Topic A4:
Advanced topics of segmentation

Takeaki Sunada[1]

[1]Simon Business School
University of Rochester

Advanced topic 1: Rationale behind the logit specification

- Let's start from one question we left unaddressed - why do we use the ratio of exponential as our model?
- Indeed, our expression of choice probability is based on a model of consumer's decision making developed in Economics (”discrete choice model”).
- The model provides the rationale of why we use the ratio of exponentials, and also why $-\frac{\beta_0^j}{\beta_1}$ represents each consumer's willingness to pay.
- For simplicity, let's forget about segmentation and consider a simple logit model.

A discrete choice model

- Consider a situation where a consumer chooses one option from $\{1, ..., j, ..., J\}$ discrete alternatives (e.g. product, brand).
- Let's try to create a simple model that describes how a consumer may make decisions.
- The simplest formulation is probably the trade-off between the "value" that a consumer gets from getting product j, and the price she has to pay to get it (P^j). She compares between the two, and she may choose the alternative that gives her the highest surplus (value left with her after paying price P^j).

A discrete choice model

- If a consumer chooses an option j, she receives some value. Let's denote that value by β_0^j (called "utility" in Economics). For example, product KB gives her value β_0^{KB}.
- This β_0^j can represent any intangible satisfaction of the consumer from consuming this product. Hence:
  1. It can mean anything - different consumers value a product for different reasons.
  2. It doesn't have to be in dollar term. For example, if a consumer believes that she values a bottle of milk by "6". It doesn't have to be 6 dollars. Different consumers may use different scales. The important assumption is that consumers know the product value for her. It doesn't matter if she can measure it in dollar terms.

A discrete choice model

- Suppose also that by paying the price P^j, each consumer feels loss of $\beta_1 P^j < 0$.
- Again, $\beta_1 P^j$ doesn't have to be in dollar values. $\beta_1 P^j$ represents the feeling of dissatisfaction from having to spend money, and different people may have different scales.
- In this course, we assume that β_1 (how consumers evaluate losing P^j) is not product-specific. Dissatisfaction of losing 5 dollars is the same, whether it is spent for KB, or KR.

A discrete choice model

- Suppose that consumers compare between their value β_0^j and the loss $\beta_1 P^j$ in making purchase decisions.
- In other words, they see some surplus (value left with them after paying price P) from each product, which is

  $$u^j = \beta_0^j + \beta_1 P^j.$$

  Let's assume that they pick the one with the highest u^j. i.e.

  $$\beta_0^j + \beta_1 P^j > \max\{\beta_0^{j'} + \beta_1 P^{j'}\},$$

  for all $j' \neq j$ that is available, then they choose product j.

A discrete choice model

$$\beta_0^j + \beta_1 P^j > \max\{\beta_0^{j'} + \beta_1 P^{j'}\},$$

- This looks like a simple approximation of how consumers choose between J alternatives. However, there's one problem with it.
- In reality, how consumers value the product likely varies across choice occasions - idiosyncratic factors just affect our decisions so much.
- This means that β_0^j varies from time to time.

Include randomness

- Allowing β_0^j to be time-varying will complicate the model notation, so let's consider an alternative way.
- Suppose that consumers product valuation at each shopping trip consists of two terms: $\beta_0^j + \epsilon^j$. β_0^j represents each consumer's inherent product valuation (invariant across time), and ϵ^j captures any idiosyncratic change in taste across occasion.

$$u^j = \underbrace{\beta_0^j + \beta_1 P^j}_{\text{True consumer type}} + \underbrace{\epsilon^j}_{\text{Idiosyncratic}}.$$

A discrete choice model

- At each shopping trip, consumers choose j if j provides the highest surplus. i.e.

$$\beta_0^j + \beta_1 P^j + \epsilon^j > \max\{\beta_0^{j'} + \beta_1 P^{j'} + \epsilon^{j'}\}.$$

- ϵ^j and $\epsilon^{j'}$ fluctuate across different shopping occasions. So even if a consumer's inherent valuation $\beta_0^j + \beta_1 P^j$ remains the same over time, that consumer may switch across different j's due to ϵ shocks.
- Formally, ϵ^js are called a random variable, which makes whether or not each consumer buys j a *probabilistic event*. For a given $\beta_0^j + \beta_1 P^j$ for each j, a consumer chooses j with some probability, $Pr(y = j \mid P)$.

