Homework 2 – Estimating a Logit Model

Jin Miao
Empirical Models in Marketing

February 21, 2018

Question 1 Likelihood Construction. The indirect utility for each individual household i to choose brand (alternative) j in a particular purchase trip (time) t is

$$U_{ijt} = V_{ijt} + \epsilon_{ijt}$$

where

$$V_{ijt} = \alpha_{ij} + \beta_1 price_{ijt}$$

ϵ_{ijt} are IID Extreme Value Distributed. The Cumulative Distrbution Function of ϵ_{ijt} is

$$F(\epsilon_{ijt}) = e^{-e^{-\epsilon_{ijt}}}$$

Consequently,

$$P(choice_{ij} = 1) = P(V_{i1t} + \epsilon_{i1t} > V_{i2t} + \epsilon_{i2t}) = P(\epsilon_{i1t} - \epsilon_{i2t} > V_{i2t} - V_{i1t})$$

where

$$V_{i2t} - V_{i1t} = \alpha_{i2} + \beta_1 price_{i2t} - (\alpha_{i1} + \beta_1 price_{i1t}) = (\alpha_{i2} - \alpha_{i1}) + \beta_1 (price_{i2t} - price_{i1t})$$

The propbability that each individual household i chooses brand (alternative) j in a particular purchase trip (time) t is

$$P(choice_{it} = j) = \frac{e^{V_{ijt}}}{\sum_{k=1}^{2} e^{V_{ikt}}}$$

Taking segmentation into consideration, α_{i2} is set as the benchmark level for both segments such that

$$P(choice_{it} = 1|\kappa = 1) = \frac{e^{V^1_{i1t}}}{e^{V^1_{i1t}} + e^{V^1_{i2t}}} = \frac{e^{\alpha^1_{i1} + \beta^1_1 price_{i1t}}}{e^{\alpha^1_{i1} + \beta^1_1 price_{i1t}} + e^{\beta^1_1 price_{i2t}}}$$

$$P(choice_{it} = 2|\kappa = 2) = \frac{e^{V^2_{i2t}}}{e^{V^2_{i1t}} + e^{V^2_{i2t}}} = \frac{e^{\beta^2_1 price_{i2t}}}{e^{\alpha^2_{i1} + \beta^2_1 price_{i1t}} + e^{\beta^2_1 price_{i2t}}}$$

Given Segment $m = 1, 2$, the likelihood for observing the choice history is

$$L_m = \prod_{i=1}^{300} \prod_{t=1}^{10} \prod_{j=1}^{2} P(choice_{it} = j|m)^{Y_{ijt}}$$

where Y_{ijt} is an indicator variable such that

$$Y_{ijt} = \begin{cases} 0, & \text{if brand j was chosen by household i at time t} \\ 1, & \text{if brand j was not chosen by household i at time t} \end{cases}$$

Denote the segment probabilities as $\kappa_1 \geq 0$ and $\kappa 2 \geq 0$ such as

$$\kappa_1 + \kappa_2 = 1$$

The complete data likelihood assuming two latent segments is

$$L = \sum_{m=1}^{2} \kappa_m L_m = \sum_{m=1}^{2} \kappa_m (\prod_{i=1}^{300} \prod_{t=1}^{10} \prod_{j=1}^{2} P(choice_{it} = j|m)^{Y_{ijt}})$$

Question 2 Model Estimation. I estimate this model with R. The output is shown as follows:

Table 1: Segmentation Selection based on AIC and BIC

	ll	AIC	BIC
1	3771.227	3775.23	3787.24
2	3349.481	3359.48	3389.51
3	3180.037	3196.04	3244.09
4	3175.588	3197.59	3263.66

These estimates are robust to different starting points. Thus, jointly based on AIC and BIC, I choose Latent Class Model with 3 segments.

Question 3 Segment Interpretation. (Results are shown with Table 2)

Based on the estimated weights for different segments (share1 and share2 variables), the first segment takes up 18.6% of total customers and the segment cluster takes up about 22.1% of the total market share.

