OM386 Marketing Analytics II
Assignment 4

By: Callie Gilmore (cgg756)

Due: April 14th, 11:59pm

Censored Regression

In this exercise, we will apply censored regression to the dataset "CreditCard_SOW_Data2.csv". The dataset has the following variables.

ConsumerID	ID's of the sampled consumers
SOW	The card's share of wallet in the consumer's total monthly spending
Promotion	Index of monthly promotion activity –higher index indicates more promotions
Balance	The customer's unpaid balance at the beginning of the month

1). In this data set, the share of wallet (SOW) can be 0% or 100%. Therefore, the share of wallet is considered a truncated variable at 0% (censored at 0) or 100% (censored at 1). We would like fit the following regression model

$SOW_{ij}^{*} = \beta_0 + \beta_1 \times Balance_{ij} + \beta_2 \times Promotion_{ij} + \varepsilon_{ij}$

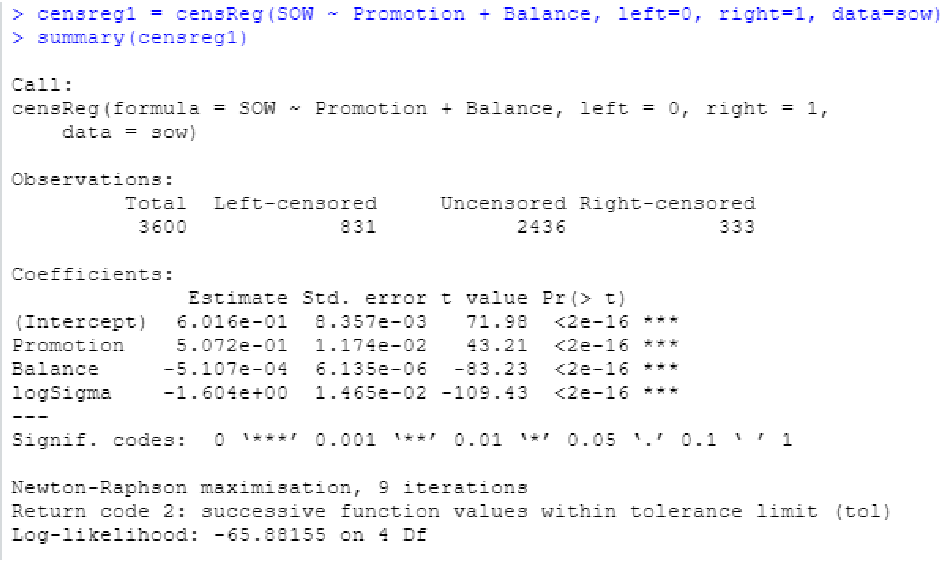
$SOW_{ij} = SOW_{ij}^{*}$ if $0 < SOW_{ij}^{*} < 1$

$SOW_{ij} = 0$ if $SOW_{ij}^{*} \leq 0$

$SOW_{ij} = 1$ if $SOW_{ij}^{*} \geq 1$

$\varepsilon_{ij} \sim N(0, \sigma^2)$ (We can reparametrize $\tau = 1/\sigma^2$, which is often named the precision)

Please use the R function censReg() in library(censReg) to fit this model by MLE. Copy and paste the summary of the results here. Interpret the parameters β_1 β_2 in the model.



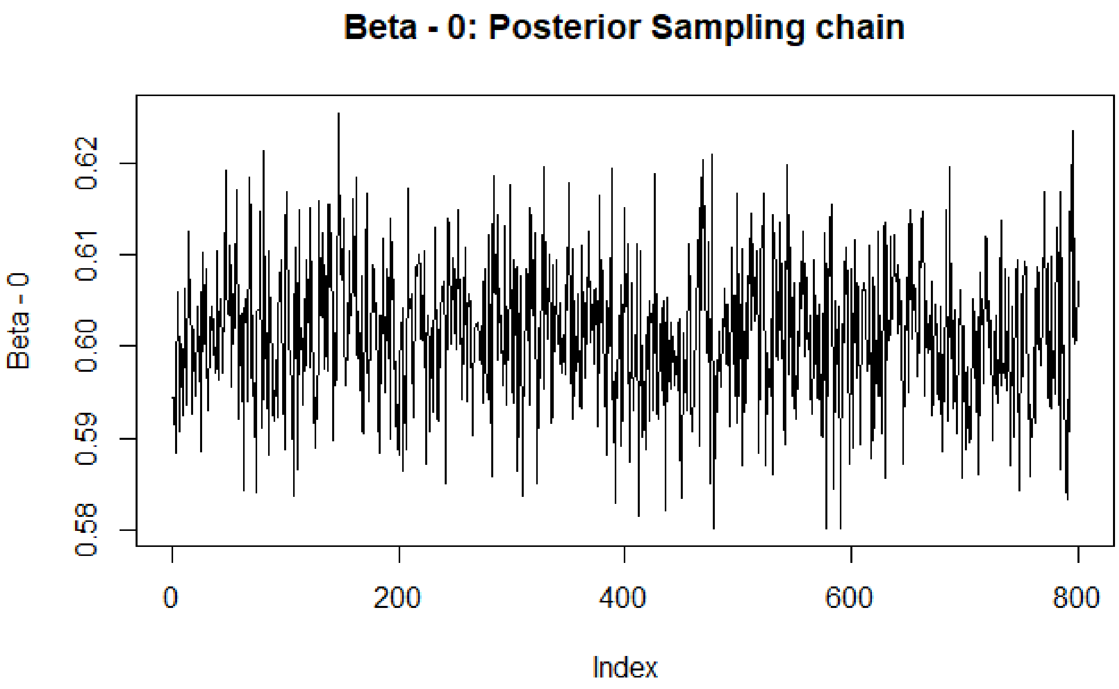
```
> censreg1 = censReg(SOW ~ Promotion + Balance, left=0, right=1, data=sow)
> summary(censreg1)

Call:
censReg(formula = SOW ~ Promotion + Balance, left = 0, right = 1,
    data = sow)

Observations:
         Total  Left-censored     Uncensored Right-censored
          3600            831           2436            333

Coefficients:
              Estimate Std. error t value Pr(> t)
(Intercept)  6.016e-01  8.357e-03   71.98  <2e-16 ***
Promotion    5.072e-01  1.174e-02   43.21  <2e-16 ***
Balance     -5.107e-04  6.135e-06  -83.23  <2e-16 ***
logSigma    -1.604e+00  1.465e-02 -109.43  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Newton-Raphson maximisation, 9 iterations
Return code 2: successive function values within tolerance limit (tol)
Log-likelihood: -65.88155 on 4 Df
```

