Topic 4:
Introduction to market segmentation

Takeaki Sunada[1]

[1]Simon Business School
University of Rochester

Heterogeneous demand

- The logit model we have discussed so far takes the following form:

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

- We assumed that *everyone* follows this choice probability - for a given P^{KB} and P^{KR}, $Pr(y = KB \mid P)$ is the same across all consumers. This is because β_0^j and β_1 are the same across consumers.
- In practice, different consumers may have different propensity of choosing KB even under the same prices: loyalty, habit formation, or simply, different preferences and different price sensitivities.

Heterogeneous demand

- Today we consider incorporating heterogeneity through consumers' observed characteristics (e.g. demographics). In other words, we allow β_0^j and β_1 to vary across consumers with different demographic background (but the same among those with same demographics).
- We consider segmentation through consumers' latent (unobserved) types next time.

Why observed characteristics?

- Demographic-based targeting is a common business practice:
  - Senior discount, student discount
  - Distribute coupons by mail (discount for people living in a particular neighborhood)
  - Amazon family
- Sometimes observed characteristics do capture different demand structures. Households with young kids have different demand patterns; Consumers with different income have different price sensitivity, etc.

Targeting vs positioning

- However, note that allowing segmentation in the demand system does not necessarily need to result in targeting.
- "Targeting" is the term used when the firm sells the same product at different prices across different segments of consumers.
- There's also "positioning", in which the firm sells multiple products with different characteristics (Kiwi Bubble vs Kiwi Regular) to serve different segment of consumers.
- Hence understanding segmentation helps product-line pricing even without the intent of targeting.

Heterogeneity through observed characteristics

- To do segmentation based on observed characteristics, we need to estimate different demand across consumers with different characteristics.
- In the context of multinomial logit, this means we estimate different β_0^j and β_1 across different consumers.

Heterogeneity through observed characteristics

- There are two ways to achieve this:
  - Estimate separate multinomial logit models across subsample of consumers with different observed characteristics.
  - Use all observations to estimate one model, but include observed characteristics explicitly as part of the model.
- We cover both approaches.

Demographic data

id	fam_size	fem_age	fem_educ	fem_smoke	male_age	male_educ	male_smoke	dogs
1	3	4	4	1	3	6	0	1
2	2	6	4	0	7	9	0	1
3	2	6	4	0	6	4	0	0
4	3	3	6	0	3	4	0	1
6	4	3	5	0	3	5	0	0
7	2	3	4	0	7	9	0	0
8	2	5	5	1	6	3	1	1
9	1	6	4	0	7	9	0	0

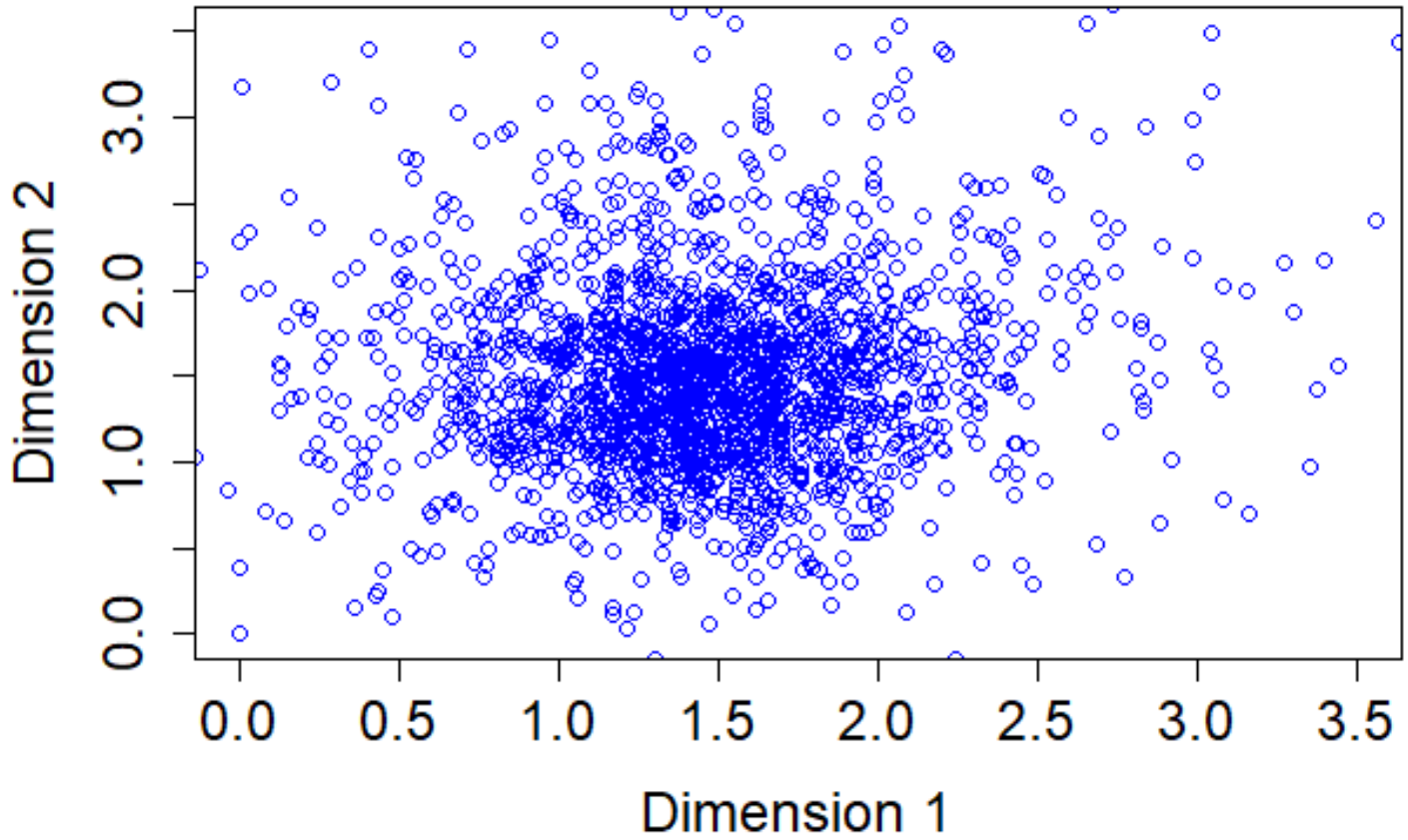
Using demographic variables to cluster segments

- First, let's consider the first approach - estimate separate logit models for each subgroup of people, defined by demographic variables.
- We have a lot of demographic variables in the data. Which one should we use? We cover a way to figure it out later.
- To begin with, we apply a K-mean clustering using all demographic variables. i.e. we assume that all demographic variables matter.

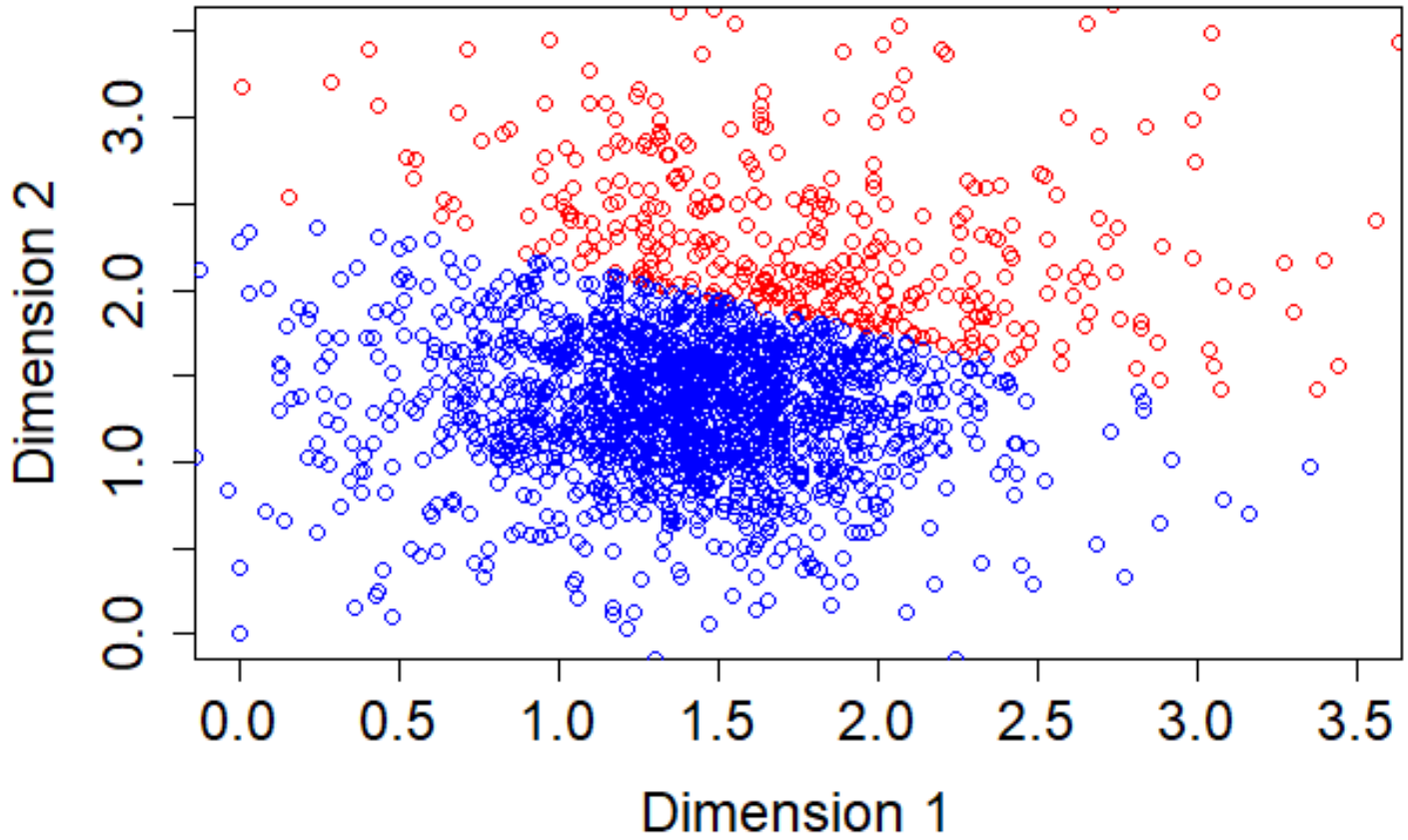
K-mean clustering

- K-mean clustering automatically clusters observations into subgroups based on the similarity of their characteristics with one another.
- It is one of the simplest forms of machine-learning - find a pattern from a bunch of attributes and group similar observations.
- In R, it is available in "kmeans" function.

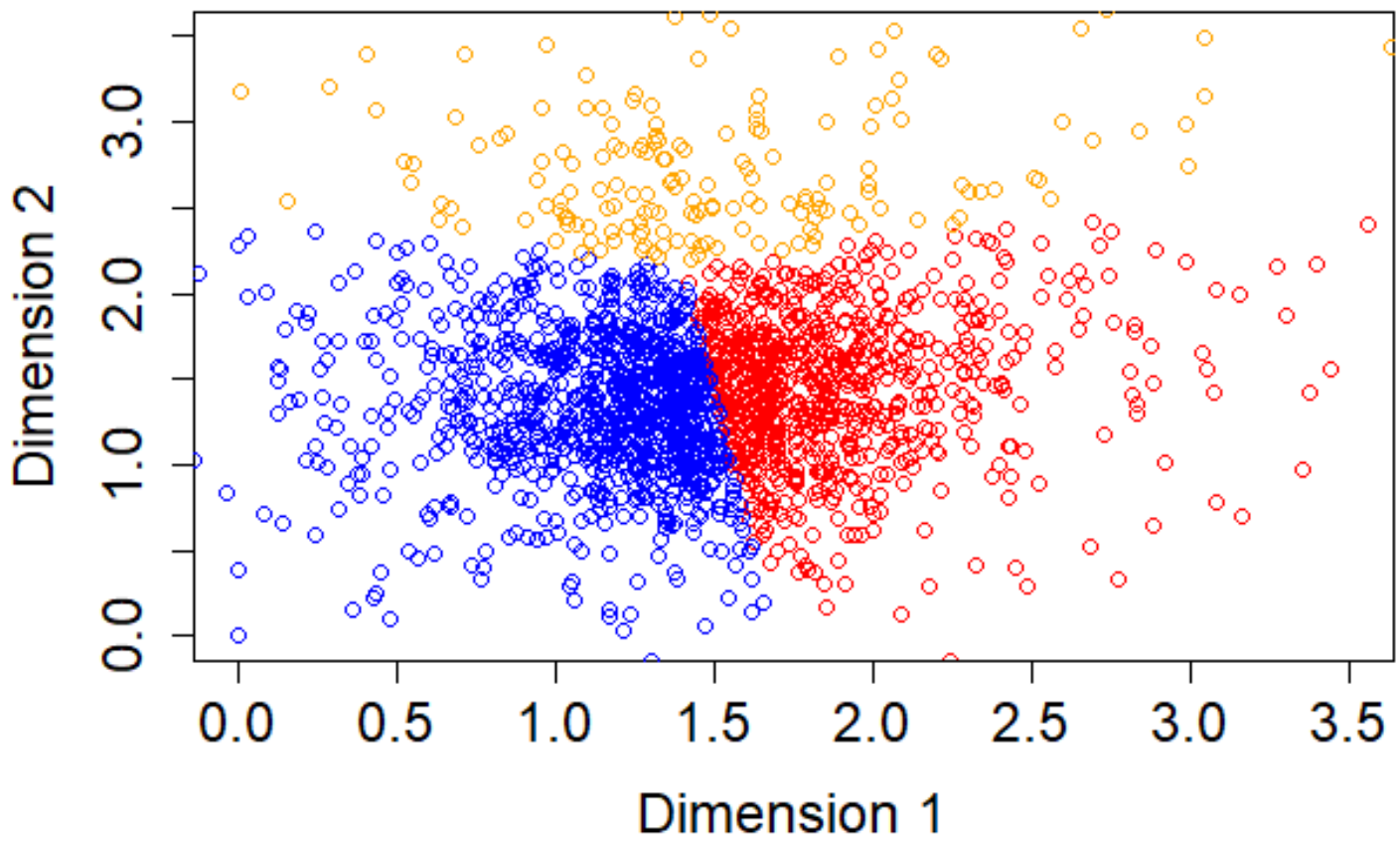
K-mean clustering with two-dimensional scatterplot



