Chapter 4

Methodology And Survey Data

4.1 Introduction

In this chapter, the methodology and survey data used in this study will be explained. Section 4.2 gives a brief explanation of the model used to evaluate the demand for intercity rail passenger services in comparison with intercity bus services, followed by the coding scheme used. The specification of the models shall be revealed in section 4.4, along with the description of methods implemented to analyse these models. Finally, this chapter will review the passenger survey conducted and the questionnaire design that were used for the purpose of the survey.

4.2 The Binary Logit Model

The models that will be used in this study are called binary logit models. It describes the relationships of independent variables to a dichotomous dependent variable. The binary logit model is popular because the probability of an event occurs is always some number between 0 to 1 due to the logistic function. Specifically, the binary logit models in this study describe the mode choice between intercity rail passenger services and bus services. It gives an insight into the variables that influence the mode choice between the two services. The probability for choosing intercity rail passenger services shall always be between 0 to 1.

The probability for choosing intercity rail passenger services is given by equation (1).

$$P_i = E(D=1 \mid X_i) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}} \quad (1)$$

$$P_i = \frac{1}{1 + e^{-z}} \quad (2)$$

where $z = \alpha + \sum \beta_i X_i$

= linear sum of α(constant) plus the sum of β_i times X_i

$i = 1,2,3,4 \ldots n$

D is a dichotomous variable which takes the value of "1" for a commuter who uses intercity rail passenger services and "0" for a commuter who uses intercity bus services. X_i represents the independent variables. The variables α and β_i represent unknown parameter that will be estimated using the data obtained about D and X's. When the parameters of α and β_i and the values of X_1 through X_n are determined, the probability of choosing intercity rail passenger services for travel can be estimated.

So the probability for not choosing intercity rail passenger services is given by equation (3) shown below.

$$1 - P_i = \frac{1}{1+e^{z}} \quad (3)$$

Therefore, the odds for choosing intercity rail passenger services are given by equation (4).

$$\frac{P_i}{1 - P_i} = \frac{1 + e^{z}}{1 + e^{-z}} = e^{z} \qquad (4)$$

where $P_i/1 - P_i$ is the odds ratio for choosing intercity rail passenger services or the ratio of probability for choosing intercity rail passenger services to the probability for not choosing intercity rail passenger services. Equation (5) is acquired by taking the natural log of equation (4). L_i is the log of the odds ratio and also known as the logit. Models like equation (5) are logit models.[28]

$$L_i = \ln\left(\frac{P_i}{1 - P_i}\right) = z = \alpha + \sum \beta_i X_i \qquad (5)$$

4.3 Coding Scheme

The explanatory variables used in this study are socioeconomic and demographic variables. These variables are categorical variables so they have to be expressed in the form of dummy variables so that the influence of each category of that variable towards the choice of mode can be evaluated. The coding scheme used in documenting the data is called the "Indicator-Variable Coding Scheme". The

[28] Damodar N.Gujarati, Basic Econometrics, 3rd ed (Singapore: Mc Graw-Hill ,Inc, 1995) 554-555

reference category is coded as "0" for every new variable created for each category of a socioeconomic and demographic variable.[29]

4.4 Specifications of Model

According to David G. Kleinbaum(1994), the variable α cannot be estimated from a case-control or cross-sectional study but the β's can be estimated. Most computer packages will provide values to all parameter involved including α. The value provided for the constant does not really estimate α.This value estimates other parameters of no real interest. The binary logit models in this study, the constant is not included in the equation. This is shown in equation (6). [30]

$$\ln\left(\frac{\widehat{Pi}}{1-Pi}\right) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots \beta_i X_i + \varepsilon \qquad (6)$$

$i = 1,2,3,4\ldots n$

There are five binary logit models in this study. Each model will be regressed on different combinations of socioeconomic and demographic independent variables. From these models, we hope to find the model that will provide the most description for the demand for intercity rail passenger services. All explanatory variables used in this study are expressed in the form of dummy variables so that comparisons within

---

[29]Marija J.Norusis, SPSS Advanced Statistics 6.1 (Chicago: SPSS Inc, 1994) 12

[30] David G.Kleinbaum,Logistic Regression : A Self Learning Text (New York :Springer-Verlag , 1994) 14-15

that category can be made and it will provide an insight into the variables that influence the mode choice between intercity rail passenger services and bus services.

