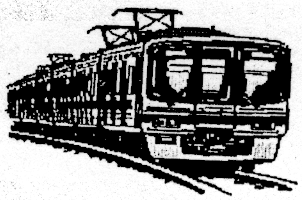
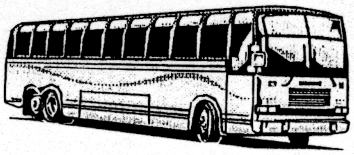
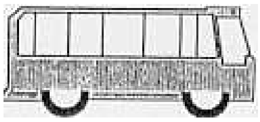
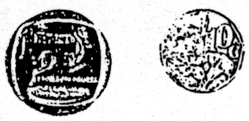
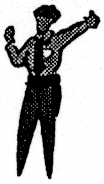
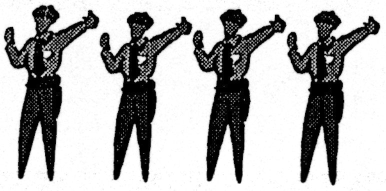
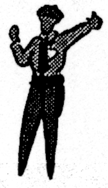
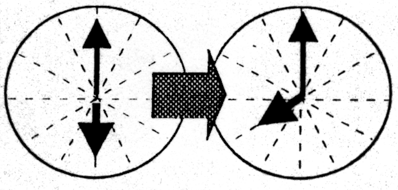
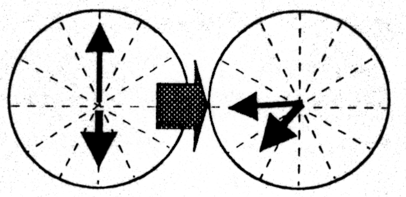
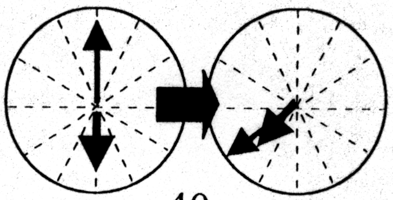
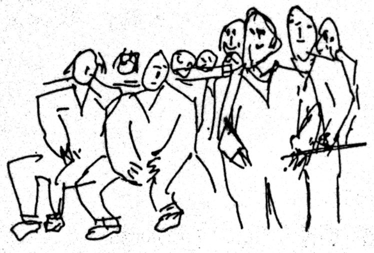
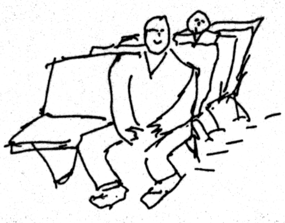
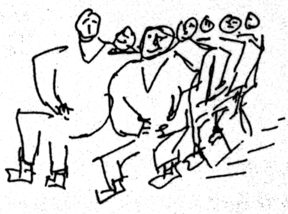
Here are another three alternatives which 1 would you choose? | train | bus | minibus |



Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TT1	TSea	BF	BC	BSec	BT1	BSea	MF	MC	MSec	MT1	MSea	Choice
	1	P2	35		2	3.60	1	65	1	2	5.30	1	105	1	1	7.20	2	75	1	

Interviewer		Lit- 3/5 pict2	Respondent	□	□	□

Here are another three alternatives which 1 would you choose? **train** | **bus** | **minibus**

	Only		Only
			
Cost	 1.10	7.80	3.80
Security			
Travel time	 120	 105	 40
Seating			

Int #	Educ	Pres	Des	Q #	TF	TC	TSec	TTt	TSea	BF	BC	BSec	BTt	BSea	MF	MC	MSec	MTt	MSea	Choice
	1	2	35		1	1.10	1	120	1	2	7.80	2	105	2	1	380	1	40	1	

.. a In the last question what was the cost of the bus alternative?............ □

Interviewer □□ Lit- Respondent □□

SECTION 4: STATED PREFERENCE FOLLOW-UP

In this section I will ask you questions about how easy or difficult you found the previous set of questions?

4.1 *Did you find the questions*

very confusing	1	a bit confusing	2	absolutely clear	3

4.2 *Did you find the questions*

Interesting	1	tiring	2	boring	3

4.3 *Which method of presentation is better?*

Verbal	1	Pictorial	2

4.4 *Why?*.. □□

.... □□

4.5 *How sure are you of the choices you have made?*

Very sure	1	Quite sure	2	Not sure at all	3

4.6 *How long have you lived in Mamelodi?* □□ *years*

6 *Where did you live before coming to Mamelodi:*

7 *How old are you?* □□ *years...*

8 *What is the highest standard that you have passed at school?:* □□

[if the reply is Std 10, ask] *what higher qualification have you obtained)*

9 *Are you*

Single	1	*Married*	2	*Divorced*	3	*Separated*	4	*Widowed*	5

10

Male	1	Female	2

11 What is the monthly income of the household? R □□□□

SECTION 5: CONTEXT TO DECISION MAKING

Give the composition of the respondent's household by indicating the number of persons in each of the categories below who sleep in the household for at least two nights per week.

	Number of persons
Husband	
Wife	
Father	
Mother	
Father-in-law	
Mother-in-law	
Older brother	
Younger brother	
Older sister	
Younger sister	
Brother-in-law	
Sister-in-law	
Married son	
Married daughter	
Unmarried son	
Unmarried daughter	
Son-in-law	
Daughter-in-law	
Father's father	
Father's mother	
Mother's father	
Mother's mother	
Grandson	
Granddaughter	
Father's older brother	
Father's younger brother	
Mother's brother	
Father's older sister	
Father's younger sister	
Mother's sister	
Cousin	
Nephew	
Niece	
Friend	

Total number of persons in household	

Interviewer			Lit-	Respondent			

I will describe four decisions to you. I will then ask you for each decision:

- *Whether you made the decision by yourself*
- *Whether you made the decision jointly with a persons that I will mention or*
- *Whether you consulted with persons that I will mention*

	Type of transport			**School children attend**			**Size of daughter's marriage goods**			**House alterations/purchases**		
	Sole	Joint	Consul	Sole	Joint	Consul	Sole	Joint	Consul	Sole	Joint	Consul
Ego												
Husband												
Wife												
Father												
Mother												
Father-in-law												
Mother-in-law												
Older brother												
Younger brother												
Older sister												
Younger sister												
Brother-in-law												
Sister-in-law												
Married son												
Married daughter												
Unmarried son												
Unmarried daughter												
Son-in-law												
Daughter-in-law												
Father's father												
Father's mother												
Mother's father												
Mother's mother												

Interviewer		

Lit-	Respondent			

	Type of transport			School children attend			Size of daughter's marriage goods			House alterations/purchases		
	Sole	Joint	Consul	Sole	Joint	Consul	Sole	Joint	Consul	Sole	Joint	Consul
Grandson												
Granddaughter												
Father's older brother												
Father's younger brother												
Mother's brother												
Father's older sister												
Father's younger sister												
Mother's sister												
Cousin												
Nephew												
Niece												
Priest/minister												
Neighbour												
Friend												

THANK YOU VERY MUCH FOR YOUR ASSISTANCE