Homework 5 - Solution

Logistic regression II: Multinomial logistic regression

MSBX-5130: Customer Analytics

due: 2/25/2019

1) Setup

Homework description

- Use `mlogit()` to estimate demand for ketchup brands
- The ketchup data contains ketchup choices from consumer grocery trips
  - We observe which of 4 brands are chosen: delmonte, heinz, hunts, stb
  - We observe prices of all 4 brands for each shopping occasion
  - We also observe if the consumer resides in the southeast
  - The data are in the file `ketchup_data.csv`

The variables in the dataset are:

Variable	Description
`shopper_id`	Identifier for shopper
`trip_id`	Trip number/identifier for the shopper
`choice`	factor with levels: delmonte, heinz, hunts, stb
`price.heinz`	heinz price ($)
`price.hunts`	hunts price ($)
`price.delmonte`	delmonte price ($)
`price.stb`	stb price ($)
`southeast`	1 = if consumer resides in southeastern states

Workshop task workflow

1. Setup
   1. Download data & R Markdown files
   2. Load and describe data, tabluate market shares
   3. Transform data from wide to long format
2. Model estimation and comparison (yogurt data)
   1. Intercepts only
   2. Intercepts + `price`
   3. Intercepts + `price` + `southeast`
3. Prediction of choice probabilties
   1. Using `fitted()`
4. Marginal effects
   1. Computation of marginal effects
      1. Own regressor effect
      2. Cross regressor effect
   2. Application of marginal effects
      1. Average effect on choice probability from $0.25 increase in own `price`
      2. Average effect on choice probability from $0.25 increase in other `price`
      3. Average effect on choice probability from 1% increase in own `price`

1.1) Download data & R Markdown file

If you have not already done so, download the data file `ketchup_data.csv` (link in the Homework 5 assignment on Canvas). Also download this R markdown file, `HW5.Rmd`.

Now launch RStudio, and change the working directory to where you have downloaded the previously mentioned files.

1.2) Load and describe data, tabluate market shares

Tasks:

- First, read in the data from the csv file (`ketchup_data.csv`) to a dataframe called `DF`
- Next, use `describe()` to summarize the moments of `DF`
- Next, print the first 6 observations of `DF`
- Finally, calculate the brand choice shares (fraction of total unit purchases by brand)
  - Hint: the `table()` function can be useful for this task

```r
DF = read.csv("ketchup_data.csv")
library(psych)
describe(DF)
```

```
               vars    n   mean     sd median trimmed    mad  min     max
shopper_id        1 4956 936.72 549.10 923.00  931.67 702.01 1.00 1956.00
trip_id           2 4956   3.10   3.07   2.00    2.44   1.48 1.00   24.00
choice*           3 4956   2.62   0.90   2.00    2.59   0.00 1.00    4.00
price.heinz       4 4956   1.25   0.20   1.19    1.27   0.30 0.79    1.47
price.hunts       5 4956   1.34   0.18   1.39    1.38   0.06 0.79    1.53
price.delmonte    6 4956   1.43   0.11   1.46    1.45   0.04 0.89    1.49
price.stb         7 4956   0.92   0.07   0.95    0.93   0.06 0.75    0.99
southeast         8 4956   0.20   0.40   0.00    0.13   0.00 0.00    1.00
                 range  skew kurtosis   se
shopper_id     1955.00  0.06    -1.19 7.80
trip_id          23.00  2.49     7.73 0.04
choice*           3.00  0.39    -1.02 0.01
price.heinz       0.68 -0.53    -0.99 0.00
price.hunts       0.74 -1.82     2.59 0.00
price.delmonte    0.60 -3.44    11.93 0.00
price.stb         0.24 -1.24     0.60 0.00
southeast         1.00  1.48     0.20 0.01
```

```r
head(DF)
```

```
  shopper_id trip_id choice price.heinz price.hunts price.delmonte
1          1       1  heinz        1.19        1.39           1.49
2          1       2  heinz        0.99        1.36           1.39
3          1       3  hunts        1.46        1.43           1.49
4          1       4  hunts        1.46        1.43           1.45
5          1       5    stb        1.46        1.36           1.39
6          2       1  heinz        0.99        1.36           1.47
  price.stb southeast
1      0.89         1
2      0.95         1
3      0.99         1
4      0.99         1
5      0.95         1
```

```
6      0.95         0
```

```r
dim(DF)
```

```
[1] 4956    8
```

```r
# choice shares
table(DF$choice)/length(DF$choice)
```

```

  delmonte      heinz      hunts        stb 
0.05165456 0.50968523 0.20560936 0.23305085 
```

Discussion:

- Inspect the data and the summary statistics. How many ketchup choices do we observe? Is this dataset a panel dataset or a cross-section?