Beta1 (Balance Coefficient) is negative which suggests that at the beginning when the customer's unpaid balance is increasing, the share of the wallet decreases. But since the Beta1 value is very close to 0, it ultimately does not have a significant impact on the share of the wallet.

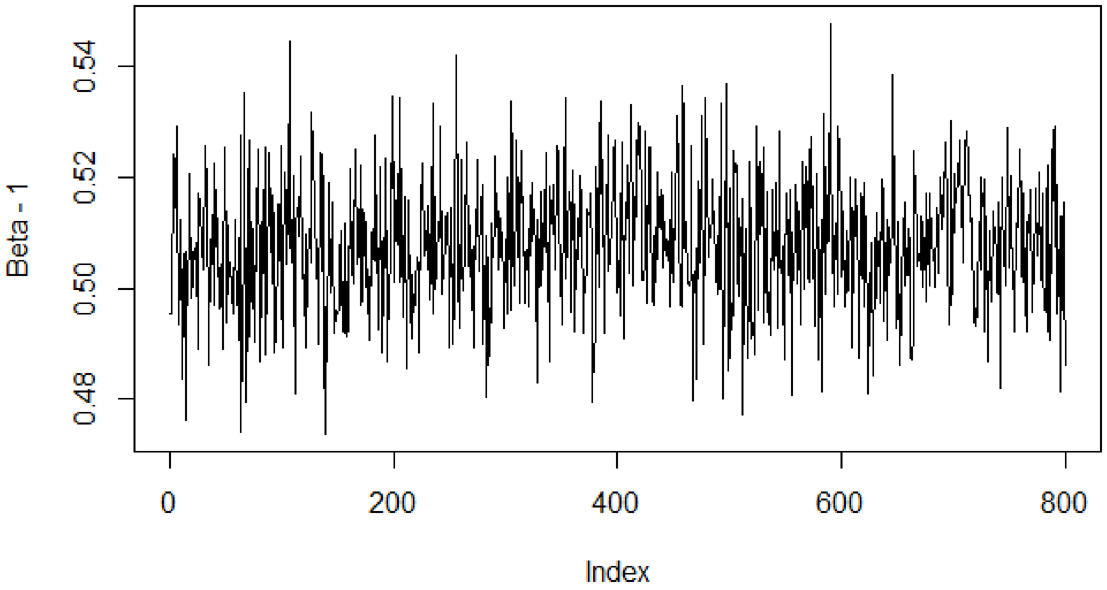
Beta2 (Promotion Coefficient) is positive which suggests that the Promotions are directly related to the share of the wallet so as the promotions increase so does the customer's share of the wallet.

2). Next, we will fit the model above using Bayesian estimation, which involves sampling the latent SOW_{ij}^* when the observed SOW_{ij} =0 or SOW_{ij} =1. The R code using MCMC (Gibbs sampling) for this inference problem is in "Assignment-4_blankcode-2021.r". Please read the code carefully and fill in the code for sampling the latent SOW_{ij}^*, the regression coefficients and the precision (inverse of the variance) of the error. You may use the rtruncnorm( ) function in the library(truncnorm) to sample from truncated normal distributions. This method is called data augmentation, which is widely applied in Bayesian statistics for missing data.

