R Notebook: Multinomial Logit

Packages

Make sure the following packages are installed before proceeding:

1. xtable
2. knitr
3. mlogit
4. caret
5. e1071

```
library("xtable") # processing of regression output
library("knitr") # used for report compilation and table display
library("ggplot2") # very popular plotting library ggplot2
library("ggthemes") # themes for ggplot2
suppressMessages(library("mlogit")) # multinomial logit
library("caret")
```

```
## Loading required package: lattice
```

Multinomial logit

Multinomial logit, in contrast to simple binomial logisic regression, is used for modeling choices among multiple alternatives.

Once the choice model has been estimated, we can use the parameter estimates to assess relative importance of different attributes in predicting the probability of choice.

Data

You will work with provided *trasportation_data.csv* file.

```
data <- read.csv(file = "transportation_data.csv")
```

The file contains data 210 travelers making a choice between 4 different modes of transport (plane, train, bus, car). Each traveler made a choice only once. Each alternative is a new row, so there are 4 rows per traveler – sequentially. Thus, the file contains 210 * 4 = 840 rows.

```
kable(head(data,8))
```

TRAVELER	MODE	TTME	INVC	INVT	HINC
1	0	69	59	100	35
1	0	34	31	372	35
1	0	35	25	417	35
1	1	0	10	180	35
2	0	64	58	68	30
2	0	44	31	354	30
2	0	53	25	399	30
2	1	0	11	255	30

In the table above, for example, the first four rows (1:4) form a choice set for the first traveler. Rows (5:8) are the choice set of the second traveler, and so on.

Column 1 “Traveler” is traveler’s id, column “Mode” helps identify which alternative was chosen by the traveler from the choice set – it contains 1 if the row represents the mode of transportation the traveler chose, and 0 otherwise. Per traveler, rows map to modes of transportaion in the following order:

1 - plane (air) 2 - train 3 - bus 4 - car

TTME, INVC, INVT are variables that describe the options, whereas HINC describes the traveler.

- TTME = terminal waiting time for plane, train and bus (minutes); 0 for car.

- INVC = in-vehicle cost (dollars).
- INVT = travel time (minutes).
- HINC = household income ($1000s).

All variables are treated as continuous.

Descriptive statistics

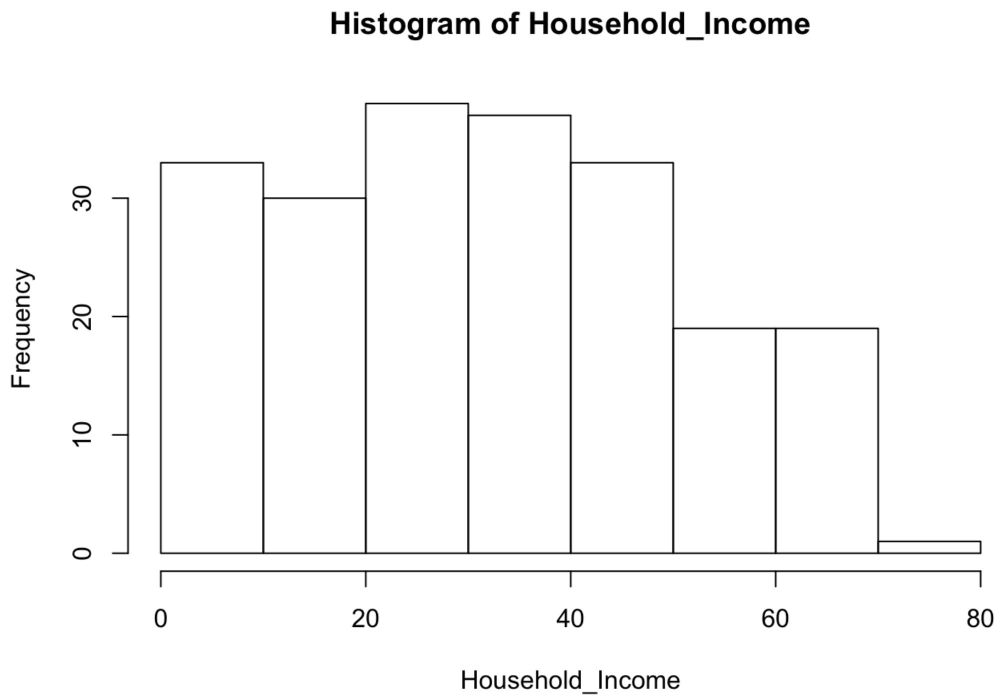
```
transp_dec<-rbind(
colSums(data[seq(1, nrow(data), 4), ])/210,
colSums(data[seq(2, nrow(data), 4), ])/210,
colSums(data[seq(3, nrow(data), 4), ])/210,
colSums(data[seq(4, nrow(data), 4), ])/210)
transp_dec<-transp_dec[,c(2:5)]
colnames(transp_dec) <- c('CHOICE SHARE','AVG. WAITING TTME', 'AVG. COST', 'AVG. TRAVEL TIME')
kable(transp_dec)
```