In the first model, the dichotomous dependent variable will be regressed on gender, race, age and marital status of respondents as explanatory variables. The first binary logit model is shown in equation (7).

$$\ln\left(\frac{\hat{Pi}}{1-Pi}\right) = \beta_1 GENDER + \beta_2 MALAY + \beta_3 CHINESE + \beta_4 INDIAN + \beta_5 ALESS20 + \beta_6 A20-30 + \beta_7 A31-40 + \beta_8 A41-50 + \beta_9 A51-60 + \beta_{10} SINGLE + \beta_{11} MARRIED + \varepsilon \quad (7)$$

As for the second binary logit model, the model uses the occupation and personal income of respondents as explanatory variables. The second binary logit model is shown in equation (8).

$$\ln\left(\frac{\hat{Pi}}{1-Pi}\right) = \beta_1 SELF + \beta_2 GOVT + \beta_3 PRIVATE + \beta_4 STUDENT + \beta_5 PENSION + \beta_6 HOUSEWIFE + \beta_7 INONE + \beta_8 ILESS500 + \beta_9 I501-1000 + \beta_{10} I1001-1500 + \beta_{11} I1501-2000 + \beta_{12} I2001-2500 + \varepsilon \quad (8)$$

While the third binary logit model uses vehicle ownership and purpose of travel as explanatory variables. The third binary logit model in shown in equation (9).

$$\ln\left(\frac{\hat{Pi}}{1-Pi}\right) = \beta_1 VHC+\beta_2 OFFICIAL+\beta_3 HOLIDAY+\beta_4 PERSONAL+\varepsilon \quad (9)$$

The fourth binary logit model is a combination of explanatory variables from the first and second binary logit model. By doing so, we'll be able to compare the consistency of the results obtained in the first and second binary logit models. The fourth binary logit model is shown in equation (10).

$$\begin{aligned}\ln\left(\frac{\hat{Pi}}{1-Pi}\right) = {} & \beta_1 GENDER + \beta_2 MALAY + \beta_3 CHINESE + \beta_4 INDIAN + \\ & \beta_5 ALESS20 + \beta_6 A20\text{-}30 + \beta_7 A31\text{-}40 + \beta_8 A41\text{-}50 + \beta_9 A51\text{-}60 + \\ & \beta_{10} SINGLE + \beta_{11} MARRIED + \beta_{12} SELF + \beta_{13} GOVT + \\ & \beta_{14} PRIVATE + \beta_{15} STUDENT + \beta_{16} PENSION + \\ & \beta_{17} HOUSEWIFE + \beta_{18} INONE + \beta_{19} ILESS500 + \beta_{20} I501\text{-}1000 + \\ & \beta_{21} I1001\text{-}1500 + \beta_{22} I1501\text{-}2000 + \beta_{23} I2001\text{-}2500 + \varepsilon \quad (10)\end{aligned}$$

The fifth binary logit model is the complete model, which includes explanatory variables from the first, second, and third binary logit models. It allows a comparison to be made between different combinations of socioeconomic and demographic variables with the complete model in terms of consistency of results obtained. The fifth binary logit model is shown in equation (11). The definition of the variables used in these models is shown in Table 4.1.

$$\ln\left(\frac{\hat{Pi}}{1-Pi}\right) = \beta_1 GENDER + \beta_2 MALAY + \beta_3 CHINESE + \beta_4 INDIAN + \beta_5 ALESS20 + \beta_6 A20\text{-}30 + \beta_7 A31\text{-}40 + \beta_8 A41\text{-}50 + \beta_9 A51\text{-}60 + \beta_{10} SINGLE + \beta_{11} MARRIED + \beta_{12} SELF + \beta_{13} GOVT + \beta_{14} PRIVATE + \beta_{15} STUDENT + \beta_{16} PENSION + \beta_{17} HOUSEWIFE + \beta_{18} INONE + \beta_{19} ILESS500 + \beta_{20} I501\text{-}1000 + \beta_{21} I1001\text{-}1500 + \beta_{22} I1501\text{-}2000 + \beta_{23} I2001\text{-}2500 + \beta_{24} VHC + \beta_{25} OFFICIAL + \beta_{26} HOLIDAY + \beta_{27} PERSONAL + \varepsilon \quad (11)$$

Table 4.1: Definition of variables

Variables	1	0
GENDER	If female	If male
MALAY	If Malay	Otherwise
CHINESE	If Chinese	Otherwise
INDIAN	If Indian	Otherwise
ALESS20	If age less than 20	Otherwise
A20-30	If age 20-30	Otherwise
A31-40	If age 31-40	Otherwise
A41-50	If age 41-50	Otherwise
A51-60	If age 51-60	Otherwise
SINGLE	If single	Otherwise
MARRIED	If married	Otherwise
SELF	If self employed	Otherwise
GOVT	If government servant	Otherwise
PRIVATE	If working in private sector	Otherwise
STUDENT	If student	Otherwise
PENSION	If pensioner	Otherwise
HOUSEWIFE	If housewife or not working	Otherwise
INONE	If no income	Otherwise
ILESS500	If income less RM500	Otherwise
I501-1000	If income RM501-1000	Otherwise
I1001-1500	If income RM1001-1500	Otherwise
I1501-2000	If income RM1501-2000	Otherwise
I2001-2500	If income RM2001-2500	Otherwise
VHC	If does not own any vehicle	If own any vehicle
OFFICIAL	If official business	Otherwise
HOLIDAY	If holiday	Otherwise
PERSONAL	If personal	Otherwise

4.5 Method of Analysis

The estimated coefficients in these binary logit models allow us to evaluate the variables that influence the mode choice between intercity rail passenger services and bus services.