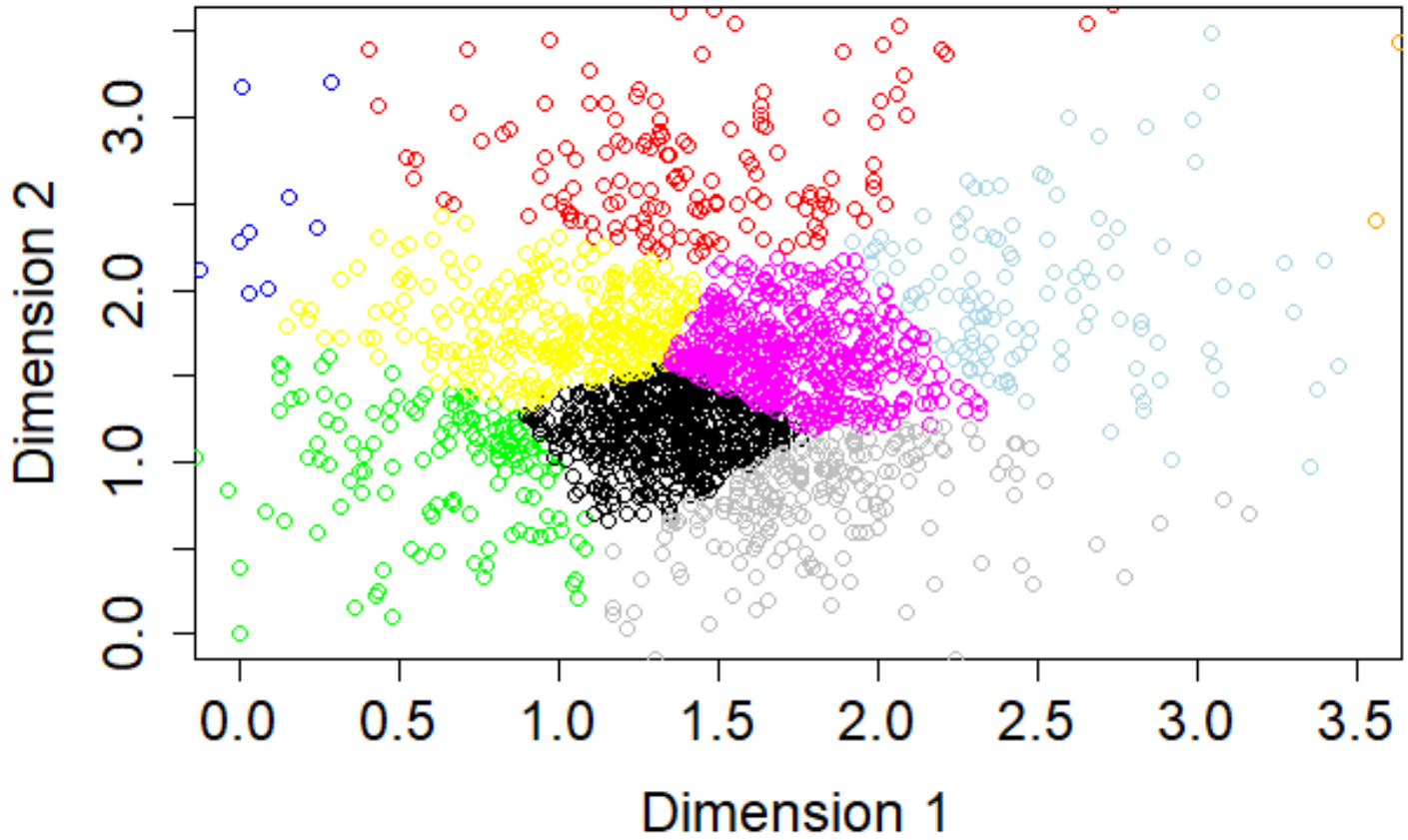
K-mean clustering with two-dimensional scatterplot



K-mean clustering with two-dimensional scatterplot



K-mean clustering with two-dimensional scatterplot



K-mean clustering in R

- In the code, we load the demographic data set, apply "kmeans" to it and assign a segment to each consumer (say "cluster ID").
- We merge that cluster ID with the original data. We then run mlogit-gmnl for each subsample with the same cluster ID.

Estimate multinomial logit with K-mean clusters

```
#Clustering
demo_cluster = kmeans(x=demo[, 2:9], centers = 5, nstart = 1000)
```

- "kmeans" takes three arguments.
  1. "x", which assigns the data we want to create segments based on. In our case, all columns of demographic data (excluding consumer ID).
  2. "centers", which assigns the number of segments.
  3. "nstart", which is always set to some large number.
- It then classifies each individual into segments and assign a cluster ID.

Estimate multinomial logit with K-mean clusters

```
#Clustering
demo_cluster = kmeans(x=demo[, 2:9], centers = 5, nstart = 1000)

# now combine cluster identity into the raw data
cluster_id = data.frame(id = demo$id)
cluster_id$cluster = demo_cluster$cluster
data = merge(data, cluster_id, by = "id", all.x = T)

# for those who don't fit in any segment, group them into one additional segment
data$cluster[is.na(data$cluster)] = 6

# segment share = proportion of consumers from each segment
seg.share = c( table(demo_cluster$cluster),N - sum(table(demo_cluster$cluster))) / N
```

- We merge the cluster ID with the original data.
- Some consumers are not assigned cluster / missing demographic info. Assign the 6th cluster to them.
- Finally, calculate the proportion of each cluster.

Clustering results

```
Cluster means:
  fam_size  fem_age fem_educ  fem_smoke male_age male_educ male_smoke      dogs
1 3.833333 3.083333 6.750000 0.08333333 3.083333  6.333333  0.1666667 0.4166667
2 1.615385 4.538462 7.692308 0.07692308 5.923077  8.230769  0.1538462 0.2307692
3 3.928571 3.428571 4.535714 0.21428571 3.607143  4.500000  0.2142857 0.6428571
4 2.263158 5.263158 4.947368 0.26315789 5.684211  4.842105  0.3157895 0.3157895
5 1.785714 5.214286 4.071429 0.21428571 6.928571  8.928571  0.0000000 0.2857143
```

```
> seg.share
   1    2    3    4    5
0.12 0.13 0.28 0.19 0.14 0.14
```

Run "gmnl" for each segment

```
#Write a for-loop.
for (seg in 1:6) {
  # During each loop, pick subset of data of consumers from each segment.
  data.sub = subset(data, cluster == seg)

  #Using that data, the rest remains the same.
  mlogitdata=mlogit.data(data.sub,varying=4:7,choice="choice",shape="wide")

  #Run MLE.
  mle= gmnl(choice ~  price, data = mlogitdata)
  mle
  #Store the outcome in the coef.est matrix.
  coef.est[seg, 2:5] = mle$coefficients
}
```

- "for" loop is a convenient way to run the same task muliple times.
- At each iteration, pick subset of consumers from each segment, run gmnl and store the coefficients.

Estimate multinomial logit with K-mean clusters

	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

- Compared to a single-segment model with 4 parameters (β_0 for three products and β_1), we can represent a more flexible demand system with 4 $\times$ 6 (number of segments) parameters.

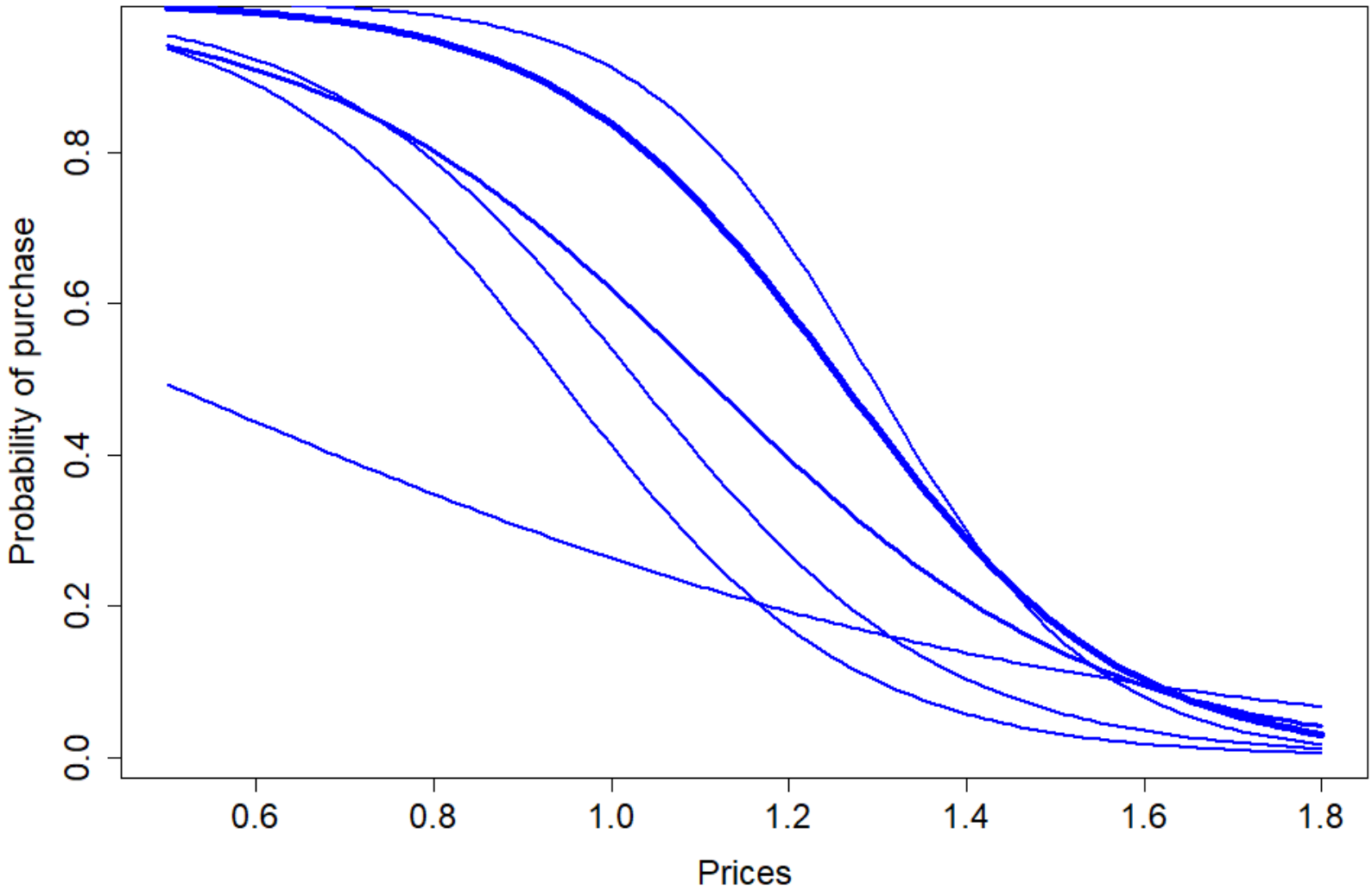
How do we pick the number of clusters?

- The optimal number of clusters depends on the trade-off between flexibility and sample size.
- More clusters = fewer consumers per cluster, based on which we estimate cluster-specific parameters. If we increase the number of clusters, the model becomes more flexible, but the estimated parameters become less reliable.
- This is the exact same trade-off that we had in the regression environments.

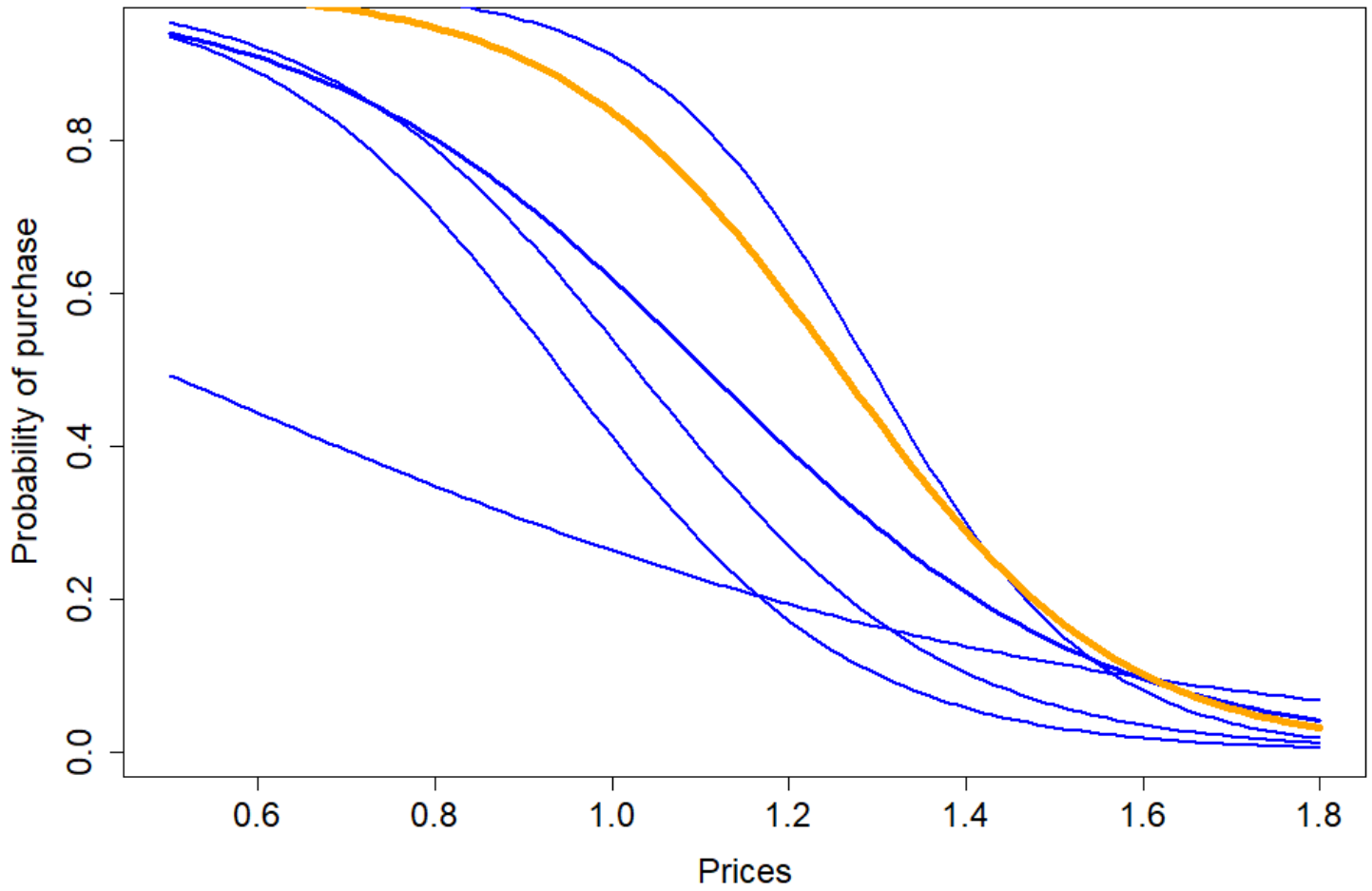
Notation

- Now that $Pr(y = KB)$ (suppressing its dependence on P) is segment-specific.
- We henceforth denote the choice probability of each segment by $Pr_k(y = KB)$, and logit parameters of that segment by β_{0k}^{KB}, β_{1k}, etc.