CHOICE SHARE	AVG. WAITING TTME	AVG. COST	AVG. TRAVEL TIME
0.2761905	61.00952	85.25238	133.7095
0.3000000	35.69048	51.33810	608.2857
0.1428571	41.65714	33.45714	629.4619
0.2809524	0.00000	20.99524	573.2048

```
Household_Income <- data[seq(1, nrow(data), 4), 6]
summary(Household_Income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    2.00   20.00   34.50   34.55   50.00   72.00
```

```
hist(Household_Income)
```



MNL model estimation - product attributes only

Based on the provided data, we will estimate the multinomial logit model.

First, we estimate a model that is trained just using 3 variables that describe the alternatives (TTME, INVC, INVT) and intercept corresponding to trasportation mode (normalizing plane intercept to zero, \(\beta_{01}=0\)). That is, \[\begin{align*} V_j = &

\beta_{0j}+\beta_{1}\text{TTME}_j + \beta_{2}\text{INVC}_j + \beta_{3}\text{INVT}_j \end{align*}\] \(U_j = V_j + \text{error}\). Assuming independent extreme value error distribution, traveler chooses mode of transportation \(j\) from the choice set of four alternatives with probability \[p_j = \frac{\exp(V_j)}{\exp(V_1)+\exp(V_2)+\exp(V_3)+\exp(V_4)},\ \ j\in\{1,2,3,4\}\]

Clearly, \(p_1+p_2+p_3+p_4=1\)

```
require('mlogit')
mdata <- mlogit.data(data=data,
                     choice='MODE', # variable that contains choice
                     shape='long', # tells mlogit how data is structured (every row is alternative)
                     varying=3:5, # only select variables that describe the alternatives
                     alt.levels = c("plane", "train", "bus", "car"), # levels of the alternatives
                     id.var='TRAVELER') # consumer id
head(mdata,6)
```

```
##          TRAVELER  MODE TTME INVC INVT HINC
## 1.plane         1 FALSE   69   59  100   35
## 1.train         1 FALSE   34   31  372   35
## 1.bus           1 FALSE   35   25  417   35
## 1.car           1  TRUE    0   10  180   35
## 2.plane         2 FALSE   64   58   68   30
## 2.train         2 FALSE   44   31  354   30
```

```
set.seed(999)
model <- mlogit(MODE~TTME+INVC+INVT,data=mdata)
summary(model)
```

```
## 
## Call:
## mlogit(formula = MODE ~ TTME + INVC + INVT, data = mdata, method = "nr", 
##     print.level = 0)
## 
## Frequencies of alternatives:
##   plane   train     bus     car 
## 0.27619 0.30000 0.14286 0.28095 
## 
## nr method
## 5 iterations, 0h:0m:0s 
## g'(-H)^-1g = 0.000192 
## successive function values within tolerance limits 
## 
## Coefficients :
##                      Estimate  Std. Error t-value  Pr(>|t|)    
## train:(intercept) -0.78666667  0.60260733 -1.3054   0.19174    
## bus:(intercept)   -1.43363372  0.68071345 -2.1061   0.03520 *  
## car:(intercept)   -4.73985647  0.86753178 -5.4636 4.665e-08 ***
## TTME              -0.09688675  0.01034202 -9.3683 < 2.2e-16 ***
## INVC              -0.01391160  0.00665133 -2.0916   0.03648 *  
## INVT              -0.00399468  0.00084915 -4.7043 2.547e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Log-Likelihood: -192.89
## McFadden R^2:  0.32024 
## Likelihood ratio test : chisq = 181.74 (p.value = < 2.22e-16)
```

Here is how likelihood ratio test is done, more explicitly.

```
model.null <- mlogit(MODE~1,data=mdata)
lrtest(model,model.null)
```

```
## Likelihood ratio test
##
## Model 1: MODE ~ TTME + INVC + INVT
## Model 2: MODE ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -192.89
## 2   3 -283.76 -3 181.74  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can also use the estimated parameters to predict the probabilities of the choice for different trasportation modes in the data. Here we print the prediction for the first traveler in the data.

```
kable(head(predict(model,mdata),1))
```

plane	train	bus	car
0.0483305	0.3255135	0.1405072	0.4856488

And now we can measure the accuracy of prediction across all data.