A discrete choice model

- Now it's time to wrap up - It turns out that if we assume that this ϵ follows a particular distribution, then the probability that a consumer buys product j is represented by a formula we already know:

$$Pr(y = j \mid P) = \frac{\exp(\beta_0^j + \beta_1 P^j)}{1 + \sum_{j'=1}^{J} \exp(\beta_0^{j'} + \beta_1 P^{j'})},$$

- This representation holds when $\epsilon^j - \epsilon^{j'}$ follows a logistic distribution - hence the name "logit models".

Consumer's willingness to pay

- Now let's discuss why $-\frac{\beta_0^j}{\beta_1}$ represents willingness to pay for j.
- Consumer's valuation for each product j in this model is β_0^j, however, it is not measured in dollar term, and hence it is not willingness to pay per se, which needs to be measured in dollar term.
- However, as we discussed earlier, taking the ratio between β_0^j and β_1 magically derives willingness to pay.

$$\text{Willingness to pay for product j} = -\frac{\beta_0^j}{\beta_1}.$$

Consumer's willingness to pay

- The intuition is as follows:
  1. β_0^j is a consumer's value for product j *in her own scale*.
  2. $\beta_1 P^j$ is her dissatisfaction from losing P^j dollars *in her own scale*.

     $\Rightarrow \beta_1$ is her dissatisfaction from losing *one dollar*, in her own scale.

     $\Rightarrow -\beta_1$ is *her valuation of one dollar in her own scale*.

- If we take the ratio, then:

$$\frac{\beta_0^j}{-\beta_1} = \frac{\text{Value from product } j \text{ in her scale}}{\text{Value from one dollar in her scale}} = \text{Value from product } j \text{ in dollar scale}$$

"her own scale" part is common between the numerator and the denominator, and hence get cancelled.

Advanced topic 2: Causality in demographics

- Earlier, we claimed that we want to estimate a causal relationship *from* the price *to* the demand.
- Now that we include consumer characteristics, and there is a bunch of coefficients associated with them. Do we need a causal relationship on them too? or a predictive relationship is enough?

Causality in demographics

- Recall the definition: we want causality from X to Y when we want to measure "what happens to Y if we manipulate X".
- Here we don't manipulate demographics - we just manipulate price (for which causality is established) and we simply want to predict its effect on different consumer segments.
- Hence predictive relationship is sufficient. In fact, we used K-mean clustering to group consumers together - a purely predictive operation (predict demand based on similarity of demographics).

Causality in demographics

- This is why we can use BIC as a criterion of our model selection, which is a concept of predictive analysis (R^2-alternative in MLE environments).
- We assumed that the price - choice probability relationship we see in the data is causal (so no need to establish it). We also don't need causal relationship regarding demographics. Hence we can focus on predictive performance of the model.
- Of course if we are concerned about potential non-causal relationship between price and choice probability, we do something similar to what we did in sales data environment (check price coefficients, include X to eliminate contamination, etc.).

Advanced topic 3: Any way to run BIC in Kmeans environment?

- Earlier, I claimed that with K-mean clustering approach, we cannot compare model fit using BIC, as parameters are estimated with multiple separate MLE (gmnl applied to each segment separately). Hence we cannot compare K-mean approach to regression-in-logit approach.
- This is really inconvenient. Is there any way to apply BIC to K-mean approach? i.e. Any way to estimate K-mean model with *one* gmnl?
- Indeed, if we apply what we have covered by now, we can do it.

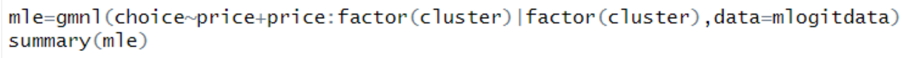
Estimate separate parameters per segment

- Running separate MLE across each clustered segment corresponds to estimating separate β_{0k}^{j} and β_{1k} for each clustered segment.
- Consider a regression-in-logit environment, where we can use BIC. If we can estimate a model in a way that β_{0k}^{j} and β_{1k} differ across clustered segments, but are the same within each segment, then we are done. Is there any way that we can do it?

Estimate separate parameters per segment

- Recall that when we run a K-means approach, we first apply "kmeans" function to demographic data to create a "cluster ID" variable, which is a categorical variable that records which consumer belongs to which segment. We then user that variable to split up the data sheet.
- Instead of splitting the data, suppose that we build a regression-in-logit model, where we include cluster ID as a categorical variable. Then parameters vary across different realization of cluster ID (=varies across clusters), and we are done.

