Homework 3 – Estimating a Logit Model

Jin Miao
Empirical Models in Marketing

March 7, 2018

Question 1 Model 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

$$V_{ijt} = \beta_{ij} + \beta_{price} Price_{ijt} + \beta_{state\_dependence} Y_{ijt-1}$$

where Y_{ijt-1} is an indicator variable such that

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

ϵ_{ijt} are IID Extreme Value Distributed so that 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

$$V_{i2t} - V_{i1t} = \beta_{i2} + \beta_{price} Price_{i2t} + \beta_{state\_dependence} Y_{i2t-1} - (\beta_{i1} + \beta_{price} Price_{i1t} + \beta_{state\_dependence} Y_{i1t-1})$$

$$= (\beta_{i2} - \beta_{i1}) + \beta_{price}(Price_{i2t} - Price_{i1t}) + \beta_{state\_dependence}(Y_{i2t-1} - Y_{i1t-1})$$

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 identification issues into consideration, β_{i2} is set as the benchmark level for both segments such that

$$P(choice_{it}=1) = \frac{e^{V_{i1t}}}{e^{V_{i1t}}+e^{V_{i2t}}} = \frac{e^{\beta_{i1}+\beta_{price}Price_{i1t}+\beta_{state\_dependence}Y_{i1t-1}}}{e^{\alpha_{i1}+\beta_{price}Price_{i1t}+\beta_{state\_dependence}Y_{i1t-1}}+e^{\beta_{price}Price_{i2t}+\beta_{state\_dependence}Y_{i2t-1}}}$$

$$P(choice_{it}=2) = \frac{e^{V_{i2t}}}{e^{V_{i1t}}+e^{V_{i2t}}} = \frac{e^{\beta_{price}Price_{i2t}+\beta_{state\_dependence}Y_{i2t-1}}}{e^{\alpha_{i1}+\beta_{price}Price_{i1t}+\beta_{state\_dependence}Y_{i1t-1}}+e^{\beta_{price}price_{i2t}+\beta_{state\_dependence}Y_{i2t-1}}}$$

Taking segmentation κ into consideration,

$$P(choice_{it}=1|\kappa) = \frac{e^{V^{\kappa}_{i1t}}}{e^{V^{\kappa}_{i1t}}+e^{V^{\kappa}_{i2t}}} = \frac{e^{\beta^{\kappa}_{i1}+\beta^{\kappa}_{price}Price_{i1t}+\beta_{state\_dependence}Y_{i1t-1}}}{e^{\alpha^{\kappa}_{i1}+\beta^{\kappa}_{price}Price_{i1t}+\beta^{\kappa}_{state\_dependence}Y_{i1t-1}}+e^{\beta^{\kappa}_{price}Price_{i2t}+\beta^{\kappa}_{state\_dependence}Y_{i2t-1}}}$$

$$P(choice_{it}=2|\kappa) = \frac{e^{V^{\kappa}_{i2t}}}{e^{V^{\kappa}_{i1t}}+e^{V^{\kappa}_{i2t}}} = \frac{e^{\beta^{\kappa}_{price}Price_{i2t}+\beta^{\kappa}_{state\_dependence}Y_{i2t-1}}}{e^{\alpha^{\kappa}_{i1}+\beta^{\kappa}_{price}Price_{i1t}+\beta^{\kappa}_{state\_dependence}Y_{i1t-1}}+e^{\beta^{\kappa}_{price}price_{i2t}+\beta^{\kappa}_{state\_dependence}Y_{i2t-1}}}$$

Question 2 Complete Data Likelihood. Given the specific segment m, 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 3 Effect of State Dependence. I estimate the homogeneous models with and without state dependence. The output is shown as follows:

Table 1: Homogeneous Model with State Dependence

	Estimate	SE	Tvalue	AIC	BIC
intercept	0.04	0.04	0.92	4150.54	4168.56
price	-0.79	0.04	-18.32	4150.54	4168.56
state_dependence	0.25	0.04	6.49	4150.54	4168.56

In terms of parameter estimates, these two models have similar parameters for brand intercept and price. State Dependence has significantly positive effect for brand choice. The estimation results support state dependence rather than variety seeking.

In terms of fit, adding State Dependence variable increases AIC and BIC. One possible explanation is that the degree of state dependence varies across different segments of customers.

Table 2: Homogeneous Model without State Dependence

	Estimate	SE	Tvalue	AIC	BIC
intercept	0.04	0.04	1.03	3775.23	3787.24
price	-0.78	0.04	-18.21	3775.23	3787.24

Question 4 Model Selection. Based on the BIC and AIC creterion, the model that best fots the data is the State Dependence Model with 3 Segments.