There are 4956 rows in the dataset, which corresponds to the number of observed ketchup choices (each choice is indicated by a value of the ***`choice`*** *variable, 1 choice per unique* ***`shopper_id`*** *and* ***`trip_id`*** *combination). The dataset is a panel of observations on shoppers over a sequenence of trips. Each trip involves a separate ketchup choice.*

- Which variables are alternative-specific and which are individual-specific? What column index values would select the alternative-specific varaibles?

There are 4 price variables, each associated with one of the 4 alternatives (delmonte, heinz, hunts, stb). Price is thus an alternative-specific variable. These variables are in columns 4 to 7 in the dataframe, so an index corresponding to those values would be 4:7.

- Is the source data in wide or long format? How can you tell?

The source data is in wide format. We can tell because each row corresponds to one choice outcome, and alternative specific variables appear in different columns.

- Which brand has the highest market share?

Heinz has the higest market share, with 51.0% of units purchased.

1.3) Transform data from wide to long format

Now transform the data from wide to long format, using the `mlogit.data()` function. Name the resulting dataframe `DF_long`. Finally, print the first 6 observations of `DF_long` and use `describe()` to summarize the moments of `DF_long` (you may ignore any warning messages).

Hint 1: You must first install and load (using `library()`) the mlogit package.

Hint 2: You must specify the `shape`, `choice`, `sep`, and `varying` parameters to `mlogit.data()`.

```r
library(mlogit)
DF_long = mlogit.data(DF,shape='wide',choice='choice',sep=".",varying=4:7)
head(DF_long)
```

```
           shopper_id trip_id choice southeast      alt price chid
1.delmonte          1       1  FALSE         1 delmonte  1.49    1
1.heinz             1       1   TRUE         1    heinz  1.19    1
1.hunts             1       1  FALSE         1    hunts  1.39    1
1.stb               1       1  FALSE         1      stb  0.89    1
2.delmonte          1       2  FALSE         1 delmonte  1.39    2
2.heinz             1       2   TRUE         1    heinz  0.99    2
```

```r
describe(DF_long)
```

```
Warning in describe(DF_long): NAs introduced by coercion

Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning
Inf

Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning
Inf

Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning
-Inf

Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning
-Inf
```

```
           vars     n    mean      sd  median trimmed     mad  min     max
shopper_id    1 19824  936.72  549.06  923.00  931.66  702.01 1.00 1956.00
trip_id       2 19824    3.10    3.07    2.00    2.43    1.48 1.00   24.00
choice*       3 19824     NaN      NA      NA     NaN      NA  Inf    -Inf
southeast     4 19824    0.20    0.40    0.00    0.13    0.00 0.00    1.00
alt*          5 19824     NaN      NA      NA     NaN      NA  Inf    -Inf
price         6 19824    1.24    0.24    1.39    1.25    0.15 0.75    1.53
chid          7 19824 2478.50 1430.71 2478.50 2478.50 1836.94 1.00 4956.00
             range  skew kurtosis    se
shopper_id 1955.00  0.06    -1.19  3.90
trip_id      23.00  2.49     7.73  0.02
choice*       -Inf    NA       NA    NA
southeast     1.00  1.48     0.20  0.00
alt*          -Inf    NA       NA    NA
price         0.78 -0.48    -1.40  0.00
chid       4955.00  0.00    -1.20 10.16
```

Discussion:

- In the long-form data, how many rows are there? What combination of variables in the long-form data uniquely identifies rows? How has the `choice` variable changed?

