MKT 382 Marketing Analytics II
Assignment 4
Due: April 15, 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.

```{r}
mall=read.csv('Mall_choice_data.csv')

library(mlogit)

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

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

$$U_{ijt} = \beta_{0j} + \beta_1 \times discount + \beta_2 \times targetig + \beta_3 \times distance + \beta_{4j} \times income + \beta_{5j} \times gender + \varepsilon_{ijt}$$

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

$$U_{ijt} = 0 + \varepsilon_{ijt} \text{ if } j = 0$$

Here, i is the index for consumers, t is the index for weeks and ε_{ijt} 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?

```
```{r}
mall.m1 = mlogit(choice ~ discount + targeting + distance|income + gender,
mall.long, reflevel="0")
summary(mall.m1)
```
```

```
Call:
mlogit(formula = choice ~ discount + targeting + distance | income +
    gender, data = mall.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
```

All β_{0j}, β_1, β_2 are not statistically significant.
β_3, β_{4j} are statistically significant.
β_{51} and β_{53} are statistically significant, but β_{52} and β_{54} are not.

β_3 indicates that the longer the distance that customers are away from the mall, the less likely they are going to shop there.
β_{4j} all indicates that the higher the income, the more likely that customers are going to shop in any of these malls instead of not shopping in the malls.

β_{51} and β_{53} indicates that, compared to being a female, if the customers are male, they are less likely to shop in mall 1 and mall 3.

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

$$\ln(\frac{S_{1t}}{S_{0t}}) = \ln(S_{1t}) - \ln(S_{0t}) = \beta_0 + \beta_1 \times Brand_1 + \beta_2 \times Sugar_1 + \beta_3 \times Caffeine_1 + \beta_4 \times Promotion_{1t} + \xi_{1t}$$

$$\ln(\frac{S_{2t}}{S_{Jt}}) = \ln(S_{2t}) - \ln(S_{0t}) = \beta_0 + \beta_1 \times Brand_2 + \beta_2 \times Sugar_2 + \beta_3 \times Caffeine_2 + \beta_4 \times Promotion_{2t} + \xi_{2t}$$

M

$$\ln(\frac{S_{11t}}{S_{0t}}) = \ln(S_{11t}) - \ln(S_{0t}) = \beta_0 + \beta_1 \times Brand_{11} + \beta_2 \times Sugar_{11} + \beta_3 \times Caffeine_{11} + \beta_4 \times Promotion_{11t} + \xi_{11t}$$

where S_{jt}, j=1,...,11is the market share of the jth soft drink and S_{0t} 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)))

```{r}
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)))
```

```{r}
View(soda.ms)
soda.ms$Brand=as.factor(soda.ms$Brand)
str(soda.ms)
```

```
'data.frame':   572 obs. of  8 variables:
 $ MarketShare   : num  0.076 0.076 0.182 0.144 0.048 0.056 0.082 0.12 0.012 0.03 ...
 $ ProductID     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Week          : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Brand         : Factor w/ 3 levels "1","2","3": 1 1 1 1 2 2 2 2 3 3 ...
 $ Sugar         : int  4 3 1 0 5 2 1 0 4 2 ...
 $ Caffeine      : int  1 1 1 0 1 0 0 1 0 1 ...
 $ Promotion     : num  0 0 0 0 0 0.3 0 0.2 0 0 ...
 $ logMktShrRatio: num  -0.5819 -0.5819 0.2914 0.0572 -1.0415 ...
```

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 ?

```{r}
lm1=lm(logMktShrRatio~ Brand+Sugar+Caffeine+Promotion,data=soda.ms)
summary(lm1)
```

```
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
```

All the beta coefficients are statistically significant.

The coefficient of intercept means that the baseline of market share ratio, Brand1/Outside Brand, is negative, which also indicates that Brand 1 has smaller market than outside brand.

The coefficients of Brand 2 and Brand 3 mean that Brand 2,3 have smaller market share compared to Brand 1.

The coefficient of Sugar indicates that the higher the sugar level, the smaller the market share.

The coefficient of Caffeine indicates that the higher the caffeine level, the larger the market share.

The coefficient of Promotion indicates that the higher the promotion level, the larger the market share.