Choice probability of KB by each cluster

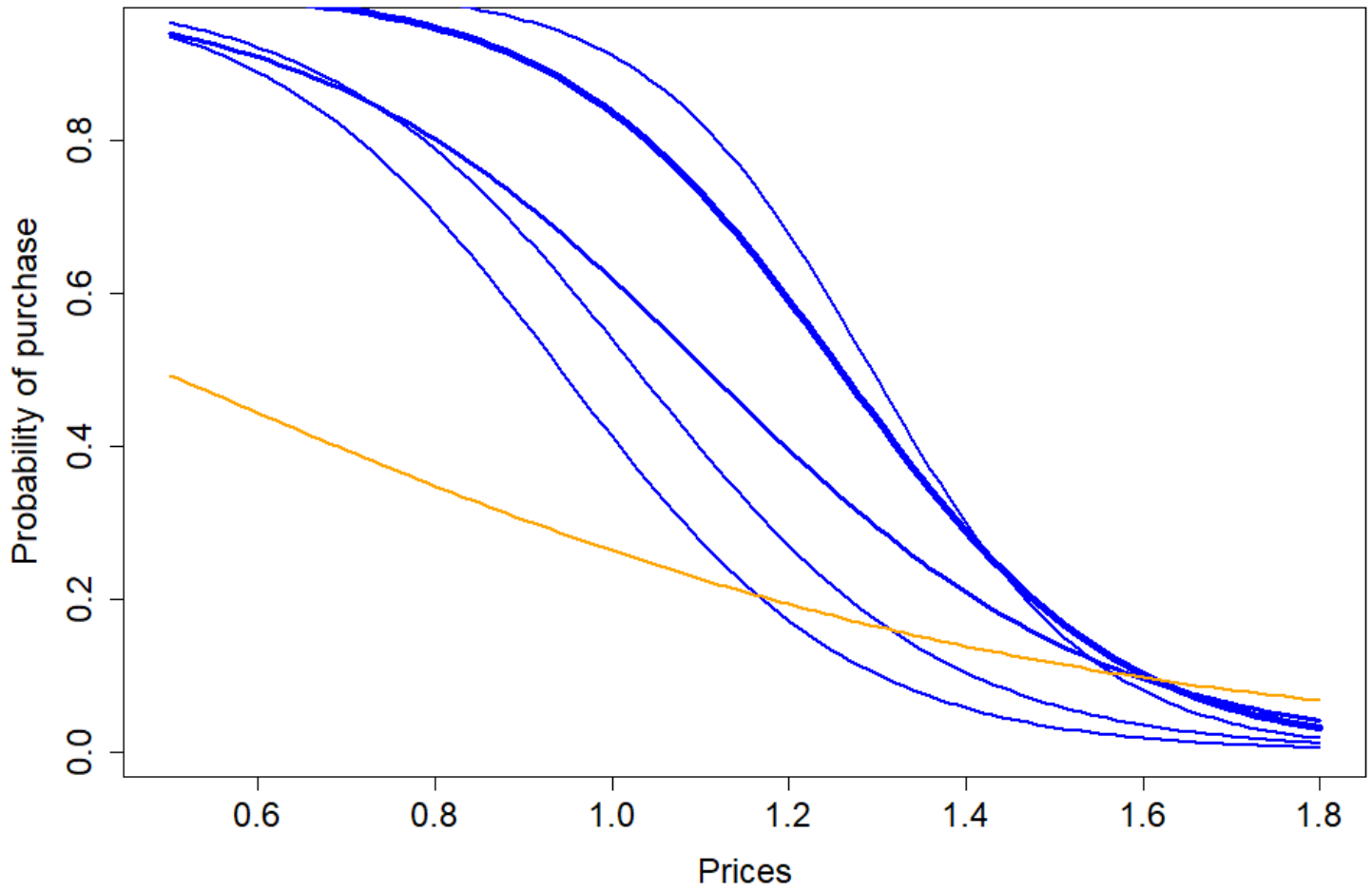


Largest segment



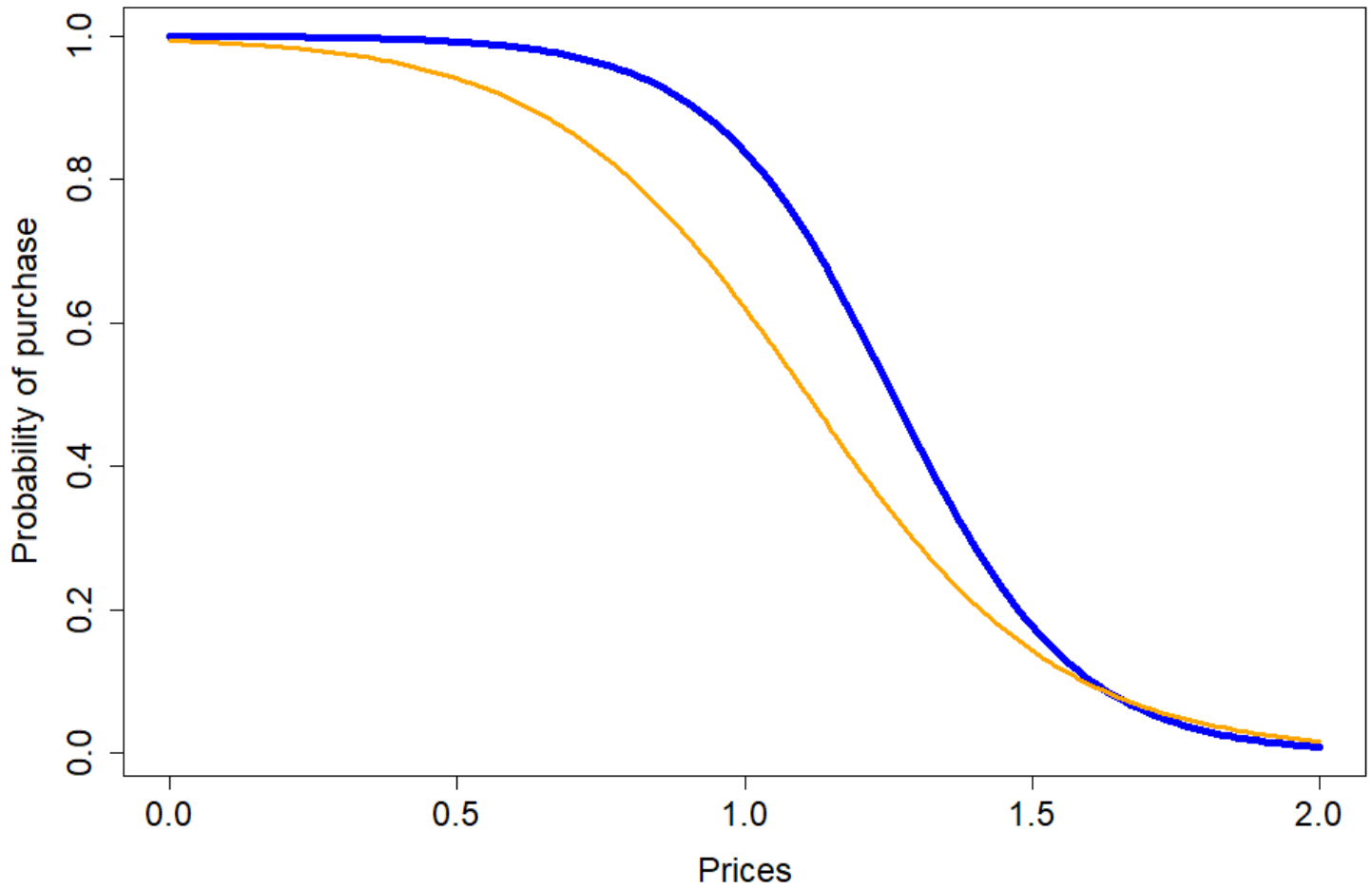
	fam_size	fem_age	fem_educ	fem_smoke	male_age	male_educ	male_smoke	dogs
3	3.928571	3.428571	4.535714	0.21428571	3.607143	4.500000	0.2142857	0.6428571

Least price-sensitive segment



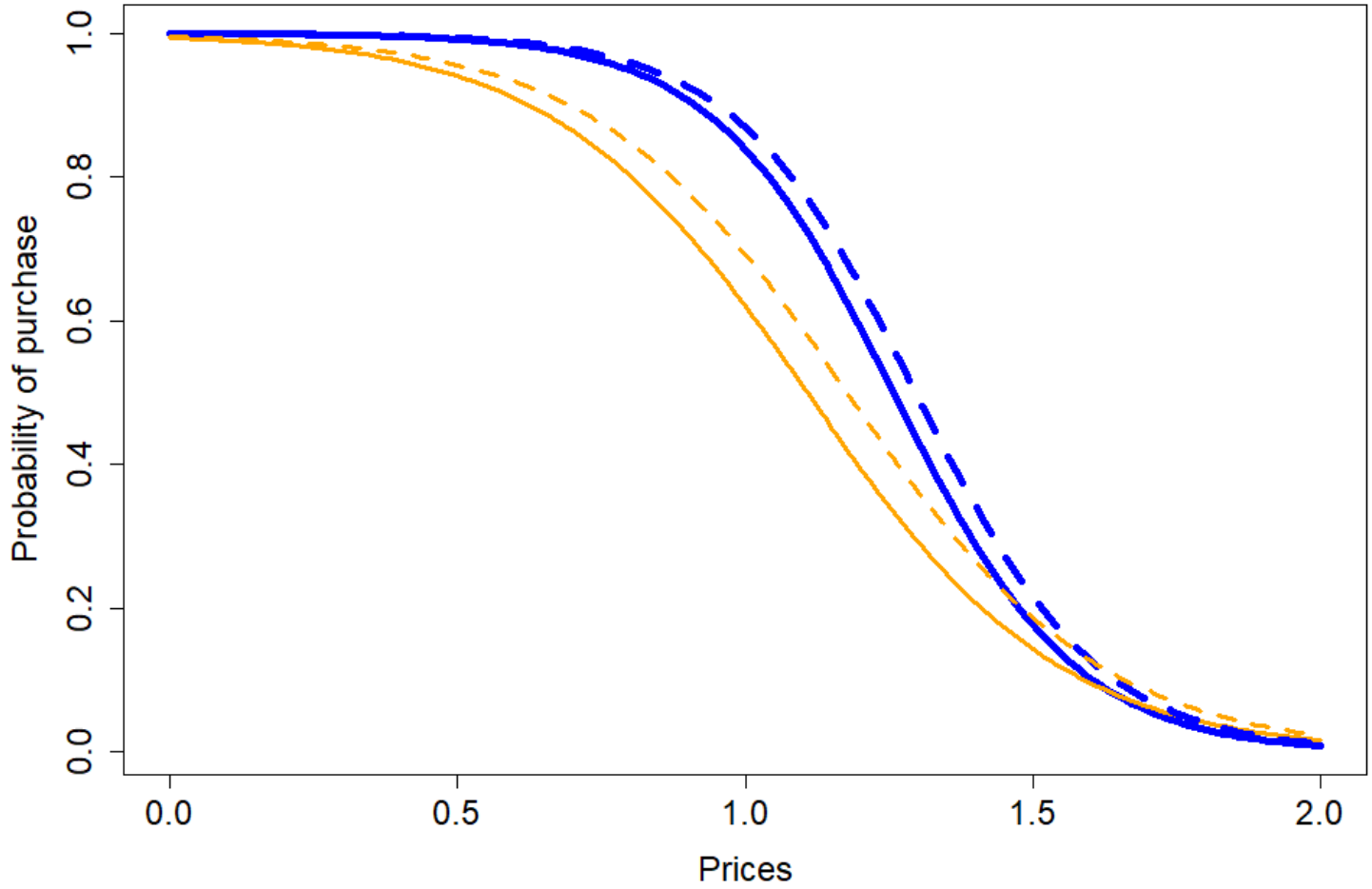
	fam_size	fem_age	fem_educ	fem_smoke	male_age	male_educ	male_smoke	dogs
2	1.615385	4.538462	7.692308	0.07692308	5.923077	8.230769	0.1538462	0.2307692

What happens if rival price increases?



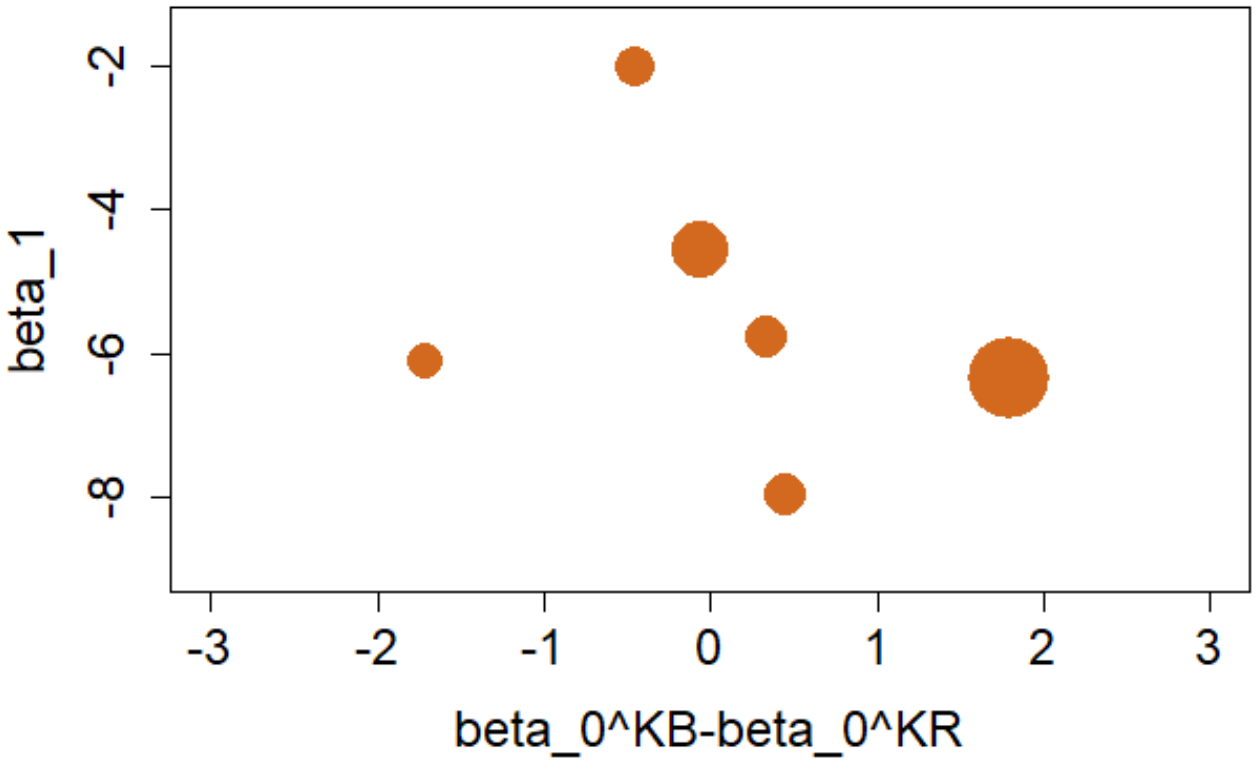
- Blue is demand for segment 3, and orange is segment 4.