The first segment has significantly negative inherent preferences towards Brand 1, whereas the second segment has significantly positive inherent preferences towards Brand 1. Without obvious brand preferences, the third segment is very price sensitive, taking up 59.3% of the total customers.

Question 4 Model Extension. To represent more accurate purchase behavior, it is useful to collect both individual-specific variables and alternative-specific ones.

In addition to price, researchers can collect more information about the attributes about the products from these two brands such as size, promotion, advertising, etc. These alternative-specific variables can be incorporated into the indirect utility U_{ijt}.

For individual-specific demographic variables such as race, gender and socioeconomic status, these demographic variables can serve as the basis for market segmentation such that we can construct the individual choice probabilities conditional on these demographic variables like $P(choice_{it} = j|gender)$.

Question Appendix R Code. (Modified from the code provided in the class)

```
library("arm")
library("numDeriv")
## library("texreg")
NX = 2 #number of parameters per segment
NS = 4 #number of consumer segments
NJ = 2 #number of brands
NT = 10 #number of period
NI = 300 #number of people

data = matrix(scan("M:/A Master of Science in Marketing Sciences/Empirical Models in Mar
data = t(data)

#Data File data2.CSV
# Col1= Customer ID
# Col2 = Time period
# Col3 = choice 1 if choice brand A and 2 if chose brand B
# Col4-5 = prices for brand A and brand B, respectively

# 1 segment1
coef.vec = rnorm(3 * NS -1) #parameters to be estimated.

log.lik <- function(coef.vec,data,ns,nj,nt,ni,nx)    #likelihood function
{
  probability = exp(coef.vec[3*(1:(ns - 1))])/(1 + sum(exp(coef.vec[3*(1:(ns - 1))])))
  prob = array(NA,nj)

  overall = 0
  ## Individual
  for (i in 1:ni)
  {
    ## Segment
    persontotal = 0
    for (mj in 1:ns)
    {
      total = 1
      ## Time Series
      for (k in 1:nt)
      {
```

```
        row = (i - 1)*NT + k
        V1 = coef.vec[3*(mj-1) + 1] + coef.vec[3*(mj-1) + 2] * data[row,4] #the systemat
        V2 = coef.vec[3*(mj-1) + 2]*data[row,5] #the systematic utility for 2

        prob[1] = exp(V1)/(exp(V1) + exp(V2)) # logit for 1
        prob[2] = exp(V2)/(exp(V1) + exp(V2)) # logit for 2
        choice = data[row,3]
        ################## Remains to be updated
        total = total * prob[choice]
      }
      if (mj < ns)
      {
        persontotal = persontotal + probability[mj] * total
      } else
      {
        persontotal = persontotal + (1 - sum(probability)) * total
      }
    }
    overall = overall + log(persontotal)
  }
  return(-overall)
}

# optimization procedure to calculate the MLE estimates

mle <- nlm(log.lik,coef.vec,data = data,ns = NS,nj = NJ,nt = NT,ni = NI,nx = NX, hessian

# calculating the Hessian to obtain stdev
mode = mle$estimate   # output parameter estimates
SE = sqrt(diag(solve(mle$hessian))) # output parameter SEs
Tvalue = mode/SE # output parameter T-values
ll = 2*mle$minimum # -2*log-likelihood
np = length(coef.vec)                    # number of parameters
AIC = 2*(mle$minimum + np)                  # calculates AIC
n = sum(NI*NT)                           # number of observations
BIC = 2*mle$minimum + np*log(n)            # calculates BIC

list(Estimate = mode,SE = SE,Tvalue = Tvalue,minus2ll = ll,AIC = AIC,BIC = BIC)
```

Table 2: Segment Statistical Summary

	Estimate	SE	T-Statistic
alpha1	**-2.14468**	0.198236	-10.8188
alpha2	**2.236913**	0.178148	12.55647
alpha3	-0.04889	0.073672	-0.66361
beta1	-0.58483	0.163236	-3.58274
beta2	-0.39891	0.146758	-2.71816
beta3	**-1.40499**	0.083662	-16.7938
share1	-1.15577	0.187439	-6.16611
share2	-0.98481	0.168446	-5.84643