Table 3: Variety Seeking Models with Segmentation

Segments	log-lik	AIC	BIC
2	3335.8	3349.8	3391.84
3	3178.96	3200.96	3267.03

Question 5 Model Interpretation. Based on the estimated weights for different segments (share1 and share2 variables), the first segment takes up 22.5% of the customers and the second segments takes up about 58.7% of the market share.

$$MarketShare_i = \frac{exp(sh\hat{a}re_i)}{\sum_{j=1}^{3} exp(sh\hat{a}re_j)}$$

Based on Table 4, consumers in the first segment typically have inherent favorable preferences for *Brand 1* whereas consumers in the third segment typically have inherent favorable preferences for *Brand 2*. The second market segment is characterized by their high price sensitivity.

Table 4: Estimates of State Dependence Model with Three Segments

Variables	Estimate	SE	Tvalue
Intercept1	**2.193**	0.290	7.565
Intercept2	-0.056	0.075	-0.753
Intercept3	**-2.240**	0.350	-6.402
beta_price1	-0.410	0.145	-2.822
beta_price2	**-1.407**	0.085	-16.558
beta_price3	-0.591	0.162	-3.652
beta_state_dependence1	0.034	0.266	0.127
beta_state_dependence2	-0.063	0.068	-0.919
beta_state_dependence3	-0.113	0.305	-0.369
share1	0.186	0.211	0.879
share2	1.145	0.191	6.009

With respects to state dependence, the estimates for variety seeking parameters from these three segments are insignificant, indicating that we do not find state dependence from this model.

Table 5: Models with versus without State Dependence

	Without State Dependence			With State Dependence		
Variables	Estimate	SE	Tvalue	Estimate	SE	Tvalue
Intercept1	**-2.145**	0.198	-10.819	**2.193**	0.290	7.565
Intercept2	**2.237**	0.178	12.556	-0.056	0.075	-0.753
Intercept3	-0.049	0.074	-0.664	**-2.240**	0.350	-6.402
beta_price1	-0.585	0.163	-3.583	-0.410	0.145	-2.822
beta_price2	-0.399	0.147	-2.718	**-1.407**	0.085	-16.558
beta_price3	**-1.405**	0.084	-16.794	-0.591	0.162	-3.652
beta_state_dependence1				0.034	0.266	0.127
beta_state_dependence2				-0.063	0.068	-0.919
beta_state_dependence3				-0.113	0.305	-0.369
share1	-1.156	0.187	-6.166	0.186	0.211	0.879
share2	-0.985	0.168	-5.846	1.145	0.191	6.009

Question 6 Effect of State Dependence and Heterogeneity. The comparison between Models (with Heterogeneity) with versus without State Dependence is shown as Table 5.

The interpretation for these two models are quite similar. There are one segment represented by inherent strong preference for *Brand 1* and another for *Brand 2*. The third segment is highly price-sensitive. As shown with bold numbers, the estimates for these three segments from the models with and without state dependence are almost the same.

Table 6: Models (with Three Segments) with versus without State Dependence

	log-lik	AIC	BIC
without State Dependence	3180.04	3196.04	3244.09
with State Dependence	3178.96	3200.96	3267.03

The insignificant estimates of state dependence variables are consistent with the AIC and BIC measures as shown in Table 6. On top of that, the best model should exclude state dependence variables.

In conclusion, the market exhibits heterogeneity but NOT state dependence.

```
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 = 3 #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

coef.vec = rnorm(4 * NS - 1) #parameters to be estimated.
log.lik <- function(coef.vec,data,ns,nj,nt,ni,nx)    #likelihood function
{
  probability = exp(coef.vec[4*(1:(ns - 1))])/(1 + sum(exp(coef.vec[4*(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
        if (k == 1)
        {
          V1 = coef.vec[4*(mj - 1) + 1] + coef.vec[4*(mj - 1) + 2] * data[row,4]
          # the systematic utility for 1
          V2 = coef.vec[4*(mj - 1) + 2] * data[row,5]
          # the systematic utility for 2
        } else if (k != 1)
        {
          lag = coef.vec[4*mj - 1]
          V1 = coef.vec[4*(mj - 1) + 1] + coef.vec[4*(mj - 1) + 2]
          * data[row,4] + lag * as.numeric(data[row - 1,3] == 1)
          # the systematic utility for 1
          V2 = coef.vec[4*(mj - 1) + 2] * data[row,5] + lag
          * as.numeric(data[row - 1,3] == 2)
          # 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]
        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 = TRUE)

# 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)
```