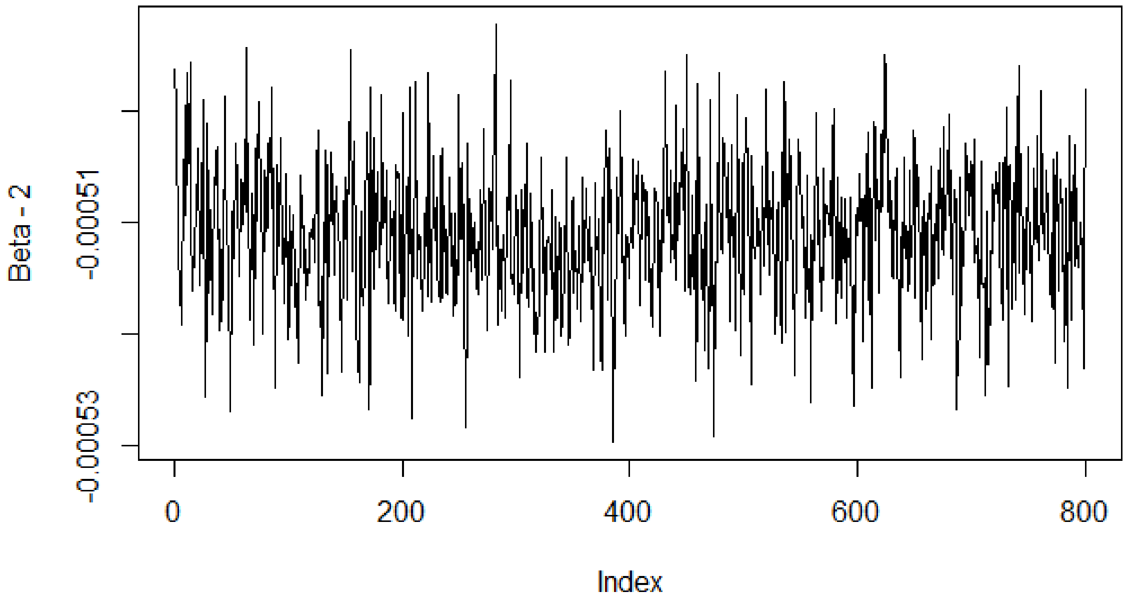
Please run the completed code. Use the plot() function to plot the posterior sampling chains and hist() to plot posterior histograms for β_0, β_1, β_2 and τ. Copy and paste the results here. Please also calculate the 95% posterior intervals for β_0, β_1, β_2 and τ. Copy and paste the results here.