The long-form data has 19824 rows in total. Rows are uniquely identified by the combination of ***`shopper_id`****,* ***`trip_id`*** *and* ***`alt`*** *– since there is 1 choice occasion per* ***`shopper_id`*** *and* ***`trip_id`*** *combination, we can also say rows are identified by the combination of choice occasion and alternative.*

2) Model building and comparison

2.1) Intercepts only

Estimate and summarize (using `summary()`) a logit model of `choice` as the outcome. Include only intercepts (for all but one alternative). Name the result `logit1`.

```r
logit1 = mlogit(choice ~ 0 | 1 | 0, data=DF_long)
summary(logit1)
```

```

Call:
mlogit(formula = choice ~ 0 | 1 | 0, data = DF_long, method = "nr",
    print.level = 0)
```

```
Frequencies of alternatives:
delmonte    heinz    hunts      stb
0.051655 0.509685 0.205609 0.233051

nr method
5 iterations, 0h:0m:0s
g'(-H)^-1g = 4.02E-08
gradient close to zero

Coefficients :
                  Estimate Std. Error t-value  Pr(>|t|)
heinz:(intercept) 2.289215   0.065591  34.901 < 2.2e-16 ***
hunts:(intercept) 1.381400   0.069911  19.759 < 2.2e-16 ***
stb:(intercept)   1.506678   0.069080  21.811 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -5755.1
McFadden R^2:  -2.2204e-16
Likelihood ratio test : chisq = -1.819e-12 (p.value = 1)
```

Discussion:

- Which alternative is the baseline alternative?

The model estimates coefficients for heinz, hunts, and stb Since no coefficient appears for delmonte, we infer that delmonte is the baseline alternative (intercept normalized to zero). delmonte is selected by default because it is the first value when sorted alpha-numerically.

- Interpret the regression coefficients

Alternative specific intercepts are interpreted as the incremental utility from choosing the alternative, relative to the baseline category, when all other predictors/regressors are zero. We can also interpret intercepts as the incremental log-odds of choosing the alternative, relative to the baseline category, when all other predictors/regressors are zero.

What do intercepts capture? Similar to fixed effects in linear models, alternative-specific intercepts capture the average effect of factors influencing the alternative's utility that are not observed or included in the model. Since there are no observed factors (regressors) in this model, the intercepts simply capture the average utility (and log-odds), relative to the baseline alternative.

heinz intercept: on average consumers receive 2.29 more units of utility when buying heinz instead of delmonte (which has expected utility of zero). The log-odds of purchasing heinz versus delmonte is 2.29.

hunts intercept: on average consumers receive 1.38 fewer units of utility when buying hunts instead of delmonte (which has expected utility of zero). The log-odds of purchasing hunts versus delmonte is 1.38.

stb intercept: on average consumers receive 1.51 fewer units of utility when buying stb instead of delmonte (which has expected utility of zero). The log-odds of purchasing stb versus delmonte is 1.51

We note that all coefficients are highly significant using the standard t-test/p-value thresholds of 2/.05

2.2) Intercepts + price

Estimate and summarize (using `summary()`) a logit model of `choice` as the outcome. Include the following regressors: `price` (a single coefficient), plus intercepts for all but one alternative. Name the result `logit2`.

```
#TBD
logit2 = mlogit(choice ~ price | 1 | 0, data=DF_long)
summary(logit2)
```

```
Call:
mlogit(formula = choice ~ price | 1 | 0, data = DF_long, method = "nr",
    print.level = 0)

Frequencies of alternatives:
delmonte    heinz    hunts      stb
0.051655 0.509685 0.205609 0.233051

nr method
5 iterations, 0h:0m:0s
g'(-H)^-1g = 7.23E-07
gradient close to zero

Coefficients :
                   Estimate Std. Error  t-value  Pr(>|t|)
heinz:(intercept)  1.743609   0.069922  24.9365 < 2.2e-16 ***
hunts:(intercept)  1.070434   0.074099  14.4459 < 2.2e-16 ***
stb:(intercept)   -0.520243   0.086088  -6.0431 1.511e-09 ***
price             -4.321122   0.112277 -38.4861 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -4896.7
McFadden R^2:  0.14914
Likelihood ratio test : chisq = 1716.7 (p.value = < 2.22e-16)
```

Discussion:

- Interpret the price coefficient

The price coefficient estimate of -4.32 means that for any alternative, a $1 increase price decreases the alternative utility and log-odds of choice by 4.32 units, relative to any other alternative. The comparison is valid across all alternatives because the price coefficient is not alternative specific.

- Why have the intercept estimates changed?