What happens if a rival price increases?



- Suppose that MB increases its price by \$1. Dashed lines are the new demand.

Different angles of looking at it



- Horizontal axis represents how much each segment of consumers prefers KB over KR. Vertical axis is their price sensitivity.
- This provides an intuitive visualization of the location of consumer preference.

Comparison to homogeneous demand

```
Coefficients:
                Estimate Std. Error  z-value  Pr(>|z|)
KB:(intercept)  6.15202    0.60211  10.2174 < 2.2e-16 ***
KR:(intercept)  6.01682    0.60275   9.9823 < 2.2e-16 ***
MB:(intercept)  5.82222    0.56675  10.2730 < 2.2e-16 ***
price          -4.93438    0.43220 -11.4170 < 2.2e-16 ***
```

- Recall that when we assume homogeneous demand, we predict that consumers have similar preference across all products.
- Assuming homogeneous demand hence sometimes masks the substantial heterogeneity present in the data.

Comparison to homogeneous demand

- Why this difference?
- Assuming homogeneous demand, we estimate a single set of parameters by aggregating all the consumers' choices. The estimates represent an average preference across consumers - "One segment choosing KB often" and "another segment choosing KR often" will cancel out.
- With segment-specific parameters, we can capture such segment-specific variations without the averaging.

Setting a single optimal price

- Now that we have recovered heterogeneous demand. Let's maximize the profit against it.
- Suppose that this market consists of 1000 people. KB is our product and we want to set a single price. Our unit cost is one dollar. What price should we set?

Setting a single optimal price

- Recall that we want to maximize the following profit (same as the homogeneous demand case):

$$\text{Profit}^{KB} = \text{Market-level demand} \times \underbrace{(P^{KB} - \text{unit cost})}_{\text{Per-unit margin}}$$

- But with heterogeneous consumers, the market-level demand is now the sum of demands from all segments:

$$\begin{aligned}\text{Market-level demand} = {} & \text{Number of segment 1 consumers} \times Pr_1(y = KB|P) \\ & + \text{Number of segment 2 consumers} \times Pr_2(y = KB|P) \\ & + \ldots\end{aligned}$$

Aggregate choice probability

- Because "the number of segment k consumers = the total number of consumers $\times$ the proportion of consumers from segment k", we have:

$$
\begin{aligned}
&\text{Market-level demand} \\
=&\text{Total number of consumers} \times w_1 \times Pr_1(y = KB) \\
&+ \text{Total number of consumers} \times w_2 \times Pr_2(y = KB) + ... \\
=&\text{Total number of consumers} \times (w_1 Pr_1(y = KB) + w_2 Pr_2(y = KB)...
\end{aligned}
$$

  where I denote by w_k the proportion of consumers from segment k.

- Note that everything inside the parenthesis is known and hence the term inside the parenthesis is computable.

Aggregate choice probability

- So to compute the market-level demand under heterogeneous segments, we compute:

$$\sum_{k=1}^{K} w_k Pr_k(y = KB \mid P).$$

  and the rest is the same as the case of no segmentation (slide deck 3).
- Note that this is a weighted average of segment-specific choice probabilities, where the weight is given by the proportion of consumers.

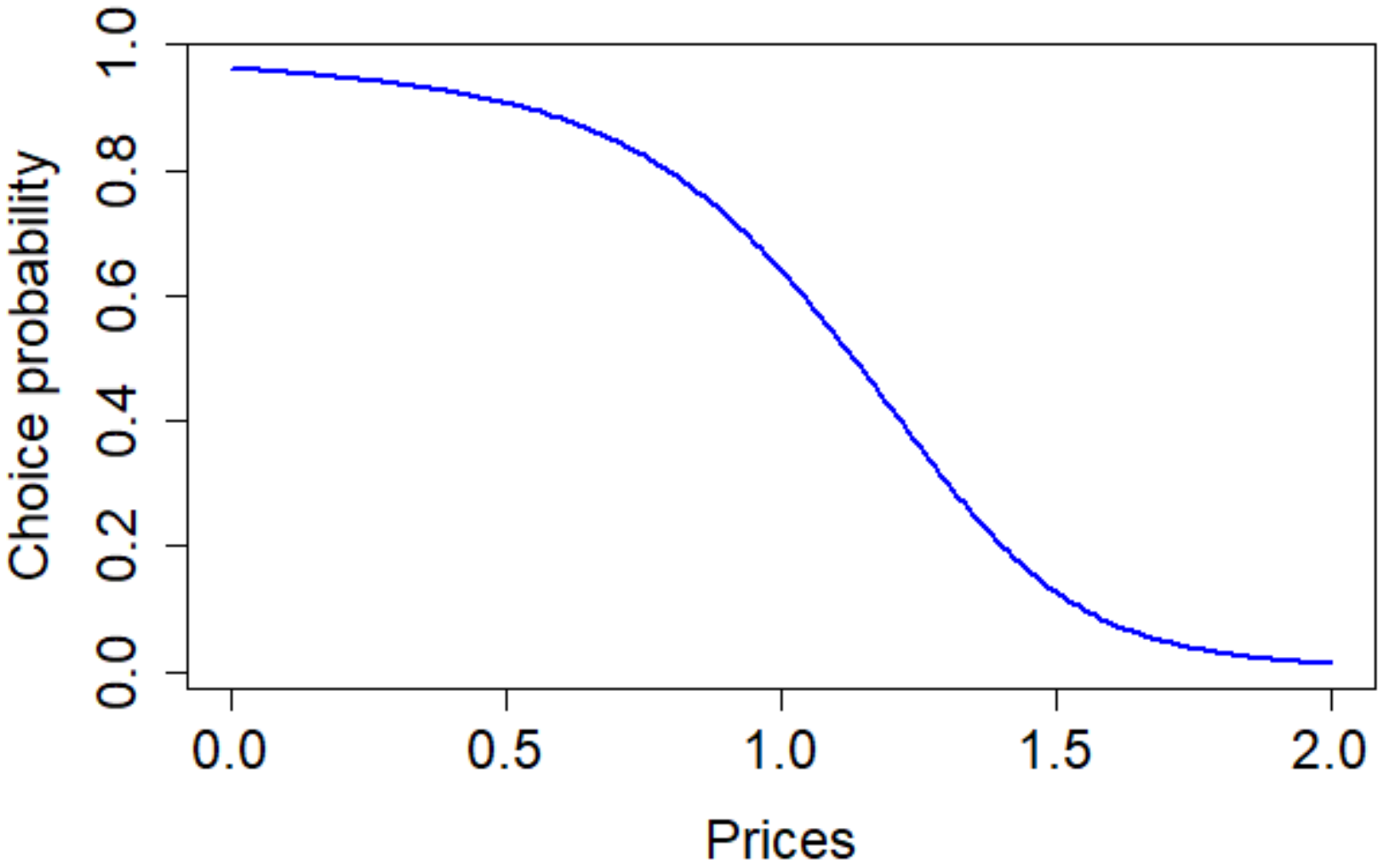
Aggregate choice probability

- From now on, I will call this weighted average of choice probabilities "aggregate choice probability" and denote it by $Pr(y = KB|P)$ (no subscript k).

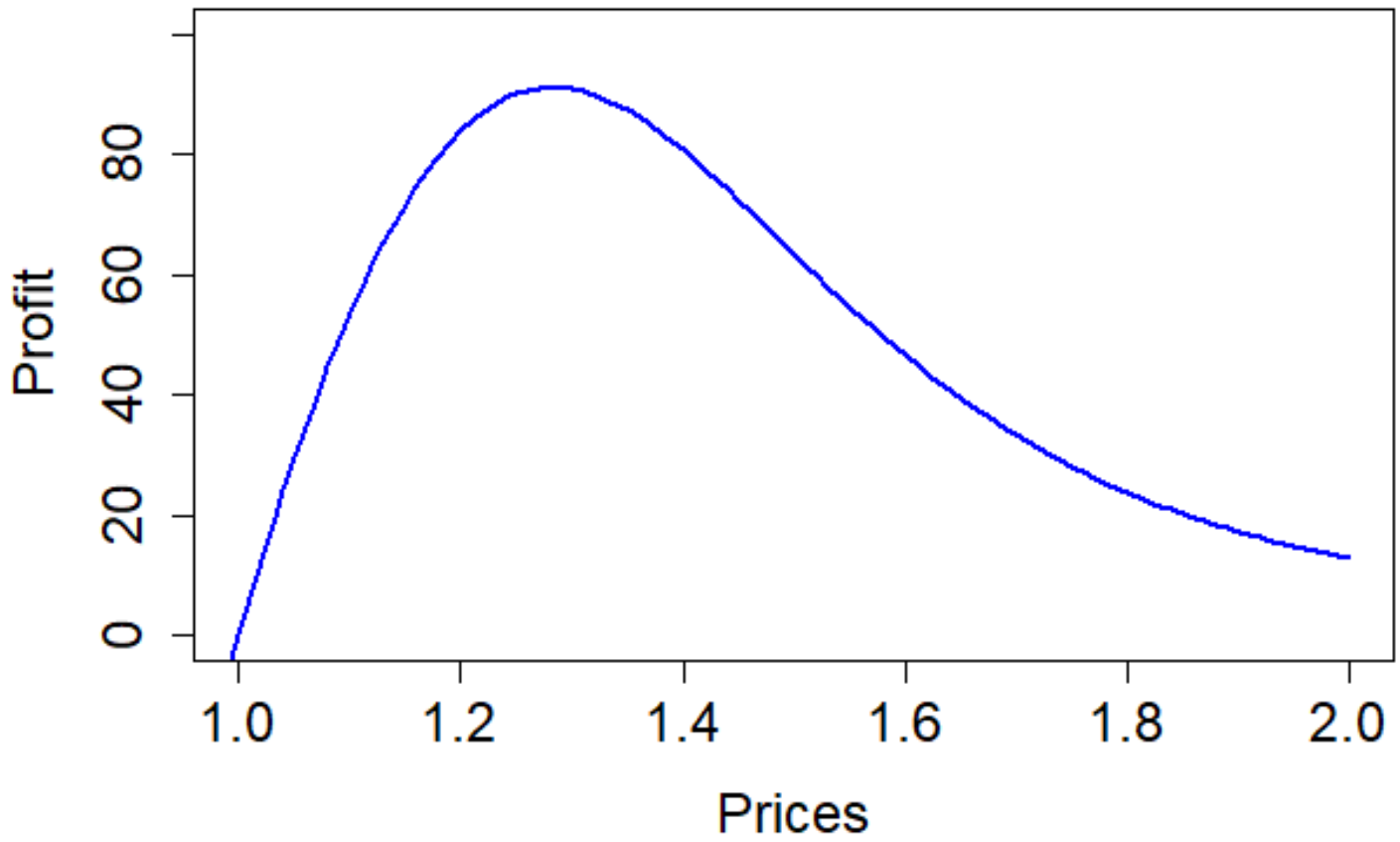
$$Pr(y = KB \mid P) = \sum_{k=1}^{K} w_k Pr_k(y = KB \mid P).$$

- Intuitively, this corresponds to the choice probability of an average person in the market. This is a measure of choice probability at the market level as a whole, in the presence of heterogeneous consumers.

Aggregate choice probability for KB



Profit



- At $P = 1.28$, the profit is maximized at 91.05.

Note: Elasticity under K-mean model

- Note that the simple elasticity formula from the homogeneous logit case does not directly apply.

$$\frac{\frac{\partial Pr(y=KB)}{Pr(y=KB)}}{\frac{\partial P^{KB}}{P^{KB}}} = -\beta_1 P^{KB}(1 - Pr(y = KB)).$$

- Because we assume segment-specific parameter values, this now holds *for each segment*.
- In order to calculate price elasticity at market-level (aggregated across segments), we cannot use this formula anymore.