```
predicted_alternative <- apply(predict(model,mdata),1,which.max)
selected_alternative <- rep(1:4,210)[data$MODE>0]
confusionMatrix(predicted_alternative,selected_alternative)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2  3  4
##          1 39  6  3  7
##          2  4 49  3  8
##          3  0  1 23  0
##          4 15  7  1 44
##
## Overall Statistics
##
##                Accuracy : 0.7381
##                  95% CI : (0.6731, 0.7962)
##     No Information Rate : 0.3
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6414
##  Mcnemar's Test P-Value : 0.2118
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity            0.6724   0.7778   0.7667   0.7458
## Specificity            0.8947   0.8980   0.9944   0.8477
## Pos Pred Value         0.7091   0.7656   0.9583   0.6567
## Neg Pred Value         0.8774   0.9041   0.9624   0.8951
## Prevalence             0.2762   0.3000   0.1429   0.2810
## Detection Rate         0.1857   0.2333   0.1095   0.2095
## Detection Prevalence   0.2619   0.3048   0.1143   0.3190
## Balanced Accuracy      0.7836   0.8379   0.8806   0.7967
```

Note that if the predictions were random, the accuracy would be 25% (for four alternatives). Our simple model is doing much better than that – although it is not perfect.

Model with demographics

Now we will estimate a model that also includes a demographic variable – household income. However, we cannot just include it as an ordinary alternative-specific variable – this is because demographics for one individual would be the same across all alternatives, and so would cancel out from the probability expression as follows (so we cannot estimate the parameter \(\beta_4\))

\[\begin{align*} p_{bus} &= \frac{\exp(\cdots_{bus} + \beta_4 \text{HINC})}{\exp(\cdots_{car} + \beta_4 \text{HINC}) + \cdots + \exp(\cdots_{plane} + \beta_4 \text{HINC})}\\ &= \frac{\exp(\cdots_{bus})\exp(\beta_4 \text{HINC})}{\exp(\cdots_{car})\exp(\beta_4 \text{HINC}) + \cdots + \exp(\cdots_{plane})\exp(\beta_4 \text{HINC})}\\ &= \frac{\exp(\cdots_{bus})}{\exp(\cdots_{car}) + \cdots + \exp(\cdots_{plane})} \end{align*}\]

To deal with this issue, we need to interact the demographic variable with a dummy code for each alternative and then estimate the model. Specifically, we are now estimating utility equation where

\[\begin{align*} V_j = & \alpha_{0j} + \alpha_{1j}HouseholdIncome +\beta_{1}\text{TTME}_j + \beta_{2}\text{INVC}_j + \beta_{3}\text{INVT}_j \end{align*}\]
with intercept terms for air normalized to zero: \(\alpha_{01}=\alpha_{11}=0\). \(\alpha_{0j}\) here has the same interpretation as an intercept term in no-demographics model – that is, inherent utility of a trasportation mode relative to travel by plane. And \(\alpha_{1j}\) now measures additional (dis)utility from a trasportation mode at higher income level (again, relative to the plane).

This is how we would estimate the model

```
model1 <- mlogit(MODE~TTME+INVC+INVT|HINC,data=mdata)
summary(model1)
```

```
##
## Call:
## mlogit(formula = MODE ~ TTME + INVC + INVT | HINC, data = mdata,
##     method = "nr", print.level = 0)
##
## Frequencies of alternatives:
##   plane   train     bus     car
## 0.27619 0.30000 0.14286 0.28095
##
## nr method
## 5 iterations, 0h:0m:0s
## g'(-H)^-1g = 0.000546
## successive function values within tolerance limits
##
## Coefficients :
##                      Estimate  Std. Error t-value  Pr(>|t|)
## train:(intercept)  1.24212398  0.81686459  1.5206  0.128360
## bus:(intercept)   -0.18436561  0.89664384 -0.2056  0.837090
## car:(intercept)   -4.24742503  1.00650942 -4.2200 2.444e-05 ***
## TTME              -0.09528341  0.01035524 -9.2015 < 2.2e-16 ***
## INVC              -0.00449878  0.00721124 -0.6239  0.532722
## INVT              -0.00366471  0.00086797 -4.2222 2.420e-05 ***
## train:HINC        -0.05589505  0.01535704 -3.6397  0.000273 ***
## bus:HINC          -0.02311070  0.01645639 -1.4044  0.160212
## car:HINC           0.00210282  0.01209542  0.1739  0.861982
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -182.22
## McFadden R^2:  0.35784
## Likelihood ratio test : chisq = 203.08 (p.value = < 2.22e-16)
```

And here is how we can use likelihood ratio test to test the second model against the first one.