Estimate separate parameters per segment



```
mle=gmnl(choice~price+price:factor(cluster)|factor(cluster),data=mlogitdata)
summary(mle)
```

- We can utilize fixed effects - we include fixed effects in both β^j_{0k} and β_{1k} by "factor(cluster)". The variable "cluster" is my "cluster ID" variable.

Estimate separate parameters per segment

- Then this is the model we are effectively estimating. Each clustered segment of consumers have the choice probability:

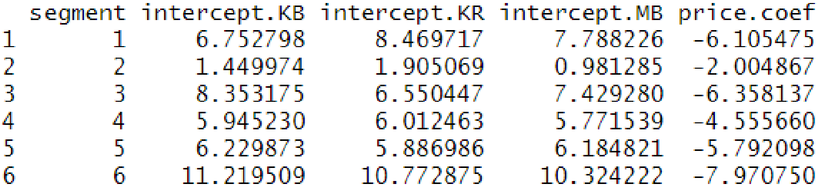
$$Pr_k(y = j \mid P) = \frac{\exp(\beta_{0k}^{j} + \beta_{1k} P^{j})}{1 + \sum_{j'} \exp(\beta_{0k}^{j'} + \beta_{1k} P^{j'})},$$

  where

$$\beta_{0k}^{j} = \beta_{base}^{j} + \beta_{seg2}^{j} \times I\{k = 2\} + \beta_{seg3}^{j} \times I\{k = 3\} + ...$$
$$\beta_{1k} = \beta_{base} + \beta_{seg2} \times I\{k = 2\} + \beta_{seg3} \times I\{k = 3\} + ...$$

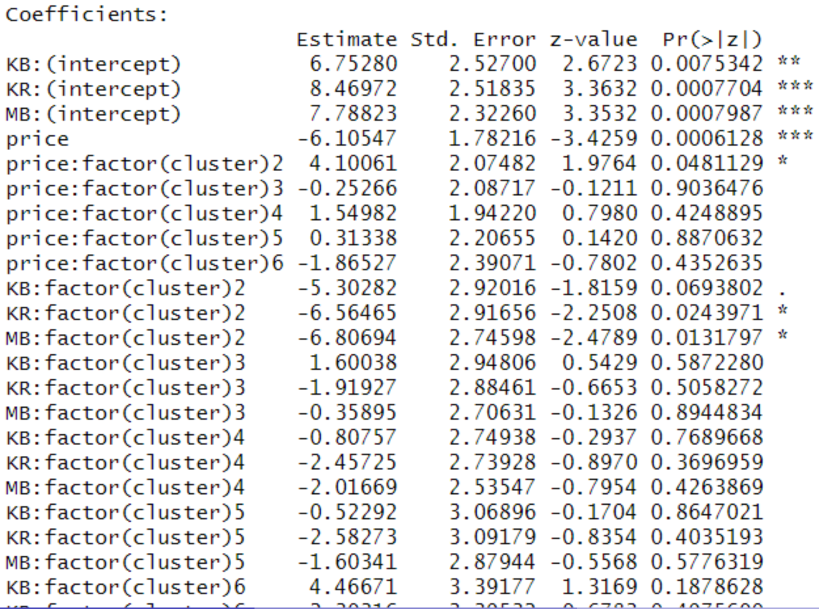
- Then $\beta_{01}^{j} = \beta_{base}^{j}$, $\beta_{02}^{j} = \beta_{base}^{j} + \beta_{seg2}^{j}$, and so on.

Repost: Parameter estimates from K-mean approach



```
  segment intercept.KB intercept.KR intercept.MB price.coef
1       1     6.752798     8.469717     7.788226  -6.105475
2       2     1.449974     1.905069     0.981285  -2.004867
3       3     8.353175     6.550447     7.429280  -6.358137
4       4     5.945230     6.012463     5.771539  -4.555660
5       5     6.229873     5.886986     6.184821  -5.792098
6       6    11.219509    10.772875    10.324222  -7.970750
```