Note: Elasticity under K-mean model

- The aggregate elasticity under segmentation is a bit involved:

$$\frac{\frac{\partial Pr(y=KB)}{Pr(y=KB)}}{\frac{\partial P^{KB}}{P^{KB}}} = -\frac{P^{KB}}{Pr(y=KB)} \sum_{k=1}^{K} w_k \beta_{1k} Pr_k(y=KB)(1 - Pr_k(y=KB)).$$

- Subscript k represents segment-specific objects, and I denote the proportion of each segment by w_k. $Pr(y = KB)$ is the aggregate choice probability defined in slide 31 (dependence on P suppressed for brevity).

$$Pr(y=KB) = \sum_{k=1}^{K} w_k Pr_k(y=KB).$$

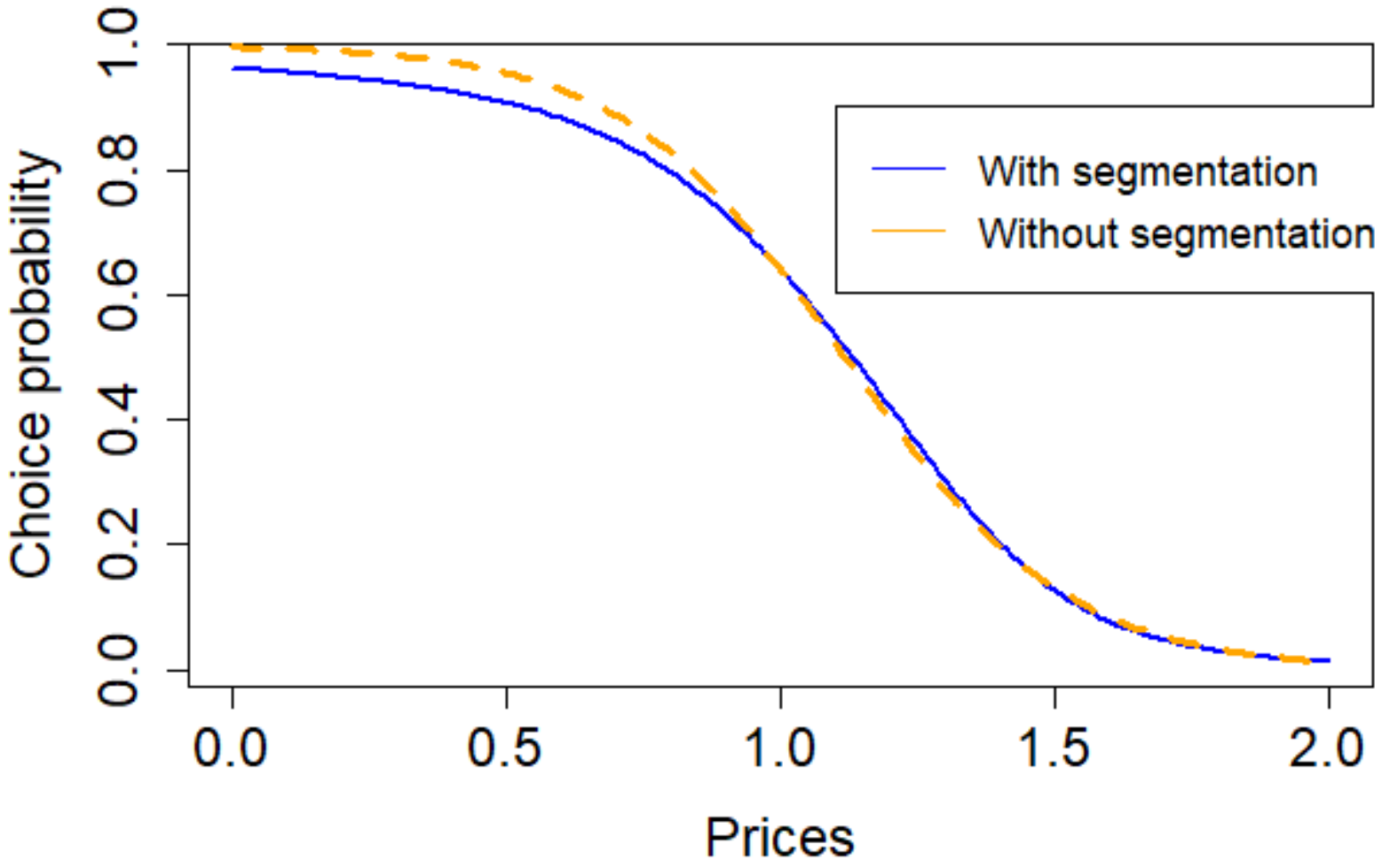
Note: Elasticity under K-mean model

- Here's the formula of cross elasticity - this is "percentage change of demand for KB when price of KR moves by one percent".

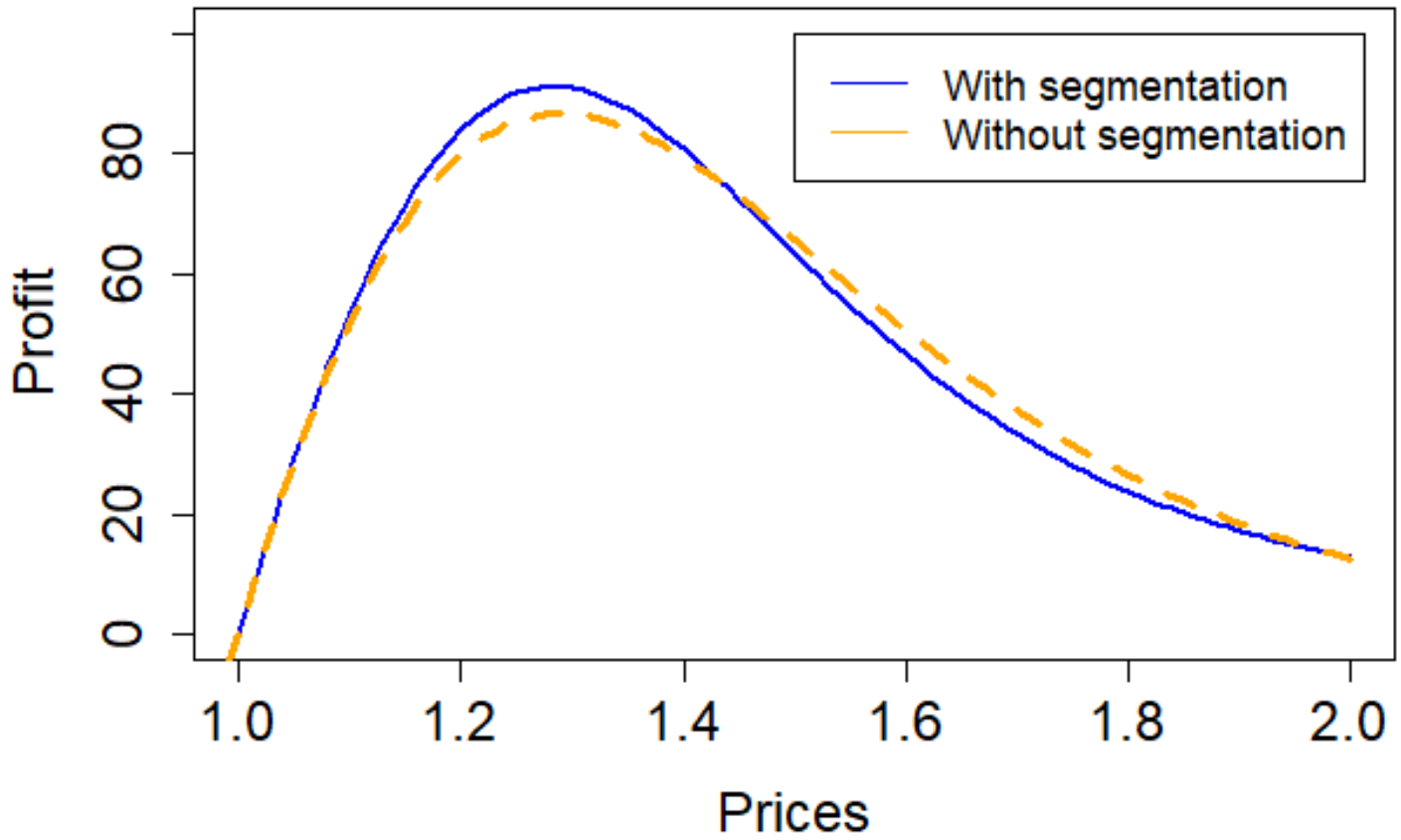
$$\frac{\frac{\partial Pr(y=KB)}{Pr(y=KB)}}{\frac{\partial P^{KR}}{P^{KR}}} = -\frac{P^{KR}}{Pr(y=KB)} \sum_{k=1}^{K} w_k \beta_{1k} Pr_k(y=KB) Pr_k(y=KR),$$

- Double-check to make sure that you don't get messed up between KB and KR!
- Cross-elasticity between other products is defined analogously (just replace "KB" and "KR" with other products).

Compare with no segmentation



Compare with no segmentation



- Optimal prices almost identical ($P = 1.29$ without segmentation).

Positioning and targeting

- With one product and one price, understanding segmentation may result in small change in profit. Even if we know the heterogeneity of demand, we cannot do much with it if we don't do targeting (single price) or positioning (single product).
- With product positioning (Project 2) or targeted pricing, we can materialize the gains from understanding the demand better. Let's consider how we can use the estimated demand to exercise targeting.

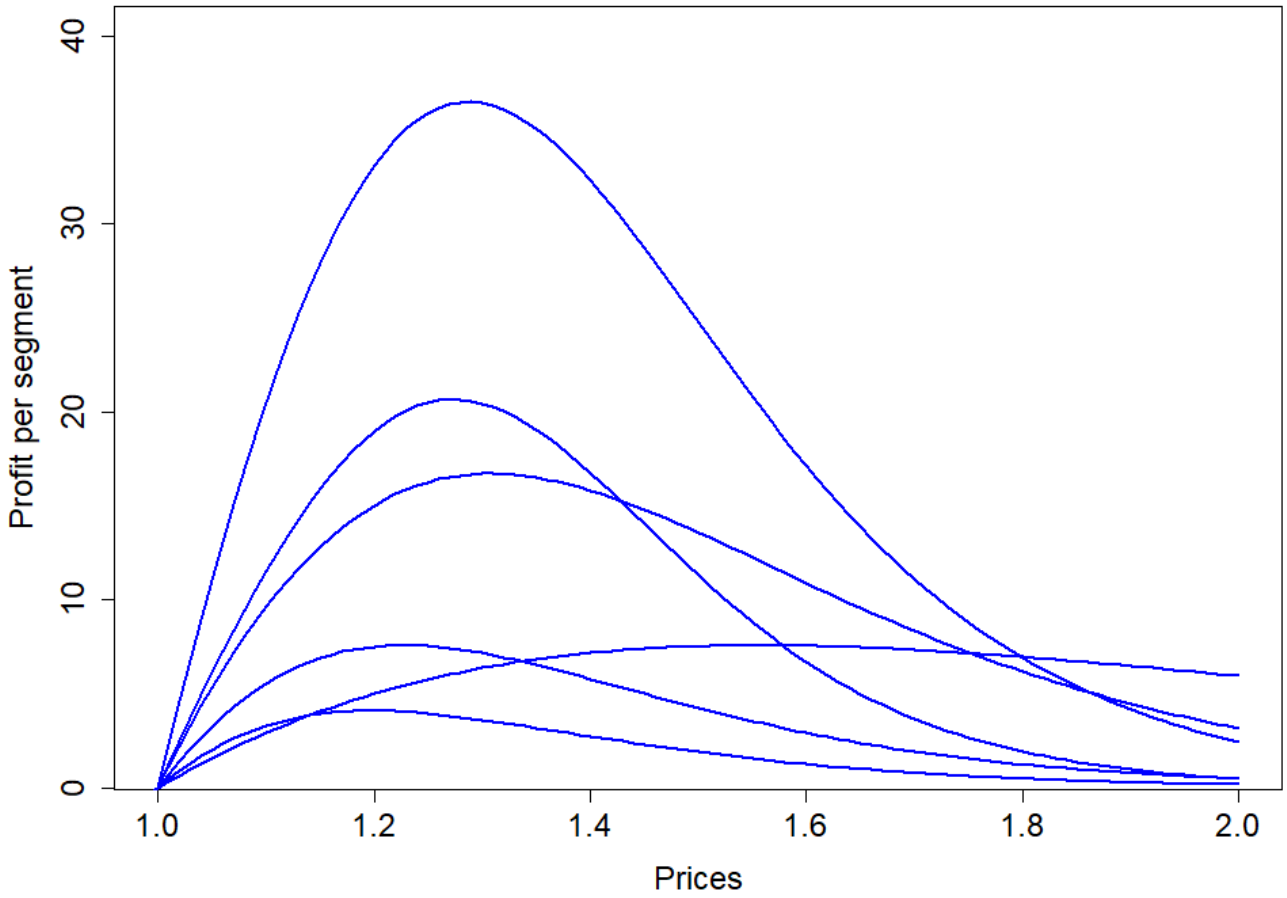
Profit per segment

- In order to consider optimal targeting, let's calculate "profit per segment".
- i.e. If we set different prices across different segments of consumers, we are essentially splitting the whole market into "market for each segment". What is the profit from each submarket?
- It is

$$\underbrace{\text{Number of consumers} \times w_k}_{\text{Number of segment } k \text{ consumers}} \times Pr_k(y = KB \mid P_k) \times (P_k^{KB} - \text{unit cost}).$$

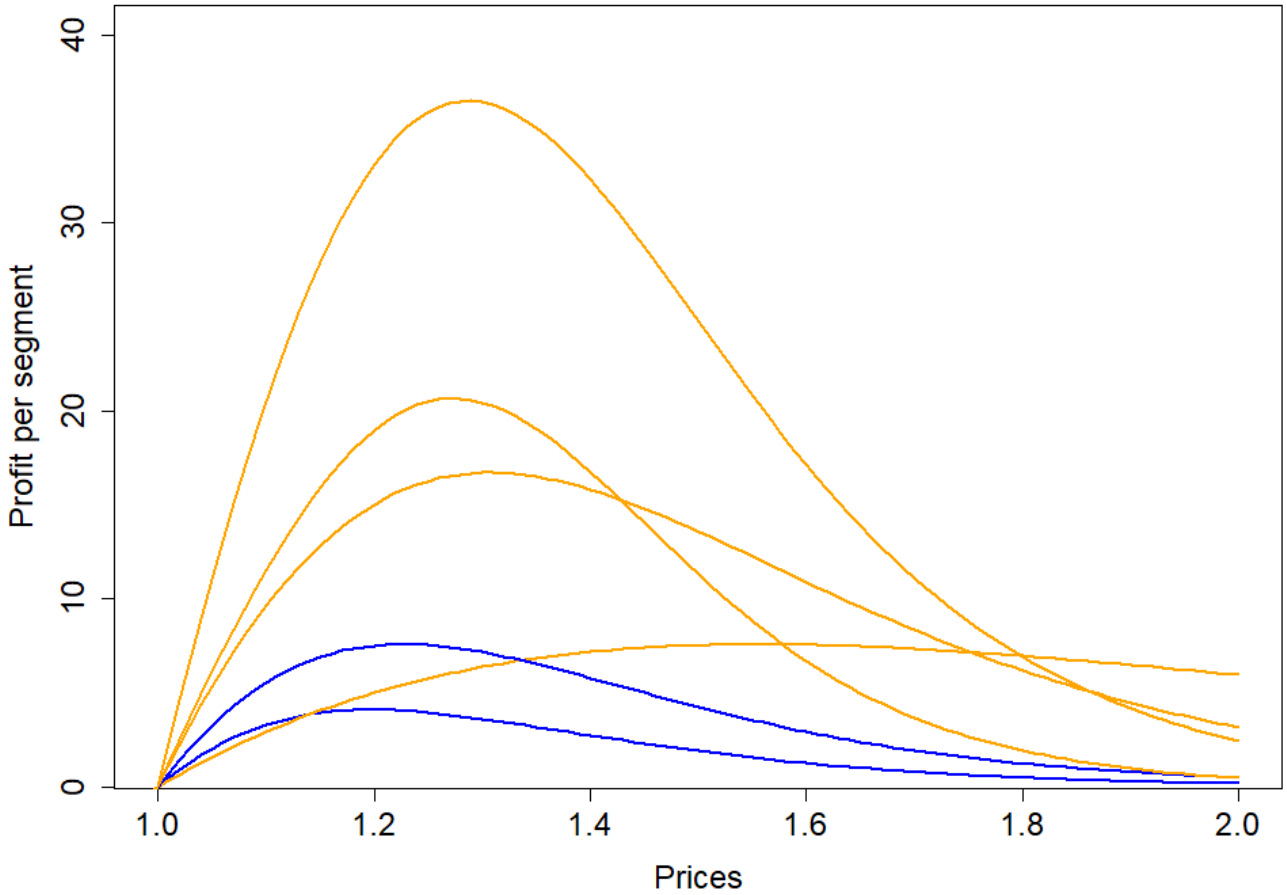
- w_k is the proportion of segment k consumers in the market.