The intercept estimates have changed, because some of the variation in outcome probabilities has been absorbed by the price coefficient. Intercepts now capture unobserved factors after accounting for product prices.

- Is this a better fitting model than the intercepts-only (2.1)? How do you know?

Yes, we have a higher (less negative) log-likelihood and a higher pseudo-R^2.

2.3) Intercepts + price + southeast

Estimate and summarize (using `summary()`) a logit model of `choice` as the outcome. Include the following regressors: `price` (single coefficient) and `southeast`, plus intercepts for all but one alternative. Be sure to place `southeast` in the appropriate portion of the `mlogit()` formula. Name the result `logit3`.

```
#TBD
logit3 = mlogit(choice ~ price | 1+southeast | 0, data=DF_long)
summary(logit3)
```

```
Call:
mlogit(formula = choice ~ price | 1 + southeast | 0, data = DF_long,
    method = "nr", print.level = 0)

Frequencies of alternatives:
delmonte    heinz    hunts      stb
0.051655 0.509685 0.205609 0.233051

nr method
7 iterations, 0h:0m:0s
g'(-H)^-1g = 1.05E-06
successive function values within tolerance limits

Coefficients :
                   Estimate Std. Error  t-value  Pr(>|t|)
heinz:(intercept)  1.732663   0.072985  23.7399 < 2.2e-16 ***
hunts:(intercept) -0.037445   0.090357  -0.4144    0.6786
stb:(intercept)   -0.799171   0.092545  -8.6355 < 2.2e-16 ***
price             -4.583948   0.124274 -36.8859 < 2.2e-16 ***
heinz:southeast   -0.162705   0.309039  -0.5265    0.5986
hunts:southeast    3.776106   0.299897  12.5914 < 2.2e-16 ***
stb:southeast      1.529874   0.299641   5.1057 3.296e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -4206.1
McFadden R^2:  0.26915
Likelihood ratio test : chisq = 3098 (p.value = < 2.22e-16)
```

Discussion:

- Interpret the regression coefficients for `southeast`.

The heinz:southeast estimate implies that southeastern residents on average receive 0.16 fewer units of utility for heinz ketchup, relative to the baseline alternative of delmonte Similarly, southeastern residents log-odds of purchasing heinz (relative to delmonte) is lower (than for non southeastern residents) by 0.16. However, this effect is not statistically significant at the 5% level.

The hunts:southeast estimate implies that southeastern residents on average receive 3.78 more units of utility for hunts ketchup, relative to the baseline alternative of delmonte Similarly, southeastern residents log-odds of purchasing hunts (relative to delmonte) is higher (than for non southeastern residents) by 3.78

The stb:southeast estimate implies that southeastern residents on average receive 1.53 more units of utility for stb ketchup, relative to the baseline alternative of delmonte Similarly, southeastern residents log-odds of purchasing stb (relative to delmonte) is higher (than for non southeastern residents) by 1.53

- Is this a better fitting model than the previous one (in 2.2)?

Yes, we have a higher (less negative) log-likelihood and a higher pseudo-R^2, which has improved to 0.269 from 0.149.

3) Prediction of choice probabilties

3.1) Using `fitted()`

Use the `fitted()` function to calculate (all) the outcome choice probabilites, using the estimates from model `logit3`. Store the fitted probabilties in a matrix called `P1`.

Print the first 6 rows of `P1`. Next, print the sample average choice probabilities (column averages).

```r
#TBD
P1 = fitted(logit3, outcome = FALSE)
head(P1)
```

```
       delmonte      heinz      hunts        stb
[1,] 0.00840349 0.15977870 0.55875034 0.27306747
[2,] 0.01053568 0.31681078 0.50823630 0.16441724
[3,] 0.01213409 0.06691986 0.67163627 0.24930977
[4,] 0.01454050 0.06675684 0.67000019 0.24870246
[5,] 0.01463433 0.05103197 0.70595378 0.22837992
[6,] 0.01708542 0.87234404 0.02724894 0.08332161
```

```r
colMeans(P1)
```

```
  delmonte      heinz      hunts        stb
0.05165456 0.50968523 0.20560936 0.23305085
```

```r
# data choice shares
table(DF$choice)/length(DF$choice)
```

```

  delmonte      heinz      hunts        stb
0.05165456 0.50968523 0.20560936 0.23305085
```

Discussion:

- How do the average sample choice probability predictions compare to observed choice probabilities (shares) in the data (that you computed in 1.2)?