Parameter estimates from the current model



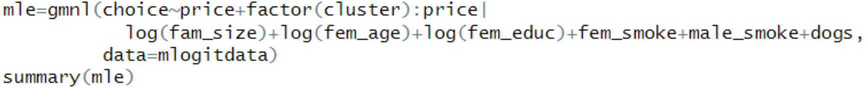
```
Coefficients:
                       Estimate Std. Error z-value  Pr(>|z|)    
KB:(intercept)          6.75280    2.52700  2.6723 0.0075342 ** 
KR:(intercept)          8.46972    2.51835  3.3632 0.0007704 ***
MB:(intercept)          7.78823    2.32260  3.3532 0.0007987 ***
price                  -6.10547    1.78216 -3.4259 0.0006128 ***
price:factor(cluster)2  4.10061    2.07482  1.9764 0.0481129 *  
price:factor(cluster)3 -0.25266    2.08717 -0.1211 0.9036476    
price:factor(cluster)4  1.54982    1.94220  0.7980 0.4248895    
price:factor(cluster)5  0.31338    2.20655  0.1420 0.8870632    
price:factor(cluster)6 -1.86527    2.39071 -0.7802 0.4352635    
KB:factor(cluster)2    -5.30282    2.92016 -1.8159 0.0693802 .  
KR:factor(cluster)2    -6.56465    2.91656 -2.2508 0.0243971 *  
MB:factor(cluster)2    -6.80694    2.74598 -2.4789 0.0131797 *  
KB:factor(cluster)3     1.60038    2.94806  0.5429 0.5872280    
KR:factor(cluster)3    -1.91927    2.88461 -0.6653 0.5058272    
MB:factor(cluster)3    -0.35895    2.70631 -0.1326 0.8944834    
KB:factor(cluster)4    -0.80757    2.74938 -0.2937 0.7689668    
KR:factor(cluster)4    -2.45725    2.73928 -0.8970 0.3696959    
MB:factor(cluster)4    -2.01669    2.53547 -0.7954 0.4263869    
KB:factor(cluster)5    -0.52292    3.06896 -0.1704 0.8647021    
KR:factor(cluster)5    -2.58273    3.09179 -0.8354 0.4035193    
MB:factor(cluster)5    -1.60341    2.87944 -0.5568 0.5776319    
KB:factor(cluster)6     4.46671    3.39177  1.3169 0.1878628    
```

Parameter estimates from the current model

- If we transform the coefficients to a proper β_{0k}^{j} and β_{1k} and store in a matrix, the estimated parameters from the current model will fit in the code you already have from Project 2.
- Moreover, we can now use BIC, as all the parameters are estimated by a single MLE.
- In practice, BIC is less reliable when the model involves any sort of discontinuity (e.g. models with categorical variables). Hence unlike the case where we compare two pure regression-in-logit models, we may not solely rely on BIC for model selection.
- Nevertheless, it is a good metric to look at.

Combining discrete clusters with regression-in-logit

- Even further, we could now then combine discrete clusters and linear components in β parameters within one model. For example, consider the following specification.



```
mle=gmnl(choice~price+factor(cluster):price|
           log(fam_size)+log(fem_age)+log(fem_educ)+fem_smoke+male_smoke+dogs,
         data=mlogitdata)
summary(mle)
```

- This assumes that β^{j}_{0k} is linear in log(demographics), and β_{1k} differs according to two clustered segments.

Combining discrete clusters with regression-in-logit

- Adding clusters as a categorical variable, on top of the demographic variables already included, may dramatically improve the fit of a regression-in-logit model.

$$\begin{aligned}\beta^j_{0k} = \beta^j_{0,fs} \times \mathit{fam\_size}_k + \beta^j_{0,fa} \times \mathit{fem\_age}_k + ... \\ + \beta^j_{seg2} \times I\{\text{k in seg 2}\} + \beta^j_{seg3} \times I\{\text{k in seg 3}\} + ...\end{aligned}$$

- Often, the way we directly include demographic variables does not allow any discontinuous jumps in β^j_{0k} as we move around the demographic variables (x, $\log(x)$, $\exp(x)$, x^2, etc. are all continuous functions). Indicator functions based on discrete segments allow for such jumps in β^j_{0k}.

Combining discrete clusters with regression-in-logit

- In this case, the demand/profit simulation with the estimated parameters is based on that of the regression-in-logit approach.
- We define the choice probability of each consumer as a function of her demographics, but also with her "cluster ID" as a categorical variable (see my code).
- As we have demographics directly entering in β^j_{0k}, we need to calculate the aggregate choice probability based on the average of every single consumer's $Pr_k(y = j \mid P)$.

Real summary

- Clustering can be nested as a part of regression-in-logit specification. Hence now we can evaluate the model fit of clustering approach by BIC. We can also compare between clustering and regression-in-logit now (but we should also use other criteria we used in Project 2).
- Moreover we can combine both approaches in one model. Categorical variables from clustering allows for discontinuous jump in how demographic variables impact consumer preference - we can pick the best of the two different sets of assumptions.
- We are now fully equipped with tools to study demographic variables.