Profit per segment



- Each line corresponds to profit from each segment. Segments with more consumers generate higher profit.

Profit per segment



- Optimal price for orange segments (2,3,4,6) around \$1.3, whereas that for blue segments (1,5) around \$1.2.

Targeted pricing

- What if we set \$1.3 as a regular price and send targeted coupons of \$0.1 to consumers from segments 1 and 5?
- Then the profit is

$$\begin{aligned}\text{Profit} = &\text{ Profit from seg 1 } (P^{KB} = 1.2) \\ &+ \text{Profit from seg 2 } (P^{KB} = 1.3) \\ &+ \text{Profit from seg 3 } (P^{KB} = 1.3) + ...\end{aligned}$$

Targeted pricing

```
#What if we set 1.30 as a regular price and offer a coupon of 10 cents to segments 1 and 5?
profitseg=profit1[pricespace==1.2]+profit2[pricespace==1.5]+profit3[pricespace==1.3]+
  profit4[pricespace==1.3]+profit5[pricespace==1.2]+profit6[pricespace==1.3]
#Profit=92.767
```

- Targeting increases profit by 1.9%.
- Depending on how we look at it, it could either seem small or large - I personally find it reasonably large.
- If we say "by just sending a coupon e-mail to some consumers we get 1.9% profit increase", then it sounds great.
- If we say "segmentation can increase profit by only 1.9%", then note that this is a pretty coarse targeting - we can certainly go finer.

Multinomial logit models with consumer characteristics

- We have studied how "estimating multinomial logit models segment-by-segment" improves our pricing strategy.
- One potential downside: its specification choice lacks good metrics we can rely on.
  - What demographic characteristics should we include? Which ones are more informative for segmentation?
  - How many segments should we choose?
- Let's turn to the other approach to incorporate heterogeneity - "estimating multinomial logit using all observations, but include consumer characteristics explicitly in the model". This approach provides a way to deal with the issue better.

Multinomial logit models with consumer characteristics

- Recall that the probability that product j gets selected is given by the following formula.

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

  where j' includes all products available in the market.
- Earlier, we assumed that β_0^j varies across clustered segments.

Multinomial logit models with consumer characteristics

- Suppose that we don't cluster the data beforehand, but instead let β_{0k}^j depend directly on each consumer's observed characteristics.
- For example, we could model β_{0k}^j as follows.

$$\beta_{0k}^j = \beta_{00}^j + \beta_{0,fs}^j \times fam\_size_k + \beta_{0,fa}^j \times fem\_age_k + ...$$

- We can do it with β_1 analogously, so we focus on β_0^j here.
- First, let's see how to estimate it in R.

Implementation in R

```
#Load demographic data
demo=fread("demo.csv",stringsAsFactors = F)
#Merge with original data
data=merge(data,demo,by="id")

#Run mlogit-gmnl
mlogitdata=mlogit.data(data,id="id",varying=4:7,choice="choice",shape="wide")

mle=gmnl(choice~price|
        fam_size+fem_age+fem_educ+fem_smoke+male_age+male_educ+male_smoke+dogs,
                                                        data=mlogitdata)
```

- I merge the two data sets (this time without clustering), and run gmnl with added demographic variables.
- In gmnl, we add demographic variables after "|".

Merged data

```
> head(data)
  id week trip price.0 price.KB price.KR price.MB choice fam_size fem_age fem_educ fem_sm
1  1   96    1       0     1.43     1.43     1.43      0        3       4        4
2  2   89    5       0     1.43     1.43     1.32     KB        2       6        4
3  2  114    8       0     1.43     1.43     1.34     MB        2       6        4
4  2   94    6       0     0.90     0.89     1.43     KR        2       6        4
5  2   96    7       0     1.43     1.43     1.43      0        2       6        4
6  2   31    4       0     1.43     0.88     1.65     KB        2       6        4
```

Estimated parameters

```
Coefficients :
                  Estimate Std. Error  z-value  Pr(>|z|)
KB:(intercept)  7.064387   2.163401   3.2654 0.0010931 **
KR:(intercept) 11.414987   2.014420   5.6666 1.456e-08 ***
MB:(intercept)  6.196771   1.751188   3.5386 0.0004022 ***
price          -5.361612   0.529047 -10.1345 < 2.2e-16 ***
KB:fam_size    -0.055775   0.242283  -0.2302 0.8179306
KR:fam_size    -0.304255   0.236921  -1.2842 0.1990692
MB:fam_size     0.241113   0.194924   1.2370 0.2161025
KB:fem_age      0.277468   0.213596   1.2990 0.1939324
KR:fem_age     -0.276498   0.247225  -1.1184 0.2633937
MB:fem_age      0.287764   0.218694   1.3158 0.1882319
KB:fem_educ    -0.217186   0.112454  -1.9313 0.0534419 .
KR:fem_educ     0.151261   0.109884   1.3765 0.1686519
MB:fem_educ     0.111546   0.107631   1.0364 0.3000252
KB:fem_smoke    1.149326   0.496002   2.3172 0.0204940 *
KR:fem_smoke   -0.251948   0.882416  -0.2855 0.7752449
MB:fem_smoke    1.169372   0.494441   2.3650 0.0180283 *
KB:male_age     0.108775   0.260818   0.4171 0.6766383
KR:male_age    -0.086156   0.264323  -0.3259 0.7444626
```

- How do we interpret these coefficients?

A regression in logit model

- This is almost running a regression within a multinomial logit model.
- Each consumer k has a unique value of β_{0k}^{j}. We specify that

$$\beta_{0k}^{j} = \beta_{00}^{j} + \beta_{0,fs}^{j} \times fam\_size_k + \beta_{0,fa}^{j} \times fem\_age_k + ...$$

  i.e. β_{0k}^{j} is linear in demographic variables and we estimate the coefficients. We are predicting β_{0k}^{j} using observed demographics by a regression.
- As such, we can interpret the estimated coefficients in the exact same way as in a regression environment.

A regression in logit model

```
Coefficients :
                  Estimate Std. Error  z-value  Pr(>|z|)
KB:(intercept)  7.064387   2.163401   3.2654 0.0010931 **
KR:(intercept) 11.414987   2.014420   5.6666 1.456e-08 ***
MB:(intercept)  6.196771   1.751188   3.5386 0.0004022 ***
price          -5.361612   0.529047 -10.1345 < 2.2e-16 ***
```

- Baseline β_{00}^{j} corresponds to "our best guess of β_{0k}^{j} when all the demographic variables are zero".

A regression in logit model

KB:fam_size	-0.055775	0.242283	-0.2302	0.8179306
KR:fam_size	-0.304255	0.236921	-1.2842	0.1990692
MB:fam_size	0.241113	0.194924	1.2370	0.2161025
KB:fem_age	0.277468	0.213596	1.2990	0.1939324
KR:fem_age	-0.276498	0.247225	-1.1184	0.2633937
MB:fem_age	0.287764	0.218694	1.3158	0.1882319
KB:fem_educ	-0.217186	0.112454	-1.9313	0.0534419

- Each coefficient of demographic represents "how one unit change of that demographic variable impacts β_{0k}^{j}".

Side: Consumer-specific β_1

- We have focused on the case where β_{0k}^j is a function of characteristics. Sometimes, we may also want to include demographic-specific β_{1k}.

```
#Demographics in beta_0
mle=gmnl(choice~price|fam_size+fem_age,data=mlogitdata)

#Demographics in beta_1
mle=gmnl(choice~price+price:fam_size+price:fem_age,data=mlogitdata)

#Demographics in both beta_0 and beta_1
mle=gmnl(choice~price+price:fam_size+price:fem_age|
                              fam_size+fem_age,data=mlogitdata)
```

- In such cases, we directly include interaction terms right after the price.

Side: categorical variables

```
#Fixed effects (slow)
mle5=gmnl(choice~price|
            factor(fem_educ)+fam_size+fem_smoke+dogs,
                                                data=mlogitdata)
```

- So far I have been assuming that all variables represent actual numbers (age, family size, etc), but some demographic variables are categorical by nature (education level, occupation, etc).
- Categorical variables can be included as a collection of dummies. However, unlike "felm", dummies in "gmnl" are really slow. So I would recommend keeping few categorical variables.
- In this slide deck, I assume all variables represent actual numbers.

Specification choice

$$\beta^j_{0k} = \beta^j_{00} + \beta^j_{0,fs} \times fam\_size_k + \beta^j_{0,fa} \times fem\_age_k + ...$$

- To tune our model optimally, we need to determine what demographic variables to include, with what functional form (linear, quadratic, log, interaction, etc.).
- Because a regression-in-logit model borrows its specification from the regression environment, we can use metrics of model fit we are familiar in the regression environments.
- This is probably a good advantage of regression-in-logit approach: there is no similar metric available for Kmeans approach to guide us.

How to choose between different model specifications?

- Because we don't estimate the model by a regression (recall that this regression-like component is just part of the model), R^2 is not available. Nevertheless, t-test and F-test can be used to check if we should include an additional term in the model.
- If we compare a model of this form:

$$\beta^j_{0k} = \beta^j_{00} + \beta^j_{0,fs} * fam\_size_k + \beta^j_{0,fa} * fem\_age_k + \beta^j_{0,fe} * fem\_educ_k + ...$$

with this:

$$\beta^j_{0k} = \beta^j_{00} + \qquad\qquad \beta^j_{0,fa} * fem\_age_k + \beta^j_{0,fe} * fem\_educ_k + ...$$

- We can do t-test on $\beta^j_{0,fs}$ in the first model to see if adding family size has significant impact on the outcome.

How to choose between different model specifications?

- However, there are also cases in which t-test and F-test doesn't work.
- How do we compare this model

$$\beta^j_{0k} = \beta^j_{00} + \beta^j_{0,fs} * fam\_size_k + \beta^j_{0,fa} * fem\_age_k + ...$$

with this?

$$\beta^j_{0k} = \beta^j_{00} + \beta^j_{0,fs} * \log(fam\_size_k) + \beta^j_{0,fa} * \log(fem\_age_k) + ...$$

- t-test is used when we add a new variable to a given model. In this case, neither model is an extension of the other model.

Bayesian information criteria (BIC)

- When models are estimated with MLE, we can compare between non-nested models using "Bayesian information criteria".

$$BIC = \text{Number of parameters} \times \log(\text{Number of observations}) - 2\log(\text{Likelihood value at the estimated parameter}).$$

- The *lower* the BIC is, the better the model fit is.
- It is the same in spirit as adjusted R^2, but in the context of more general models. Higher likelihood values are better, and adding many parameters would be punished.

Benchmark: BIC of no segmentation model

```
#No demographics - benchmark
mle0=gmnl(choice~price,data=mlogitdata)
summary(mle0)
BIC0=log(length(data$id))*length(mle0$coefficients)-2*mle0$logLik$maximum[1]
BIC0
```

```
> BIC0
[1] 1060.024
```

BIC with β_{0k}^{j} linear in demographics

```
mle1=gmnl(choice~price|
        fam_size+fem_age+fem_educ+fem_smoke+male_age+male_educ+male_smoke+dogs,
                                                              data=mlogitdata)
summary(mle1)
BIC1=log(length(data$id))*length(mle1$coefficients)-2*mle1$logLik$maximum[1]
BIC1
```

```
> BIC1
[1] 1059.547
```

- Including demographics linearly improves BIC, but only marginally.

What if we drop male characteristics?

```
#Drop male side
mle2=gmnl(choice~price|
            fam_size+fem_age+fem_educ+fem_smoke+male_smoke+dogs,
          data=mlogitdata)
summary(mle2)
BIC2=log(length(data$id))*length(mle2$coefficients)-2*mle2$logLik$maximum[1]
BIC2
```

```
> BIC2
[1] 1053.809
```

- Age and education within household are highly correlated between female and male. Dropping male characteristics don't lose model fit (gain in BIC because we have less parameters).

What if we include log(characteristics)?

```
#Inside log
mle3=gmnl(choice~price|
            log(fam_size)+log(fem_age)+log(fem_educ)+fem_smoke+male_smoke+dogs,
          data=mlogitdata)
summary(mle3)
BIC3=log(length(data$id))*length(mle3$coefficients)-2*mle3$logLik$maximum[1]
BIC3
```

```
> BIC3
[1] 1044.194
```

- Nonlinearity improves model fit.

BIC with K-mean clustering?

- BIC compares across models that are estimated by *a single MLE*. In other words, in order to use BIC, we need to estimate all parameters with one MLE.
- When we studied K-mean clustering, we ran MLE multiple times (one for each segment) to get all the parameter values - hence BIC doesn't really apply to the K-mean model.
- In the asynchronous video, I discuss one possible approach to estimate a K-mean model with a single MLE, making it possible to use BIC there.

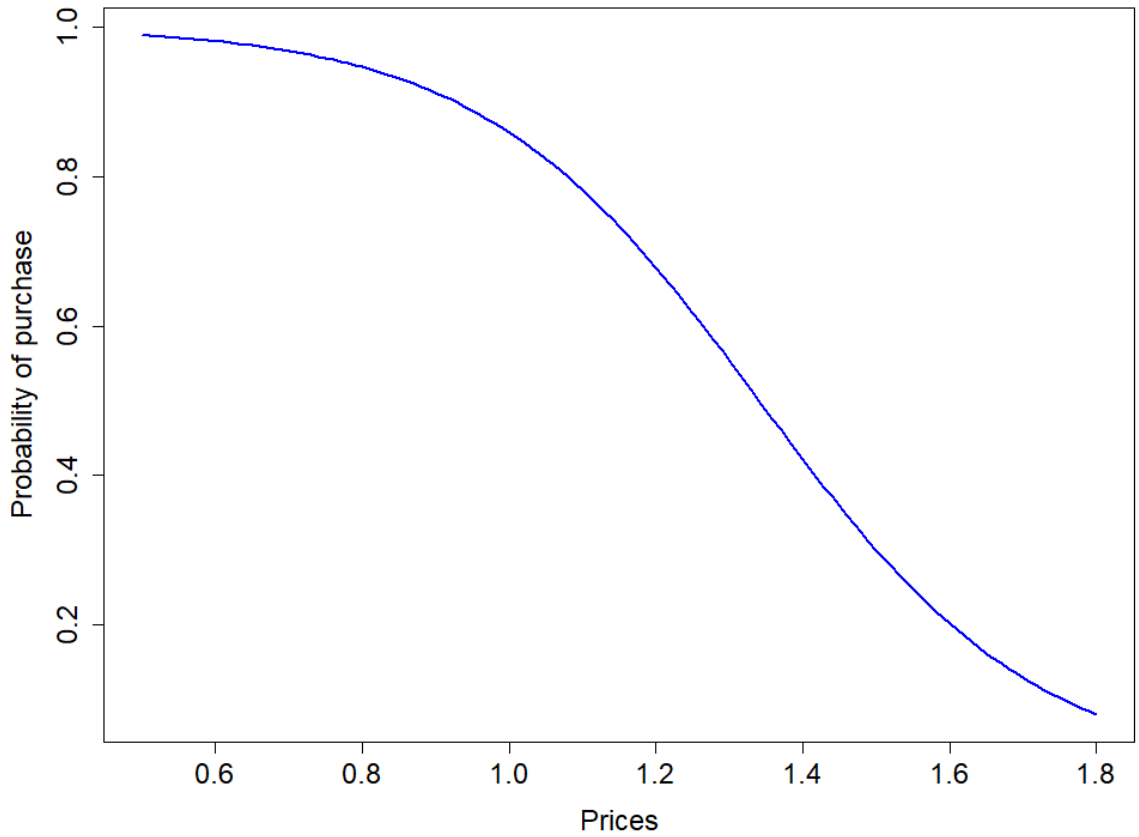
Simulating profit with regression-in-logit models

- With the estimated model, let's evaluate the predicted demand. As usual, we first code the choice probability of each consumer k as a function to be called later.
- Now $P_k(y = j \mid P)$ is a function of consumer characteristics, in addition to the prices and the model parameters. Hence our R function takes those variables as inputs.
- Then by plugging in each consumer's demographic realization, we can predict his/her choice probability.