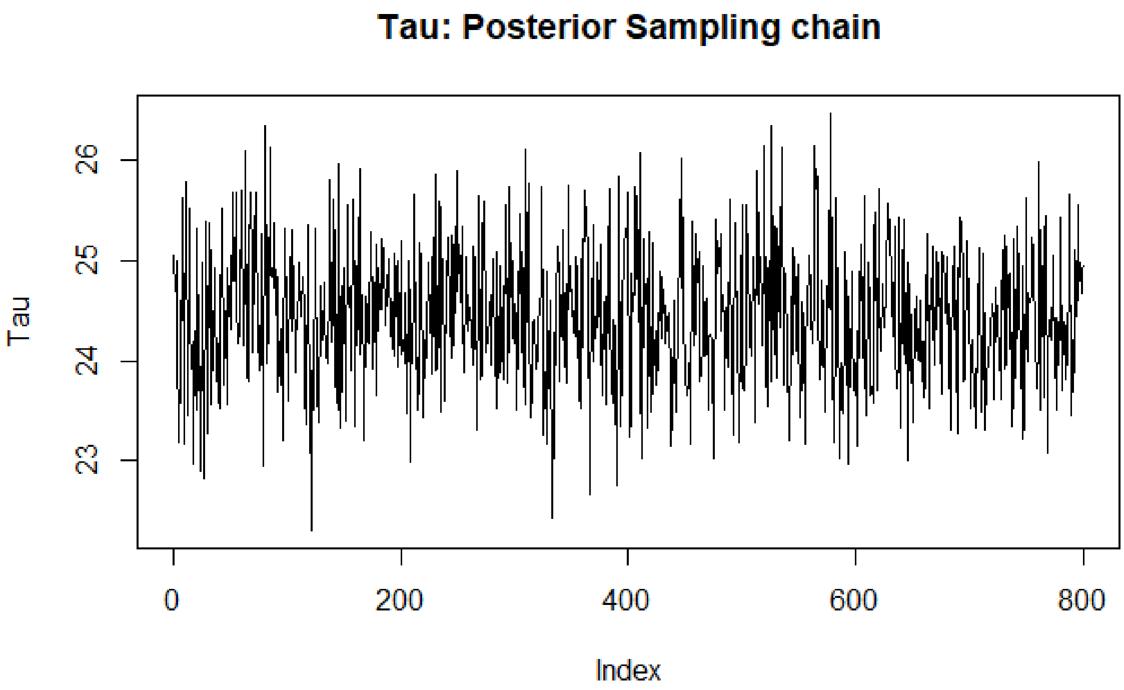


Beta - 1: Posterior Sampling chain

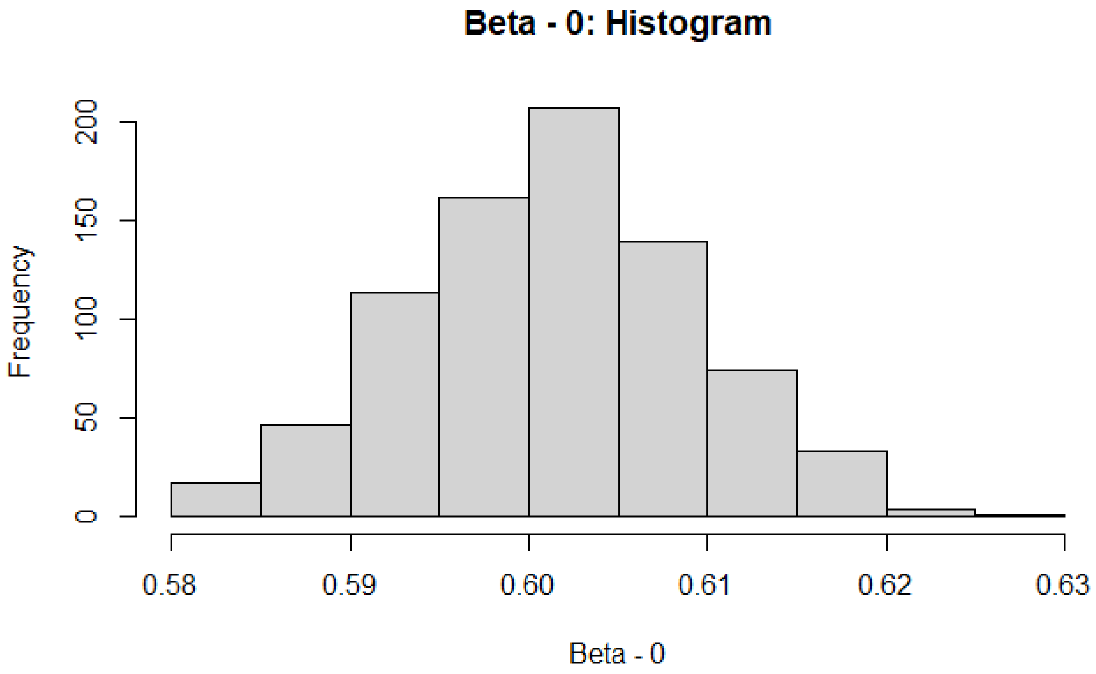


Beta - 2: Posterior Sampling chain

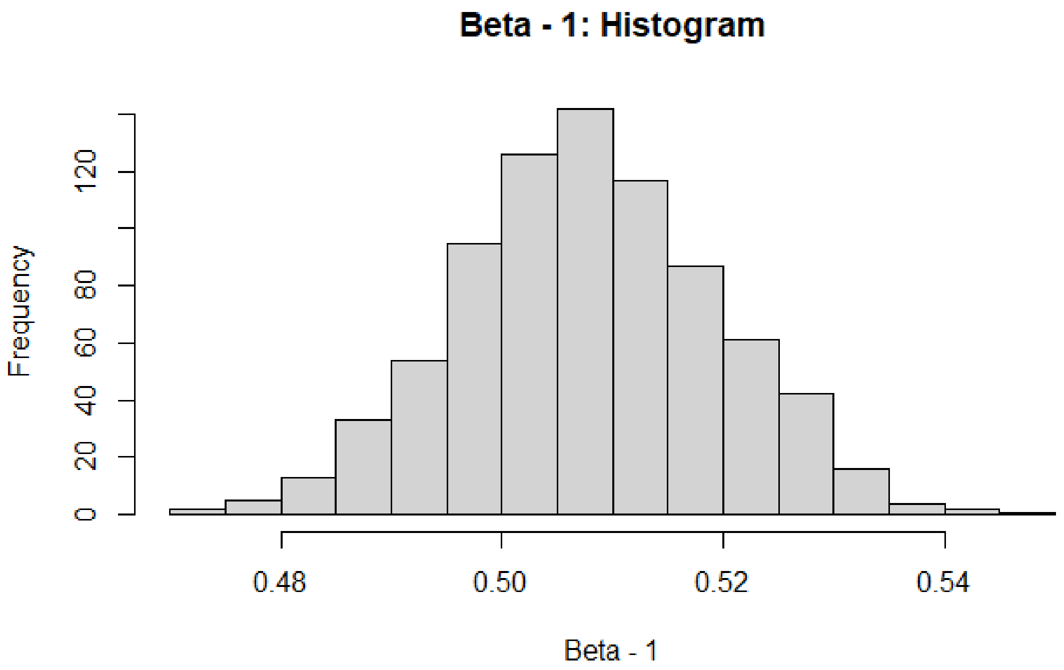


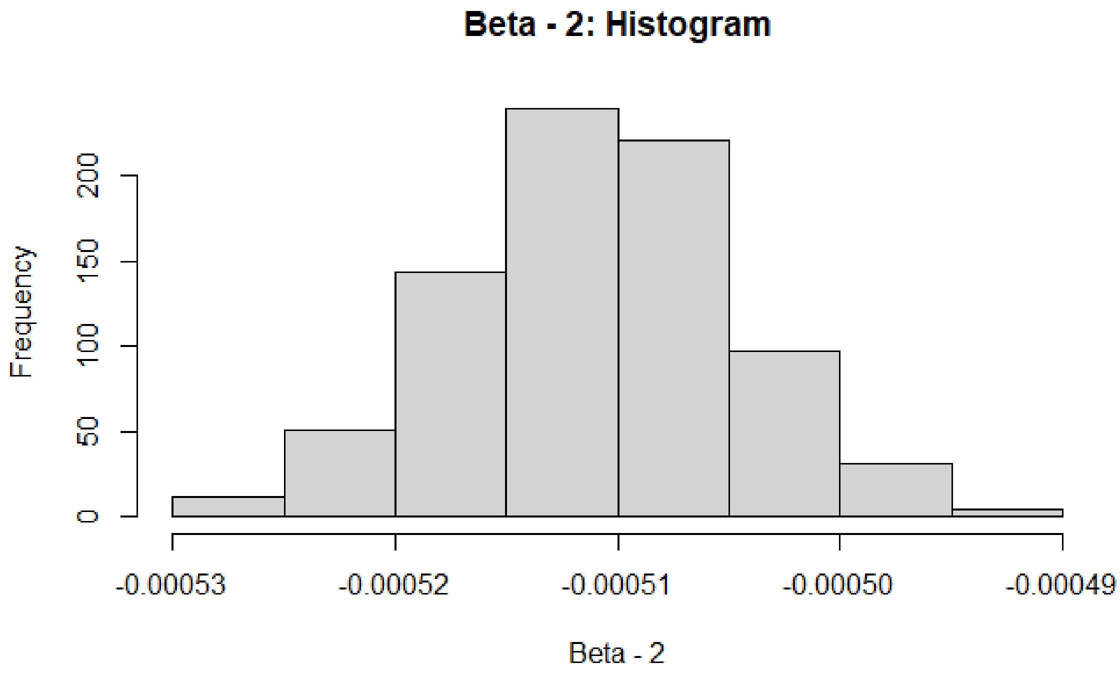


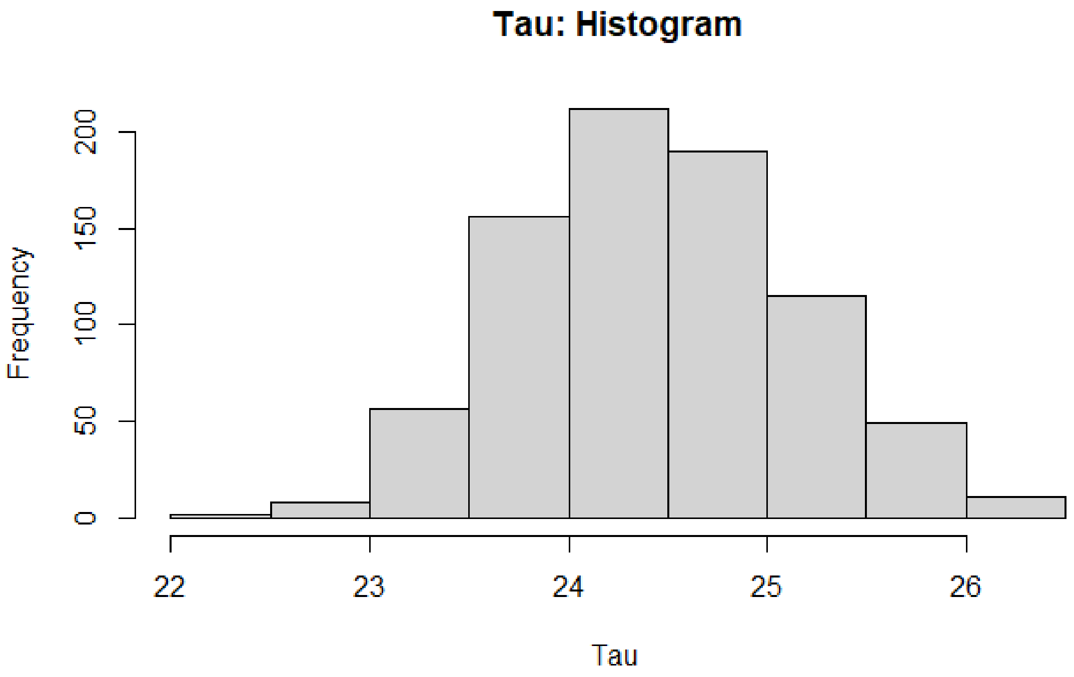
Tau: Posterior Sampling chain

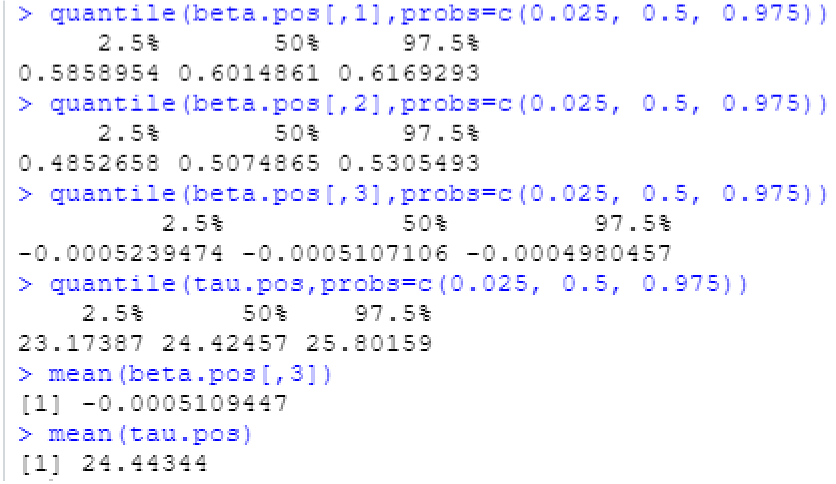


Beta - 0: Histogram


Beta - 1: Histogram
Frequency
0
20
40
60
80
120
0.48
0.50
0.52
0.54
Beta - 1


Beta - 2: Histogram
Frequency
0
50
100
150
200
-0.00053
-0.00052
-0.00051
-0.00050
-0.00049
Beta - 2





```
> quantile(beta.pos[,1],probs=c(0.025, 0.5, 0.975))
     2.5%       50%     97.5% 
0.5858954 0.6014861 0.6169293 
> quantile(beta.pos[,2],probs=c(0.025, 0.5, 0.975))
     2.5%       50%     97.5% 
0.4852658 0.5074865 0.5305493 
> quantile(beta.pos[,3],probs=c(0.025, 0.5, 0.975))