The average predicted choice probabilities match the average choice probabilities in the data to the listed precision.

4) Marginal effects

4.1) Computation of marginal effects

4.1.1) Own regressor effect

Calculate the marginal effect of changing Hunt's ketchup price on Hunt's choice probability. Do this by computing the observation-level marginal effects and then averaging over all observations.

Hint 1: Recall the formula for an own-regressor marginal effect: $m.e. = \beta_{price} p_{ik}(1 - p_{ik})$, where β_{price} is the coefficient on price from model `logit3`.

Hint 2: You can use your predicted choice probability matrix to get the values of p_{ik} – but you must be sure to select the correct matrix columns.

```
#TBD
# note Hunts is associated with the 3rd column of P1 (see 3.1 output)
me_Hunts_PriceHunts = mean(P1[,3]*(1-P1[,3])*logit3$coefficients["price"])
me_Hunts_PriceHunts
```

```
[1] -0.4308148
```

Discussion:

- Interpret the marginal effect
  - Does the sign of the effect make sense?

The marginal effect estimate of -0.43 implies that a \$1 increase in price of Hunts ketchup on average decreases the probability of choosing Hunts by 0.43, or 43%. The effect sign makes sense, as we expect increasing price to decrease the probability of choice.

4.1.2) Cross regressor effect

Calculate the marginal effect of changing Heinz ketchup's price on Hunts's choice probability. Do this by computing the observation-level marginal effects and then averaging over all observations.

Hint: Recall the formula for an cross-regressor marginal effect: $m.e. = -\beta_{price} p_{ik} p_{ij}$, where β_{price} is the coefficient on price from model `logit3`.

```
#TBD
# note Hunts is associated with the 3rd column of P1 (see 3.1 output)
# note Heinz is associated with the 2nd column of P1 (see 3.1 output)
me_Hunts_PriceHeinz = mean(-P1[,3]*P1[,2]*logit3$coefficients["price"])
me_Hunts_PriceHeinz
```

```
[1] 0.2046593
```

Discussion:

- Interpret the marginal effect
  - Does the sign of the effect make sense?

The marginal effect estimate of 0.20 implies that a \$1 increase in price of Heinz ketchup on average increases the probability of choosing Hunts by 0.20, or 20%. The effect sign makes sense, as we expect increasing price of a competitor to increase the probability of choice of the focal brand.

4.2) Application of marginal effects

Here we will use the marginal effects to evaluate the impact (on choice probabilities) of various types of price changes. Remember that evaluating changes using maginal effects is an approximation of the exact change. For the latter, we would need to calcualte the difference in two predictions, one "in sample" and the other "out of sample" using altered prices.

4.2.1) Average effect on choice probability from \$0.25 increase in own price

Estimate the average effect on Hunts's choice probability from increasing Hunt's price by \$0.25.

Hints: Given prior work, only a calculator (no data manipulation) is needed

```
#TBD
me_Hunts_PriceHunts*0.25
```

```
[1] -0.1077037
```

Discussion:

- Interpret the effect

On average, increasing Hunt's price by 25 cents leads to 10.8% decrease in Hunt's choice probability.

4.2.2) Average effect on choice probability from \$0.25 increase in other `price`

Estimate the average effect on Hunt's choice probability from increasing Heinz's price by 25 cents.

Hints: Given prior work, only a calculator (no data manipulation) is needed

```r
#TBD
me_Hunts_PriceHeinz*0.25
```

```
[1] 0.05116482
```

Discussion:

- Interpret the effect

On average, increasing Heinz's price by 25 cents leads to 5.1% increase in Hunt's choice probability.

4.2.3) Average effect on choice probability from 1% increase in own `price`

Now estimate the average effect on Hunt's choice probability from increasing Hunt's price by 1%.

Hints: The trick here is to deal with a percentage change instead of a uniform unit change. One approach is to consider a 1% increase in the sample average price as the unit change in price.

```r
#TBD
eff_4.2.3 = 0.01*mean(DF$price.hunts)*me_Hunts_PriceHunts
eff_4.2.3
```

```
[1] -0.005789548
```

Discussion:

- Interpret the effect

On average, increasing Hunt's price by 1% leads to 0.6% decrease in Hunt's choice probability.