Coding the estimated demand

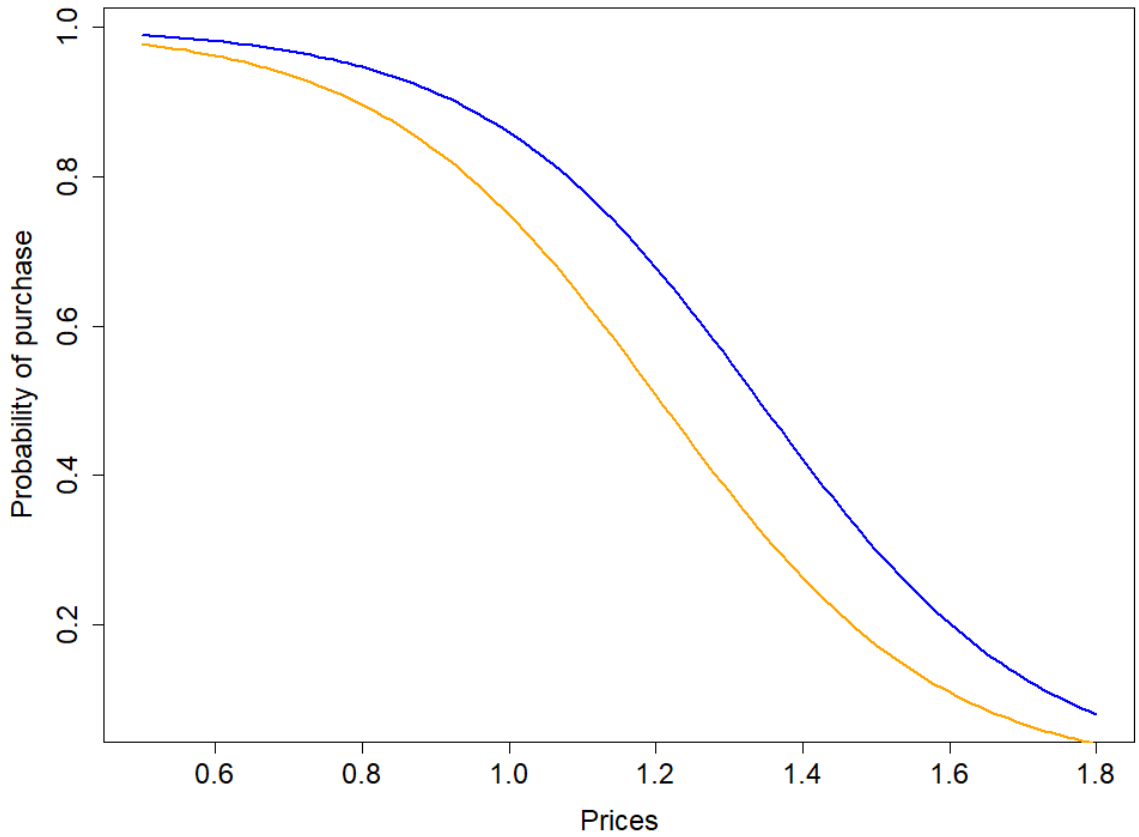
```
#Simulate demand based on log-formulation
demand=function(priceKB,priceKR,priceMB,fam_size,fem_age,fem_educ,fem_smoke,male_smoke,dogs,para){

  #Define beta_0's for each product as a function of consumer characteristics
  beta0KB=para[1]+para[5]*log(fam_size)+para[8]*log(fem_age)+para[11]*log(fem_educ)+para[14]*fem_smoke+
    para[17]*male_smoke+para[20]*dogs

  beta0KR=para[2]+para[6]*log(fam_size)+para[9]*log(fem_age)+para[12]*log(fem_educ)+para[15]*fem_smoke+
    para[18]*male_smoke+para[21]*dogs

  beta0MB=para[3]+para[7]*log(fam_size)+para[10]*log(fem_age)+para[13]*log(fem_educ)+para[16]*fem_smoke+
    para[19]*male_smoke+para[22]*dogs

  beta1=para[4]

  #Define choice probability of KB
  prob=exp(beta0KB+beta1*priceKB)/
    (1+exp(beta0KB+beta1*priceKB)+exp(beta0KR+beta1*priceKR)+exp(beta0MB+beta1*priceMB))
  return(prob)
}
```

The estimated demand for KB



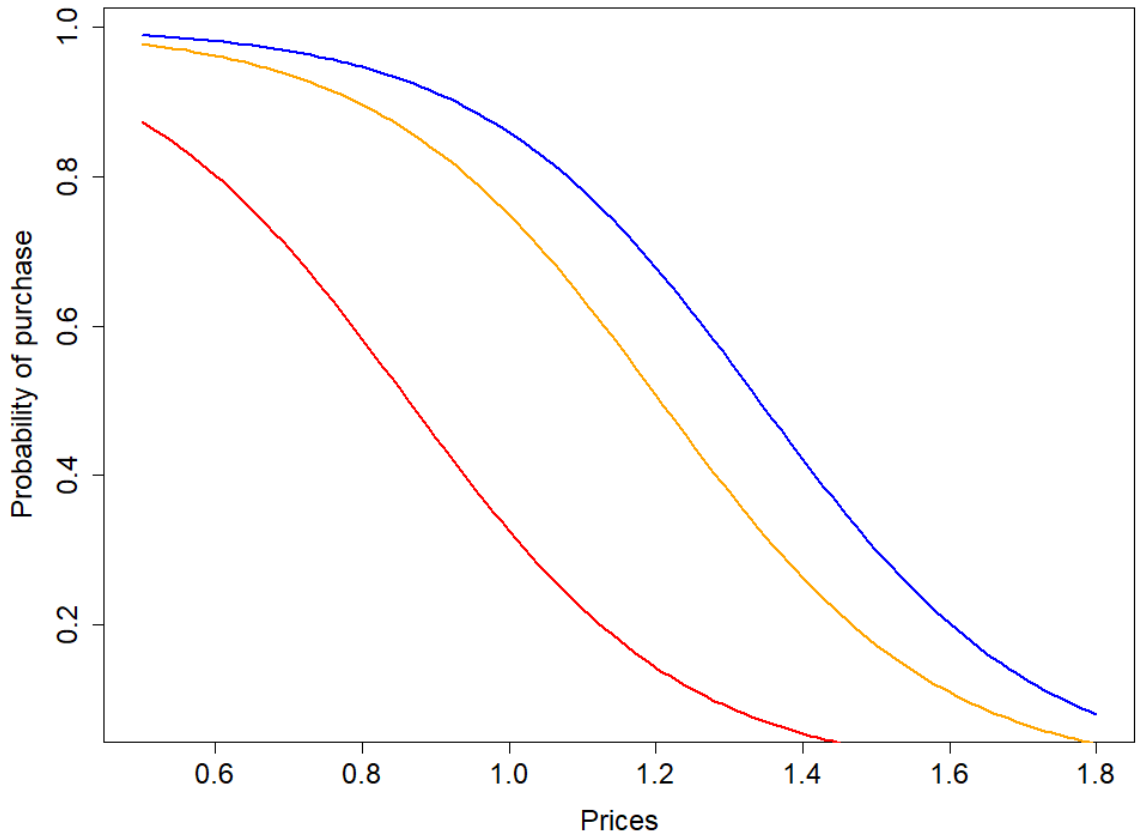
- Blue = Demand of "family size=2, female age=5 (50), female education=4, female smoke=1, male smoke=0 and dog=1".

The estimated demand for KB



- Orange= Demand of "family size=4, female age=2 (20), female education=3, female smoke=1, male smoke=0 and dog=1".

The estimated demand for KB



- Red = Demand of "family size=2, female age=2 (20), female education=3, female smoke=0, male smoke=0 and dog=0".

Simulate profit without targeting

- Let's use those choice probabilities to maximize profit. Suppose again that we only set one price (no targeting).
- Just like the K-means case, we first calculate the market-level demand by calculating the aggregate choice probability (= average of demographic-specific choice probabilities) and multiplying it with the number of consumers.
- One difference from the K-means case: calculation of the aggregate choice probability.

Calculate aggregate choice probability

- In the K-means clustering case, the aggregate choice probability is the weighted average of clustered-segment-specific choice probability. i.e. the averaging is over (say) 6 choice probabilities.
- Now we don't have any clustered segments. Every single consumer's choice probability differs from anyone else (unless two people share identical demographics). Hence the averaging needs to be over *individual consumer* now.

Calculate aggregate choice probability

- Aggregate choice probability now takes the following form:

$$Pr(y = KB \mid P) = \frac{1}{K} \sum_{k=1}^{K} Pr_k(y = KB \mid P).$$

- k now represents each consumer (K is now the number of consumers in the data), and Pr_k is the choice probability evaluated with the demographics of that consumer k.
- Hence the key is to calculate $Pr_k(y = KB \mid P)$ for *every* consumer in the data. Once it's done, the rest is just a simple average.

Demographic data

id	fam_size	fem_age	fem_educ	fem_smoke	male_age	male_educ	male_smoke	dogs
1	3	4	4	1	3	6	0	1
2	2	6	4	0	7	9	0	1
3	2	6	4	0	6	4	0	0
4	3	3	6	0	3	4	0	1
6	4	3	5	0	3	5	0	0
7	2	3	4	0	7	9	0	0
8	2	5	5	1	6	3	1	1
9	1	6	4	0	7	9	0	0

- For each consumer in our demographic data, we apply our "demand" (the ratio of exponentials) function (defined in slide 67) to compute the predicted choice probability, $Pr_k(y = KB \mid P)$.
- Usually, a simple for-loop would do.

Choice probability for each consumer k

```
> demandmat
        Pr(y=KB)
1: 0.39461203
2: 0.28638551
3: 0.19198507
4: 0.08348795
5: 0.06075368
6: 0.07736876
7: 0.55473196
8: 0.16979198
```

Elasticity with regression-in-logit approach

- Just like the case of aggregate choice probability, calculation of elasticity with regression-in-logit requires averaging over all consumers. Recall the definition of own-elasticity in the case of heterogeneous consumers:

$$\frac{\frac{\partial Pr(y=KB)}{Pr(y=KB)}}{\frac{\partial P^{KB}}{P^{KB}}} = -\frac{P^{KB}}{Pr(y=KB)}\frac{1}{K}\sum_{k=1}^{K}\beta_1 Pr_k(y=KB)(1-Pr_k(y=KB)).$$

- Note that k now represents each consumer: we now calculate inside the summation *for each individual*, and take its average.
- Cross elasticity is analogous (see Kmeans case for expression).

Targeting with regression-in-logit approach

- Targeted pricing with a regression-in-logit model is mostly identical to Kmeans case. We can draw figures similar to the ones we used in Kmeans case (I will skip them because they appear repetitive).
- However, with regression-in-logit models, every consumer's preference is different from one another - no clustering. Hence, a few things to keep in mind.
- With Kmeans segmentation, targeting means "targeting segments". Here, in principle we can offer different prices across every single consumer (whether we want it in practice is a different story).
- From data analysts' perspective, we can now use "distribution of willingness to pay" as a new metric.