         2.5%            50%          97.5% 
-0.0005239474 -0.0005107106 -0.0004980457 
> quantile(tau.pos,probs=c(0.025, 0.5, 0.975))
    2.5%      50%    97.5% 
23.17387 24.42457 25.80159 
> mean(beta.pos[,3])
[1] -0.0005109447
> mean(tau.pos)
[1] 24.44344
```

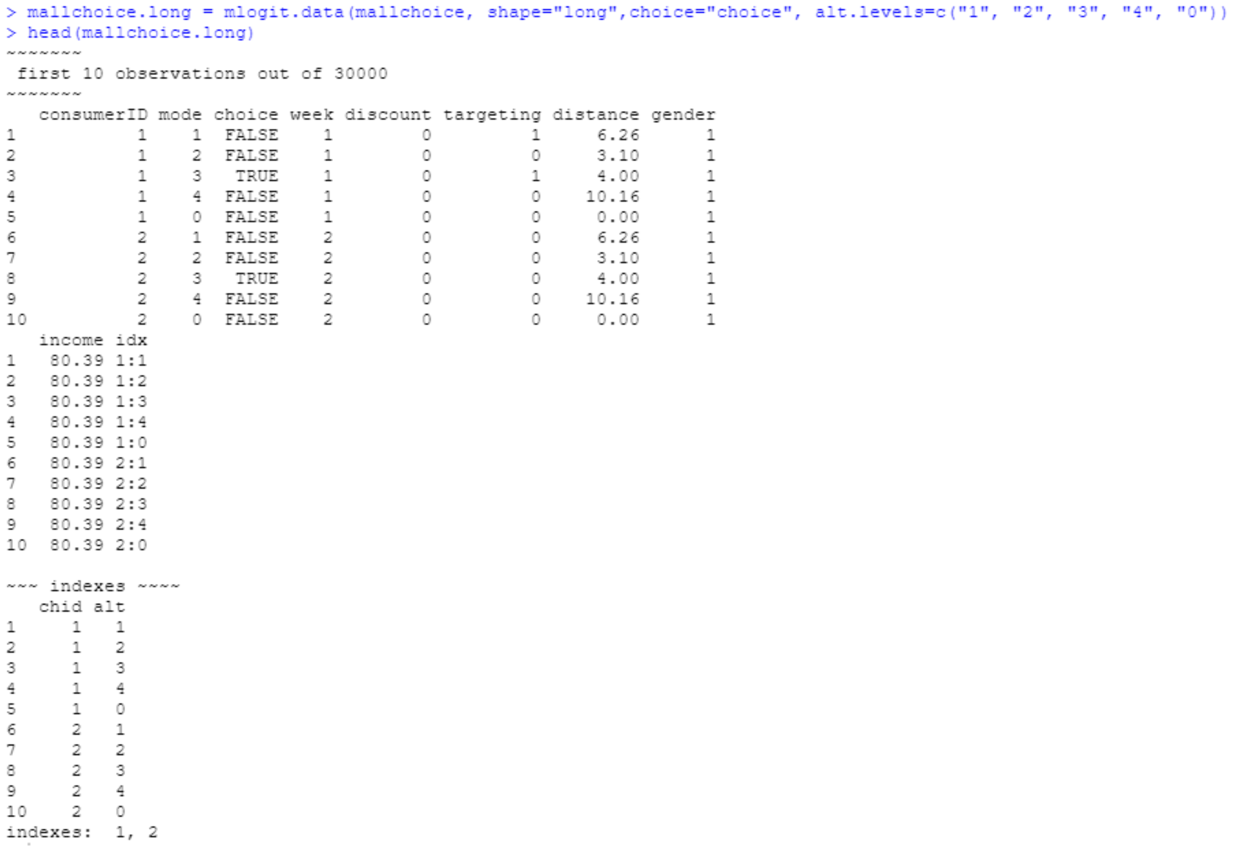
Discrete Choice Data Analysis

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 variable that indicates 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.long = mlogit.data(mallchoice, shape="long",choice="choice", alt.levels=c("1", "2", "3", "4", "0"))
> head(mallchoice.long)
~~~~~~~
 first 10 observations out of 30000
~~~~~~~
   consumerID mode choice week discount targeting distance gender
1           1    1  FALSE    1        0         1     6.26      1
2           1    2  FALSE    1        0         0     3.10      1
3           1    3   TRUE    1        0         1     4.00      1
4           1    4  FALSE    1        0         0    10.16      1
5           1    0  FALSE    1        0         0     0.00      1
6           2    1  FALSE    2        0         0     6.26      1
7           2    2  FALSE    2        0         0     3.10      1
8           2    3   TRUE    2        0         0     4.00      1
9           2    4  FALSE    2        0         0    10.16      1
10          2    0  FALSE    2        0         0     0.00      1
   income idx
1   80.39 1:1
2   80.39 1:2
3   80.39 1:3
4   80.39 1:4
5   80.39 1:0
6   80.39 2:1
7   80.39 2:2
8   80.39 2:3
9   80.39 2:4
10  80.39 2:0

~~~ indexes ~~~~
   chid alt
1     1   1
2     1   2
3     1   3
4     1   4
5     1   0
6     2   1
7     2   2
8     2   3
9     2   4
10    2   0
indexes:  1, 2
```

The dataset is long.