```
lrtest(model1,model)
```

```
## Likelihood ratio test
##
## Model 1: MODE ~ TTME + INVC + INVT | HINC
## Model 2: MODE ~ TTME + INVC + INVT
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   9 -182.22
## 2   6 -192.89 -3 21.34  8.948e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let us look at the new confusion matrix.

```
predicted_alternative <- apply(predict(model1,mdata),1,which.max)
selected_alternative <- rep(1:4,210)[data$MODE>0]
confusionMatrix(predicted_alternative,selected_alternative)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2  3  4
##          1 40  6  1  7
##          2  5 50  3  9
##          3  0  1 23  0
##          4 13  6  3 43
##
## Overall Statistics
##
##                Accuracy : 0.7429
##                  95% CI : (0.6782, 0.8005)
##     No Information Rate : 0.3
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6477
##  Mcnemar's Test P-Value : 0.2778
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity            0.6897   0.7937   0.7667   0.7288
## Specificity            0.9079   0.8844   0.9944   0.8543
## Pos Pred Value         0.7407   0.7463   0.9583   0.6615
## Neg Pred Value         0.8846   0.9091   0.9624   0.8897
## Prevalence             0.2762   0.3000   0.1429   0.2810
## Detection Rate         0.1905   0.2381   0.1095   0.2048
## Detection Prevalence   0.2571   0.3190   0.1143   0.3095
## Balanced Accuracy      0.7988   0.8390   0.8806   0.7916
```

Finally, using this model with income, we can simulate how choice share of different modes of transport will change if we reduce in-vehicle time in train by 10% (multiply it by \(0.9\)). We observe that train share increases by 5%, while bus share is most negatively affected of all modes of transport.

```
mdata.new <- mdata
mdata.new[seq(2,840,4),"INVT"] <- 0.9*mdata.new[seq(2,840,4),"INVT"]
predicted_alternative_new <- apply(predict(model1,mdata.new),1,which.max)

table(predicted_alternative)/210 # probability under original data
```

```
## predicted_alternative
##         1         2         3         4
## 0.2571429 0.3190476 0.1142857 0.3095238
```

```
table(predicted_alternative_new)/210 # probability after decrease in train travel time
```

```
## predicted_alternative_new
##         1         2         3         4
## 0.2523810 0.3380952 0.1095238 0.3000000
```

```
(table(predicted_alternative_new) - table(predicted_alternative))/table(predicted_alternative)
```

```
## predicted_alternative_new
##           1           2           3           4
## -0.01851852  0.05970149 -0.04166667 -0.03076923
```

Interaction effects

Finally, we can also interact a demographic variable with product attributes. Let us do it and see whether including corresponding terms contributes to the model's quality.

```
model2 <- mlogit(MODE~TTME+INVC+INVT+TTME:HINC+INVC:HINC+INVT:HINC|HINC,data=mdata)
summary(model2)
```

```
## 
## Call:
## mlogit(formula = MODE ~ TTME + INVC + INVT + TTME:HINC + INVC:HINC + 
##     INVT:HINC | HINC, data = mdata, method = "nr", print.level = 0)
## 
## Frequencies of alternatives:
##   plane   train     bus     car 
## 0.27619 0.30000 0.14286 0.28095 
## 
## nr method
## 5 iterations, 0h:0m:0s 
## g'(-H)^-1g = 1.24E-07 
## gradient close to zero 
## 
## Coefficients :
##                      Estimate  Std. Error t-value  Pr(>|t|)    
## train:(intercept)  2.7964e+00  1.3736e+00  2.0359 0.0417642 *  
## bus:(intercept)    1.3521e+00  1.4690e+00  0.9204 0.3573409    
## car:(intercept)   -2.2819e+00  1.7905e+00 -1.2745 0.2024902    
## TTME              -7.6820e-02  1.9786e-02 -3.8825 0.0001034 ***
## INVC              -1.2090e-02  1.6351e-02 -0.7394 0.4596505    
## INVT              -6.4867e-03  1.8918e-03 -3.4288 0.0006063 ***
## TTME:HINC         -6.1259e-04  5.6766e-04 -1.0791 0.2805255    
## INVC:HINC          2.3632e-04  4.0999e-04  0.5764 0.5643403    
## INVT:HINC          7.5168e-05  4.2332e-05  1.7757 0.0757833 .  
## train:HINC        -9.9571e-02  3.2746e-02 -3.0407 0.0023604 ** 
## bus:HINC          -6.4966e-02  3.4688e-02 -1.8729 0.0610848 .  
## car:HINC          -5.4584e-02  4.5749e-02 -1.1931 0.2328209    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Log-Likelihood: -180.1
## McFadden R^2:  0.36529 
## Likelihood ratio test : chisq = 207.31 (p.value = < 2.22e-16)
```

```
lrtest(model2,model1)
```

```
## Likelihood ratio test
## 
## Model 1: MODE ~ TTME + INVC + INVT + TTME:HINC + INVC:HINC + INVT:HINC | 
##     HINC
## Model 2: MODE ~ TTME + INVC + INVT | HINC
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  12 -180.10                     
## 2   9 -182.22 -3 4.2295     0.2377
```

We find that adding such interaction terms does not improve model significantly.