4.5.1 Wald Statistic

In each binary logit model, the test of hypothesis that a coefficient is 0 is based on the Wald statistic, which has a chi square distribution. The hypothesis is shown in equation (12). When a variable has a single degree of freedom, the Wald statistic is just the square of the ratio of the coefficient to its standard error. Meanwhile for categorical variables, the Wald statistic has degrees of freedom equal to one less than the number of categories. The significance level is shown in the column labeled "Sig". [31]

$$H_o : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0 \qquad (12)$$

4.5.2 The Model Chi Square

The model chi-square tests the null hypothesis that the coefficients for all of the terms in the current model are 0 that is comparable to the overall F test for

[31] Marija J.Norusis, SPSS Advanced Statistics 6.1 (Chicago: SPSS Inc, 1994) 5

regression. The hypotheses are shown in equation (13). The model chi-square is the difference between –2 times the log of the likelihood(-2LL) for the model with only the constant and –2LL for the current model. In this study, the constant is not included in the binary logit models so the likelihood for the model without any variable is used for comparison. [32]

H_o : The coefficients for all of the terms in the current model are 0

H_a : Not all the coefficients for all of the terms in the current model are 0 (13)

4.5.3 Prediction of the Binary Logit Model

The pseudo R squared[33] is used as a suitable statistic to measure model accuracy for the five models built in this study to evaluate the demand for intercity rail passenger services. It is measured in terms of the proportion of correct predictions. It is shown in equation (14).

$$R^2_p = \frac{\text{Number of correct predictions}}{\text{Total number of observations}} \qquad (14)$$

[32] Marija J.Norusis, SPSS Advanced Statistics 6.1 (Chicago: SPSS Inc, 1994) 11

[33] G.S.Maddala, Introduction to Econometrics 2nd Edition,(New York: MacMillan Publishing Company, 1992) 334

4.6 Survey Data

The analysis of this study uses primary data. The primary data were collected from two interview surveys – intercity train passenger survey and bus passenger survey.

4.6.1 Train Passenger Survey

The train passenger survey was conducted in the months of April and May of the year 2001. The sample of train passengers were interviewed on board coaches and while they were waiting at the platform. Passengers from different classes of coaches were approached in the survey. The train passenger survey comprised train routes from KL Sentral to Padang Besar, KL Sentral to Johor Bahru and KL Sentral to Kota Bahru. All major train stations in Peninsular Malaysia were also covered in the survey. From the survey, a total of 253 train passengers were interviewed in the process.

4.6.2 Bus Passenger Survey

This survey was also conducted in the same period with the train passenger survey. The samples of bus passengers were interviewed while they were waiting for their buses to arrive at their respective terminals. Major bus stations around Peninsular Malaysia were targeted. From the survey, a total of 272 bus passengers were interviewed in the process.

4.7 Questionnaire Design

The questionnaire was designed based on the recommendations from Stopher and Metcalf (1996) in their recent work on household travel survey, which provides a basis on how to conduct a household travel survey. The questionnaire was set in a way to fulfill the objective of the survey, which was to evaluate the factors that determined the demand for rail passenger services. The factors that would be evaluated in this study were socioeconomic and demographic variables. And then we would examine their implications for public policy on intercity rail passenger services.

In both the train passenger survey and the bus passenger survey, the questionnaire were divided into two parts. In Part A, questionnaires were set to extract socioeconomic and demographic data from respondents. The data that were collected from the respondents consisted of gender, marital status, age, occupation, race, vehicle ownership and personal income.

While part B from both surveys, the questions were more specific. The first few questions asked were related to their present journey. They were required to provide information such as place of departure, their next destination, the fare of travel and the purpose of their journey. After that, respondents were required to furnish information regarding past travel trends- frequency of travel, purpose of travel, choice of mode for travel. At a later part of the questionnaires, train passengers as

well as bus passengers were required to give their opinions on the provision of intercity rail passenger services and bus services in Peninsular Malaysia. Respondents were also required to identify the pros and cons of both service providers. A reproduction of the train passenger and bus passenger interview questionnaires can be found in Appendix C and Appendix D.