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

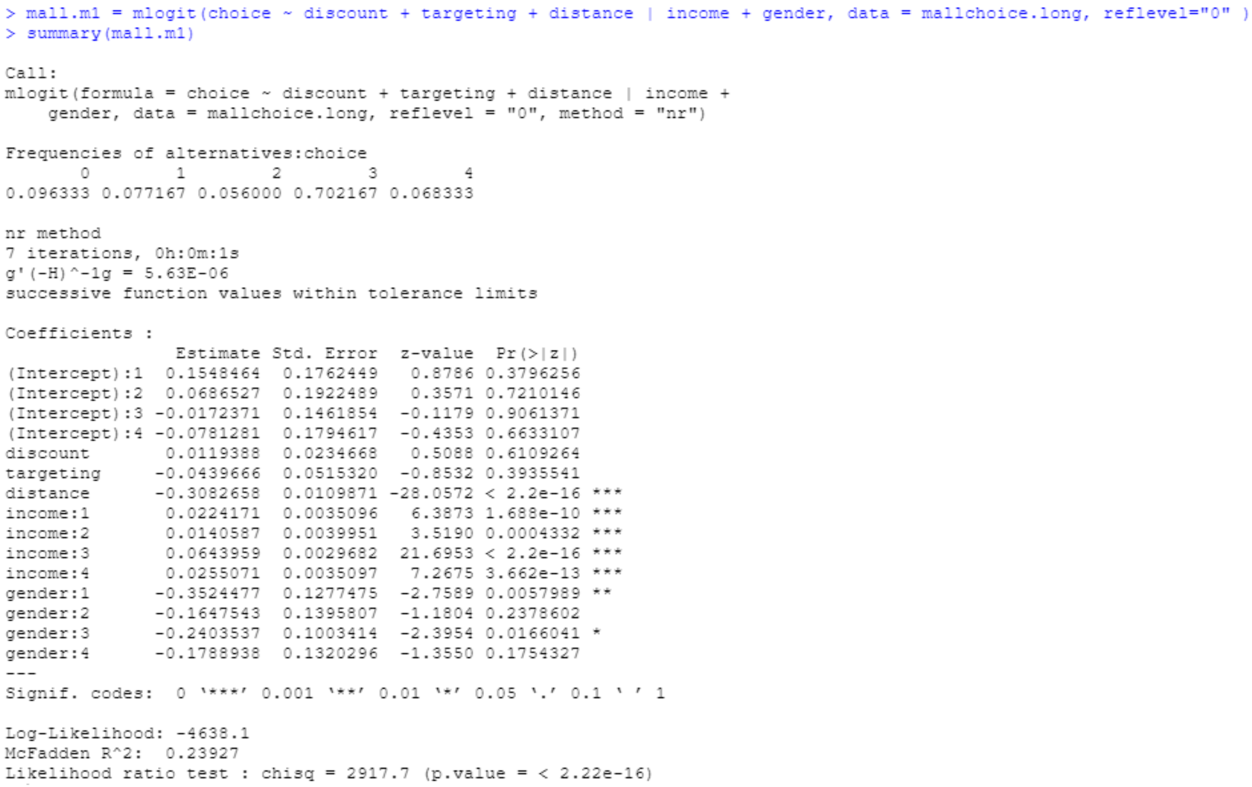
$$U_{ijt} = \beta_{0j} + \beta_1 \times discount + \beta_2 \times targetig + \beta_3 \times distance + \beta_{4j} \times income + \beta_{5j} \times gender + \varepsilon_{ijt}$$

