**OM386 Advanced Data Analytics in Marketing**

**Assignment 5**

**Due: April 20th, 11:59pm**

**Name: Mahika Bansal (mb62835)**

**Discrete Choice Data Analysis**

In this exercise, we will apply the multinomial logistic model to individual-level discrete choice data. The goal is to learn how to format the data, apply the R package "mlogit" to fit a multinomial logistic model and interpret the results.

The setting of the exercise is about consumers' choices of shopping malls. Please download the data file “Mall\_choice\_data.csv” from Canvas. Use read.csv( ) to read the data into R as a data frame. In this dataset, each of the consumers from a same city chooses a shopping mall to visit. There are 4 different shopping malls and a consumer also has the option of electing not to visit any of them. Hence, the choice set is denoted as {1, 2, 3, 4, 0}, where 1 through 4 are the ID's of the 4 malls and 0 means not visiting any of them (often called the outside option in a choice model). The columns in the dataset are as follows.

|  |  |
| --- | --- |
| customer ID | The ID of the customer |
| mallID | ID of different malls including 0 as the outside option |
| choice | A binary dummy variable that marks which alternative in the choice set is chosen |
| discount | An index which shows the level of discounts offer at the mall; a greater number means higher discount |
| targeting | Whether a consumer receives a targeting message from the shopping mall {1 = Yes, 0 = No} |
| distance | The distance between a consumer's home to the shopping malls |
| income | The income level of the customer |
| children | children indicator {1 = has children, 0 = no children} |

1). Please use the mlogit.data( ) function in the "mlogit" package to format the data so that it can be used by the mlogit( ) function. Please copy and paste your mlogit.data(...) statement here. Please also copy and paste of the first 20 rows of the resulting data frame formatted by mlogit.data( ) here.

Ans.

data = read.csv("Mall\_choice\_data.csv", header=T)

model = mlogit.data(data, shape="long",choice="choice", alt.levels=c("1","2","3","4","0"))

head(model,n=20)

Table

Description automatically generated

2).We let the utility of visiting mall *j* in or not visiting in {"1", "2", "3", "4", "0"} be

*Uijt = β0j + β1×discount + β2×targetig + β3×distance + β4j×income + β5j×children+εijt*

if *j* = 1, 2, 3, or 4, and

*Uijt = 0 + εijt* if *j* = 0

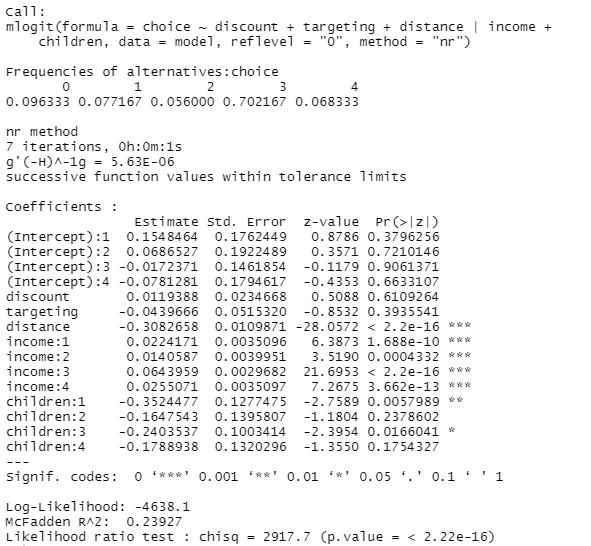
Here, *i* is the index for consumers, *t* is the index for weeks and *εijt* is assumed to have the Type-1 extreme value distribution.

Please use the appropriate statements in mlogit( ) to estimate the parameters in discrete choice model described above, using the choice "0" (not visiting) as the reference level. Copy and paste your mlogit( ) statement and the results of the regression (using summary( )) here. Please check the estimates of *β*0j, *β*1, *β*2*, β*3, *β*4j, *β*5j. Are they statistically significant? What are the interpretations of these parameters?

Ans.

model2 = mlogit(choice ~ discount + targeting + distance|income + children, model, reflevel="0")

summary(model2)



The coefficients represent the intercepts and change in utility of visiting mall wrt change in discount, targeting, distance, income & children. Based on p-values, distance, income and children (1&3) seem significant. Out of this, distance and children affect it negatively, and income positively.

**Market Share Data Analysis Based on Discrete Choice**

In this exercise, we will estimate the effects of certain characteristics of 11 different carbonated soft drinks on consumers' choices of them. Instead of using individual consumer's choice data, we will use the market share data of these soft drinks only. The data file is “Soda\_choice\_data.csv” on Canvas. The market shares of the 11 soft drinks are measure weekly for 52 weeks. Because a consumer can choose not to buy soft drinks, there is also a weekly market share for the "outside goods". The choice set is denoted as {"1", "2", ..., "11", "0"}, where 1 through 11 are the ID's of the 11 soft drinks and 0 represents the outside goods (choosing not to have soft drinks). These 11 soft drinks belong to 3 different brands, which are labeled as brand 1, 2, and 3 in the data. We have the following columns in the data.

|  |  |
| --- | --- |
| MarketShare | The market share of the soft drink |
| ProductID | The ID of the product; 0 means the outside goods |
| Week | The week indicator |
| Brand | The brand ID of the soft drink |
| Sugar | The level (1 to 5) of sugar content; a greater number means higher sugar level |
| Caffeine | The dummy for whether the drink contains caffeine {1=Yes, 0=No} |
| Promotion | Level of promotion/discount; a greater percentage means deeper discount |

1). Use read.csv( ) to read the data into R as a data frame and convert Brand into a factor. We will estimate the linear model



where *Sjt,j=*1*,...,*11is the market share of the *j*th soft drink and *S*0*t* is the market share of the outside good in week *t.*

Please follow the instructions in the class to create a new column of ln(Sjt/S0t) in the data set. Then, delete the rows of the outside good “0” from the data.

Ans.

data = read.csv('Soda\_choice\_data.csv', header = T)

data <- transform(data, lg\_mktshr = log(MarketShare) - log(1-MarketShare))

data <- data[!(data$ProductID==0),]

head(data)

Chart, table

Description automatically generated with medium confidence

2). Estimate the regression model in (1). Copy and paste the results (from the summary( ) function) here. Are *β*0, *β*1, *β*2*, β*3, *β*4  statistically significant? How do you interpret *β*0, *β*1, *β*2*, β*3, *β*4 ?

Ans.

Choice = lm(lg\_mktshr ~ Brand + Sugar + Caffeine + Caffeine + Promotion, data = data)

summary(Choice)

A screenshot of a computer

Description automatically generated with medium confidence

The coefficients represent the intercept and the change of the target variable, i.e. lg\_mktshr wrt brand, sugar, caffeine and promotion. Based on the p-values, brand, sugar and caffeine seem significant and promotion seems slightly significant. Out of these, brand and sugar affect it negatively and rest positively.