**OM 386 Marketing Analytics II**

**Assignment 6**

**Khyathi Balusu**

**Kb43582**

**Due: April 22, 11:59pm**

**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 500 consumers from a same city chooses a shopping mall to visit every week in 12 weeks. There are 4 different shopping malls and a consumer also has the option of choosing not to visit any of them in a week. 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 |
| mode | This represents the choice alternatives for a consumer |
| choice | A binary dummy variable that marks which alternative in the choice set is chosen |
| Week | A weekly time period indicator |
| 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 in that week {1 = Yes, 0 = No} |
| distance | The distance between a consumer's home to the shopping malls |
| income | The income level of the customer |
| gender | Gender indicator {1 = Male, 0 = Female} |

1). What is the format of this dataset for choice analysis, "long" or "wide"? Please use the corresponding statements in 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.

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

**install.packages("mlogit",dependencies = T)**

**library(mlogit)**

**View(Mallchoice)**

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

**head(Mallchoice.long)**

|  | **consumerID**  <int> | **mode**  <int> | **choice**  <int> | **week**  <int> | **discount**  <int> | **targeting**  <int> | **distance**  <dbl> | **gender**  <int> | **income**  <dbl> |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 1 | 0 | 1 | 0 | 1 | 6.26 | 1 | 80.39 |
| 2 | 1 | 2 | 0 | 1 | 0 | 0 | 3.10 | 1 | 80.39 |
| 3 | 1 | 3 | 1 | 1 | 0 | 1 | 4.00 | 1 | 80.39 |
| 4 | 1 | 4 | 0 | 1 | 0 | 0 | 10.16 | 1 | 80.39 |
| 5 | 1 | 0 | 0 | 1 | 0 | 0 | 0.00 | 1 | 80.39 |
| 6 | 1 | 1 | 0 | 2 | 0 | 0 | 6.26 | 1 | 80.39 |

6 rows

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×gender+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?

**Mall.m1 = mlogit(choice ~ discount + targeting + distance|income+gender, data = Mallchoice.long, reflevel = "0")**

**summary(Mall.m1)**

Call:

mlogit(formula = choice ~ discount + targeting + distance | income +

gender, data = Mallchoice.long, reflevel = "0", method = "nr")

Frequencies of alternatives:

0 1 2 3 4

0.096333 0.077167 0.056000 0.702167 0.068333

nr method

7 iterations, 0h:0m:1s

g'(-H)^-1g = 5.63E-06

successive function values within tolerance limits

Coefficients :

Estimate Std. Error z-value Pr(>|z|)

1:(intercept) 0.1548464 0.1762449 0.8786 0.3796256

2:(intercept) 0.0686527 0.1922489 0.3571 0.7210146

3:(intercept) -0.0172371 0.1461854 -0.1179 0.9061371

4:(intercept) -0.0781281 0.1794617 -0.4353 0.6633107

discount 0.0119388 0.0234668 0.5088 0.6109264

targeting -0.0439666 0.0515320 -0.8532 0.3935541

distance -0.3082658 0.0109871 -28.0572 < 2.2e-16 \*\*\*

1:income 0.0224171 0.0035096 6.3873 1.688e-10 \*\*\*

2:income 0.0140587 0.0039951 3.5190 0.0004332 \*\*\*

3:income 0.0643959 0.0029682 21.6953 < 2.2e-16 \*\*\*

4:income 0.0255071 0.0035097 7.2675 3.662e-13 \*\*\*

1:gender -0.3524477 0.1277475 -2.7589 0.0057989 \*\*

2:gender -0.1647543 0.1395807 -1.1804 0.2378602

3:gender -0.2403537 0.1003414 -2.3954 0.0166041 \*

4:gender -0.1788938 0.1320296 -1.3550 0.1754327

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -4638.1

McFadden R^2: 0.23927

Likelihood ratio test : chisq = 2917.7 (p.value = < 2.22e-16)

* Distance (*β*3), Income (*β*4j) and Gender (*β*51 and *β*53) are statistically significant.
* All the intercept coefficients (*β*0j), Discounts (*β*1) and Targeting (*β*2) are not significant
* *β*3 - Distance – negative sign shows as distance increases, probability of customers choosing this mall decreases
* *β*41 to *β*44 – Income – All have positive coeff, meaning higher income, more likely to visit malls.
* Mall3 has the highest positive coefficient. Mall 3 is chosen over other malls as income increases
* *β*51 and *β*53 – Gender: Women visit malls more than men

**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 use the following R code to reformat the data frame, so it can be used by the linear model function lm( ).

**soda = read.csv("Soda\_choice\_data.csv", header=T)**

**soda.ms = soda[soda$ProductID!=0,]**

**soda0 = soda$MarketShare[soda$ProductID==0]**

**soda0 = matrix(soda0, length(soda0), 11)**

**soda.ms$logMktShrRatio = log(soda.ms$MarketShare/as.vector(t(soda0)))**

|  | **MarketShare**  <dbl> | **ProductID**  <int> | **Week**  <int> | **Brand**  <fctr> | **Sugar**  <int> | **Caffeine**  <int> | **Promotion**  <dbl> | **logMktShrRatio**  <dbl> |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0.076 | 1 | 1 | 1 | 4 | 1 | 0.0 | -0.58192155 |
| 2 | 0.076 | 2 | 1 | 1 | 3 | 1 | 0.0 | -0.58192155 |
| 3 | 0.182 | 3 | 1 | 1 | 1 | 1 | 0.0 | 0.29135180 |
| 4 | 0.144 | 4 | 1 | 1 | 0 | 0 | 0.0 | 0.05715841 |
| 5 | 0.048 | 5 | 1 | 2 | 5 | 1 | 0.0 | -1.04145387 |
| 6 | 0.056 | 6 | 1 | 2 | 2 | 0 | 0.3 | -0.88730320 |

6 rows

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 ?

**soda.m1 = lm(logMktShrRatio ~ Brand + Sugar + Caffeine + Promotion, data = soda.ms)**

**summary(soda.m1)**

Call:

lm(formula = logMktShrRatio ~ Brand + Sugar + Caffeine + Promotion,

data = soda.ms)

Residuals:

Min 1Q Median 3Q Max

-0.87794 -0.16685 0.00523 0.15381 0.81263

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.176114 0.025637 -6.869 1.7e-11 \*\*\*

Brand2 -0.213095 0.024634 -8.650 < 2e-16 \*\*\*

Brand3 -1.021559 0.027662 -36.930 < 2e-16 \*\*\*

Sugar -0.200594 0.006366 -31.508 < 2e-16 \*\*\*

Caffeine 0.284706 0.023169 12.288 < 2e-16 \*\*\*

Promotion 0.157844 0.072373 2.181 0.0296 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2441 on 566 degrees of freedom

Multiple R-squared: 0.8435, Adjusted R-squared: 0.8422

F-statistic: 610.3 on 5 and 566 DF, p-value: < 2.2e-16

We notice that all the coefficients are statistically significant.

*β*0 – Intercept: Baseline market share

*β*1 – Brand: –negative value. Implies that as the brand id increases, the market share decreases

*β*2 – Sugar: –negative value. Implies that as the sugar level increases, the demand reduces.

*β*3 – Caffeine: –positive value. Implies that the higher the Caffeine level, the more likely are the customers to purchase it

*β*4 – Promotion: - positive value. Implies that higher the promotion/discount the more likely are the customer to purchase.