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?



```
> 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:choice
       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|)
(Intercept):1  0.1548464  0.1762449   0.8786 0.3796256
(Intercept):2  0.0686527  0.1922489   0.3571 0.7210146
(Intercept):3 -0.0172371  0.1461854  -0.1179 0.9061371
(Intercept):4 -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 ***
income:1       0.0224171  0.0035096   6.3873 1.688e-10 ***
income:2       0.0140587  0.0039951   3.5190 0.0004332 ***
income:3       0.0643959  0.0029682  21.6953 < 2.2e-16 ***
income:4       0.0255071  0.0035097   7.2675 3.662e-13 ***
gender:1      -0.3524477  0.1277475  -2.7589 0.0057989 **
gender:2      -0.1647543  0.1395807  -1.1804 0.2378602
gender:3      -0.2403537  0.1003414  -2.3954 0.0166041 *
gender:4      -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 (Beta3), Income (Beta4j) and Gender (Beta51 and Beta53) are all statistically significant.
All the intercepts (Beta0j), Discount (Beta1) and Targeting (Beta2) are not statistically significant.
Distance (Beta3) is negative which means that as the distance increases, the probability of the customers choosing the mall decreases.
Income (Beta41-Beta44) are all positive which means that higher incomes increase the probability of visiting any mall.
Mall3 has the highest positive coefficient so with higher incomes, mall 3 is chosen more.
Gender (Beta51-Beta53) display that woman visit the 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

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

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

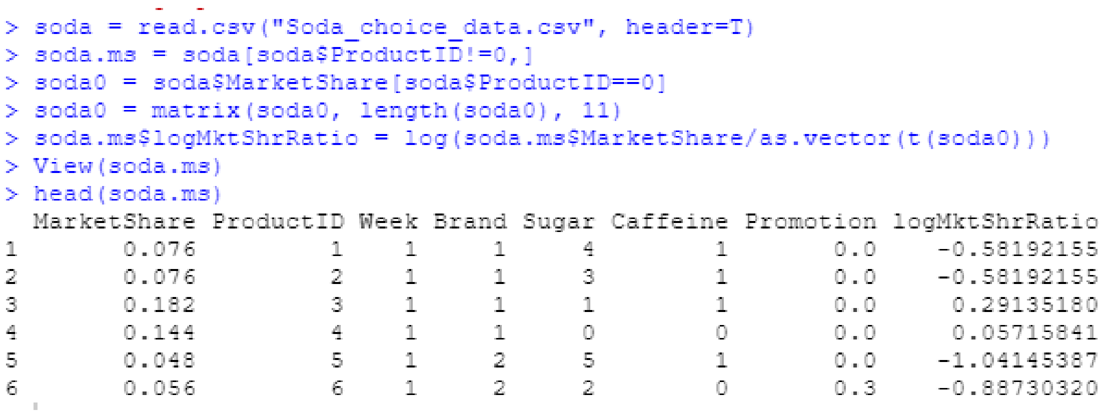
$$\vdots$$

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

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.

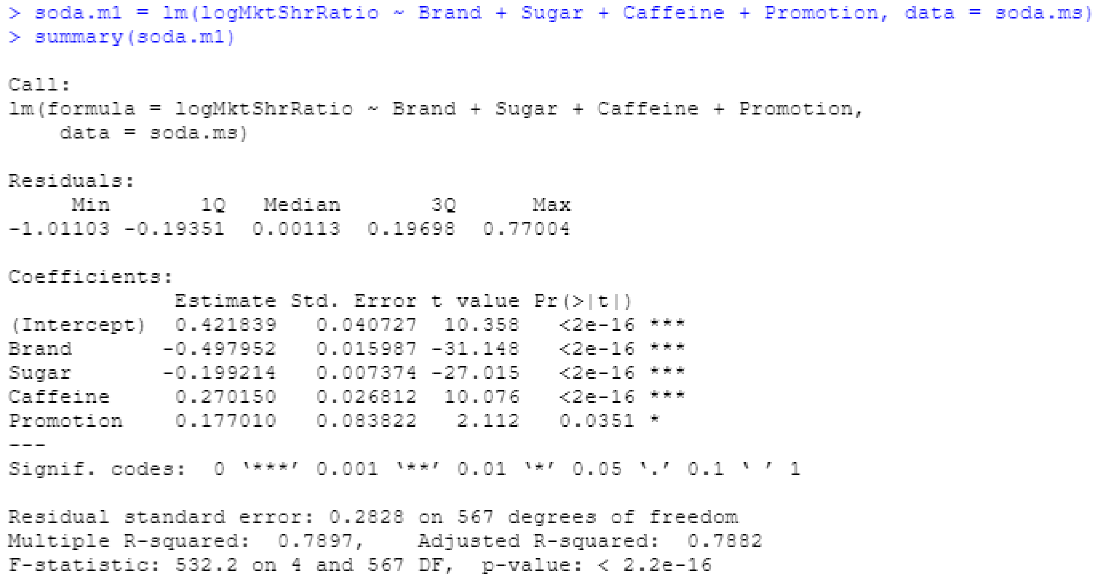
Please use the following R code to reformat the data frame, so it can be used by the linear model function lm( ).

```
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)))
```



```
> 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)))
> View(soda.ms)
> head(soda.ms)
  MarketShare ProductID Week Brand Sugar Caffeine Promotion logMktShrRatio
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
```

2). Estimate the regression model in (1). Copy and paste the results (from the summary( ) function) here. Are $\beta_0,$ β_1, β_2, β_3, β_4 statistically significant? How do you interpret $\beta_0,$ 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 
-1.01103 -0.19351  0.00113  0.19698  0.77004 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.421839   0.040727  10.358   <2e-16 ***
Brand       -0.497952   0.015987 -31.148   <2e-16 ***
Sugar       -0.199214   0.007374 -27.015   <2e-16 ***
Caffeine     0.270150   0.026812  10.076   <2e-16 ***
Promotion    0.177010   0.083822   2.112   0.0351 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2828 on 567 degrees of freedom
Multiple R-squared:  0.7897,    Adjusted R-squared:  0.7882 
F-statistic: 532.2 on 4 and 567 DF,  p-value: < 2.2e-16
```

The following coefficients are statistically significant:

Beta0 (Intercept) – baseline market share

Beta1 (Brand) – negative so as the brand id increases, the market share decreases

Beta2 (Sugar) – negative so as the sugar level increases, the less likely the customer is to purchase it

Beta3 (Caffeine) – positive so as the caffeine level increases, the more likely customers are to purchase it

Beta4 (Promotion) – positive so the higher the promotion the discount, the more likely the customers are to purchase it