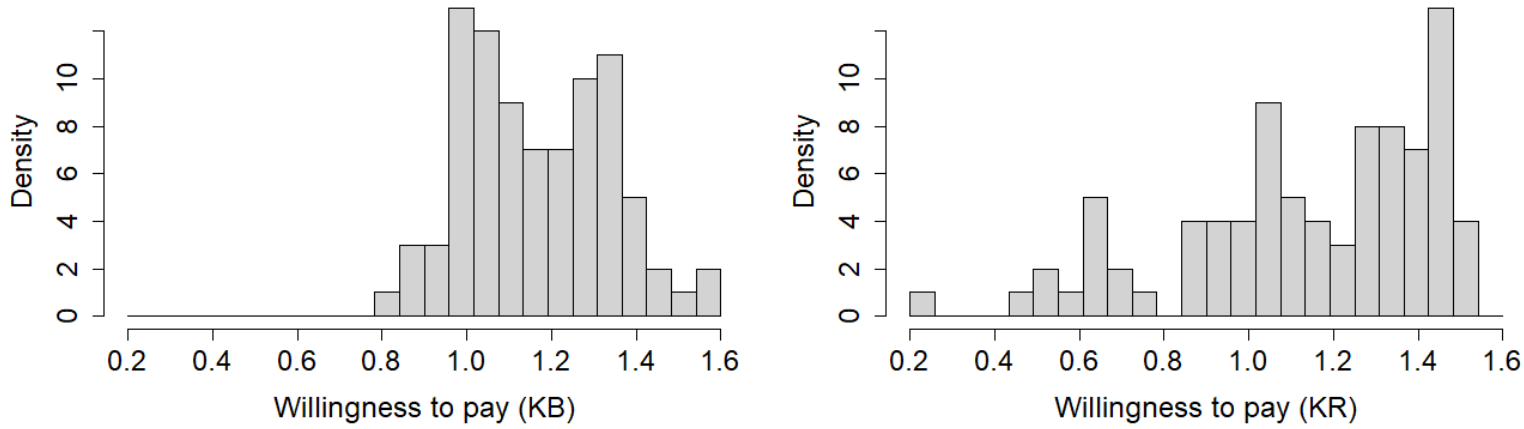
Representing willingness to pay

- Consumer willingness to pay represents how much the product is worth to each consumer, in dollar term.
- In fact, one advantage of logit models is that they have a simple form to represent willingness to pay as follows:

$$\text{Willingness to pay of consumer k for product j} = -\frac{\beta_{0k}^{j}}{\beta_1}.$$

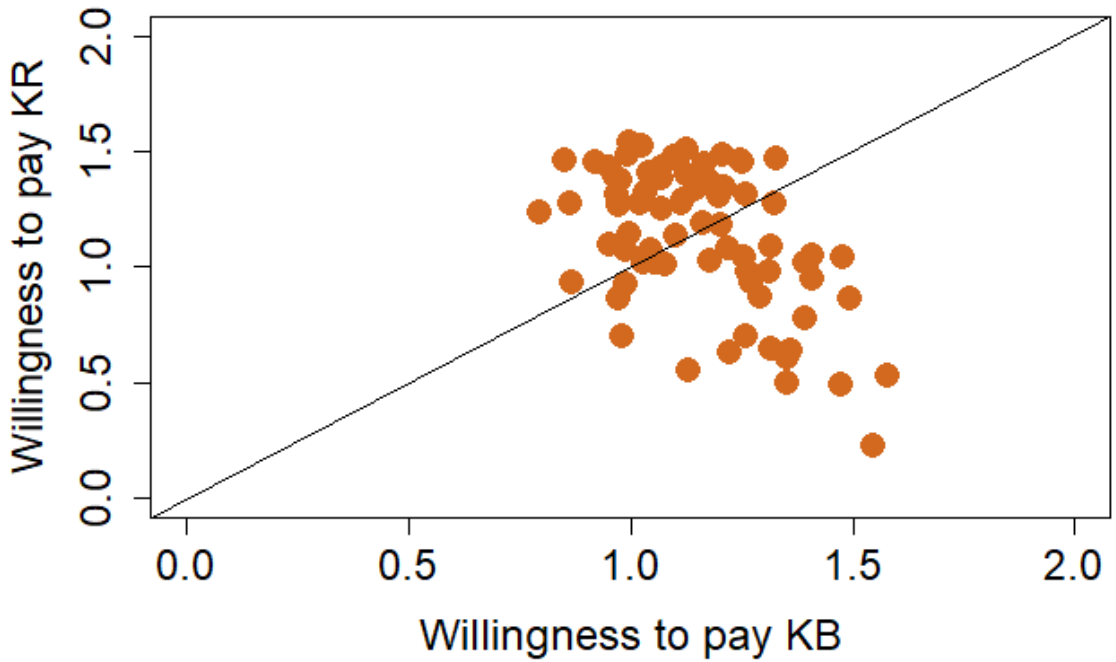
- For now, let's just memorize this - we discuss the derivation in more detail in the online content.

Distribution of willingness to pay



- Because we have estimated consumer-specific β_{0k}^{j} for each consumer k and for each product j, we know each consumer's willingness to pay for each product.
- Preference for KB and KR indeed looks different.

Relative distribution of willingness to pay



- WTPs are negatively correlated - some consumers prefer KB or KR and vice versa.
- This would likely provide a good idea of who to offer a targeted promotion.

Representing willingness to pay

- The distribution of willingness to pay gives you a great visualization tool to present the value of your product (to your boss, to your manager, etc.). Literally, it is the distribution of your product's worth among consumers.
- By comparing willingness to pay across products, we can also gain new insights about competitive structures.

Cluster or regression-in-logit?

- Now we have seen two different approaches to incorporate demographics. An obvious question is, which one should we use?
- Again, models are a collection of assumptions. The answer to this question depends on how well the model assumptions fit with the environment studied.
- Let's compare the two approaches, with particular attention to the implicit assumptions we are imposing.

Cluster or regression-in-logit?

- With K-mean clustering, we assume that β_{0k}^j and β_{1k}^j can vary across segments, but those who belong to the same segment have the same parameters.
- If we use small number of segments, each segment tends to cover a large proportion of consumers - those with really different demographics may be assigned with the same parameters.
- In general, we can make the model more flexible by increasing the number of segments. But remember the finite-sample issue - by increasing the number of segments, each segment contains less consumers. At some point, the estimate for each segment becomes unreliable.

Cluster or regression-in-logit?

- With regression-in-logit approach, we are imposing the assumptions we used in the regression environment: β_0^j can be expressed as a linear function of demographics.
- Everyone is different from one another - no clustered segment exists. However, the way β_0^j varies across demographics depends on our functional form assumption (e.g., linearity)
- In general, we can increase flexibility by adding more terms to the expression (higher-order terms, interaction terms, etc). But adding more terms will make the estimates less reliable (just like the regression case).

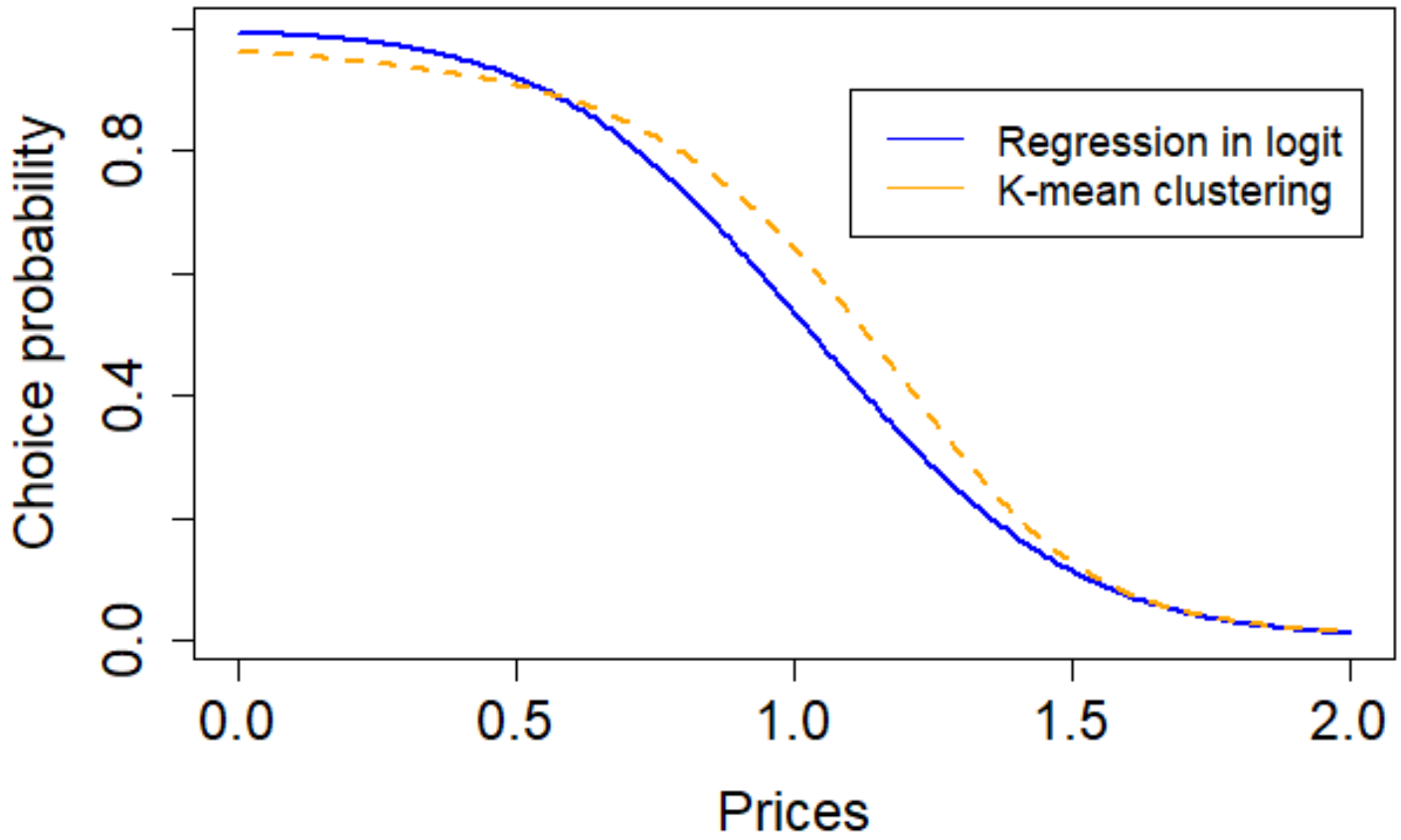
Cluster or regression-in-logit?

- Ultimately, K-mean approach and regression-in-logit approach are trying to add flexibility by imposing different kind of assumptions, facing the same "flexibility-reliability" trade-off.
- This is why we cannot really say that one is superior to the other - which one is superior depends on how suitable those assumptions (discrete clustered segments vs functional form) are to each environment we study.
- These two models are hence complements and they jointly cover most real environments.

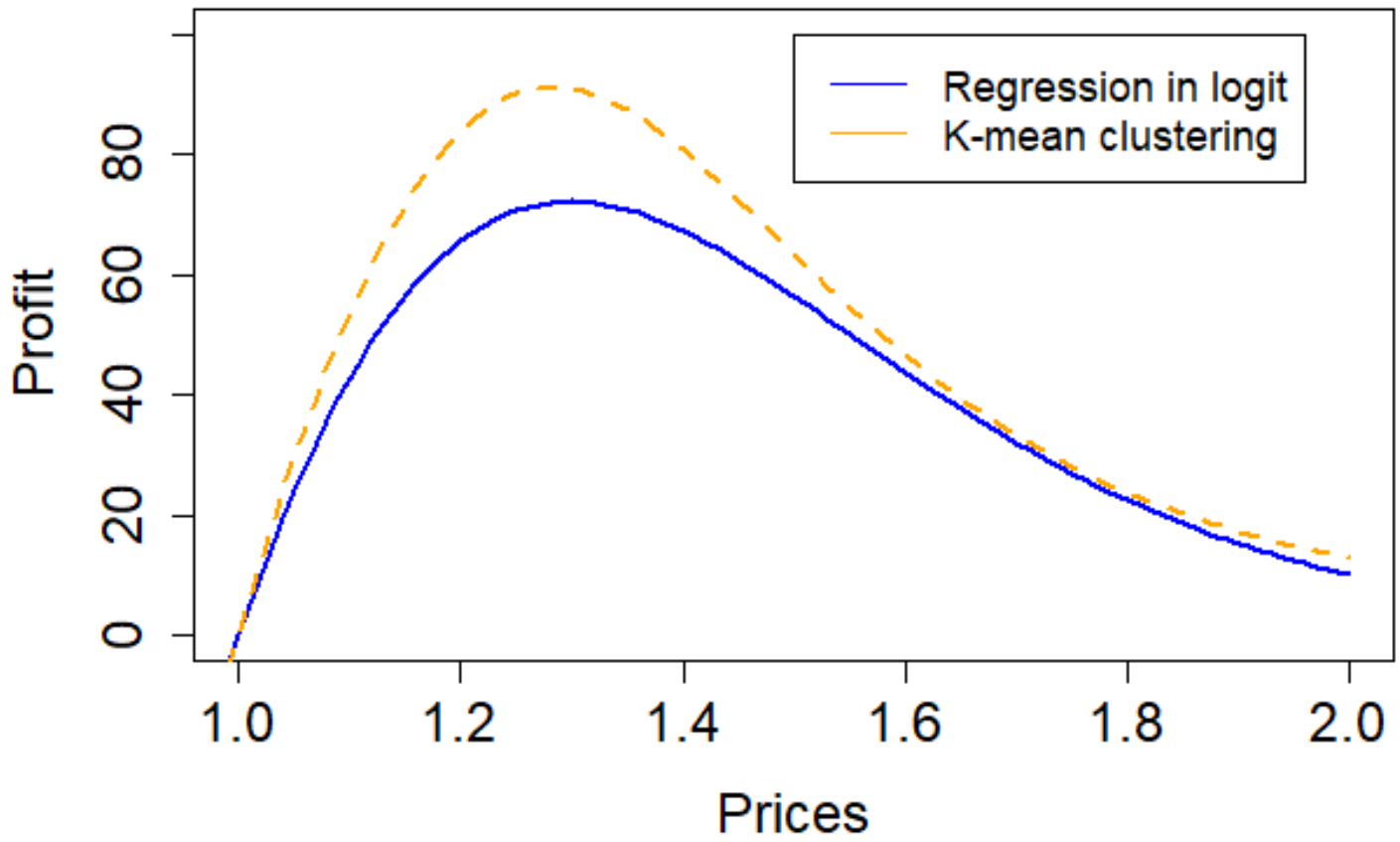
Summary

- We have covered two approaches to allow for heterogeneous demand in a multinomial logit framework.
- We can cluster observations based on the observed characteristics and estimate a multinomial logit model for each of them. Alternatively, we can include characteristics explicitly as a part of the model.
- These models are suitable for different purposes and jointly offer broad coverage of most real environments.
- In asynchronous video, I discuss some advanced materials about these models, including a way to use BIC for K-means, and to combine the two models to make a single, most flexible model.

If time allows 1: Compare predicted aggregate choice probabilities between the two models



Compare profits



- Can we conclude which model is better from these figures?

Don't select a model based on its predicted profit

- There's a widespread misunderstanding that "a model that maximizes your predicted profit is the best model". This is wrong.
- Note that a demand model can be wrong in two ways - it either underestimates the demand, or overestimates it.
- If demand is overestimated, then the profit is overly inflated. The model is basically too optimistic.
- If we choose a model based on the predicted profit, then we are choosing an overly optimistic model, not the right model.
- As we all know, model selection needs to be done by statistical evaluation: what is the BIC of each model? What is the std.error of the estimates? t-test, F-test? etc.

If time allows 2: Aggregate choice probability at the store

$$Pr(y = KB \mid P) = \frac{1}{K} \sum_{k=1}^{K} Pr_k(y = KB \mid P).$$

- Earlier I defined the aggregate choice probability (= choice probability of an average consumer) as *the average choice probability of all consumers* in the data.
- This measure is useful in evaluating the average consumer preference in a broadly defined market (state, country,etc).
- Suppose that we want to find the profit-maximizing price *at the store we collected the data from*. Is this aggregate choice probability a good measure to rely on?

Aggregate choice probability at the store

$$Pr(y = KB \mid P) = \frac{1}{K} \sum_{k=1}^{K} Pr_k(y = KB \mid P).$$

- To maximize the profit at the store, we need aggregate choice probability *at that store*.
- Simple average won't work, because some consumers in the data visit the store more frequently than others. In other words, simple average equals the average choice probability in the store only when all consumers visit the store at equal frequency.

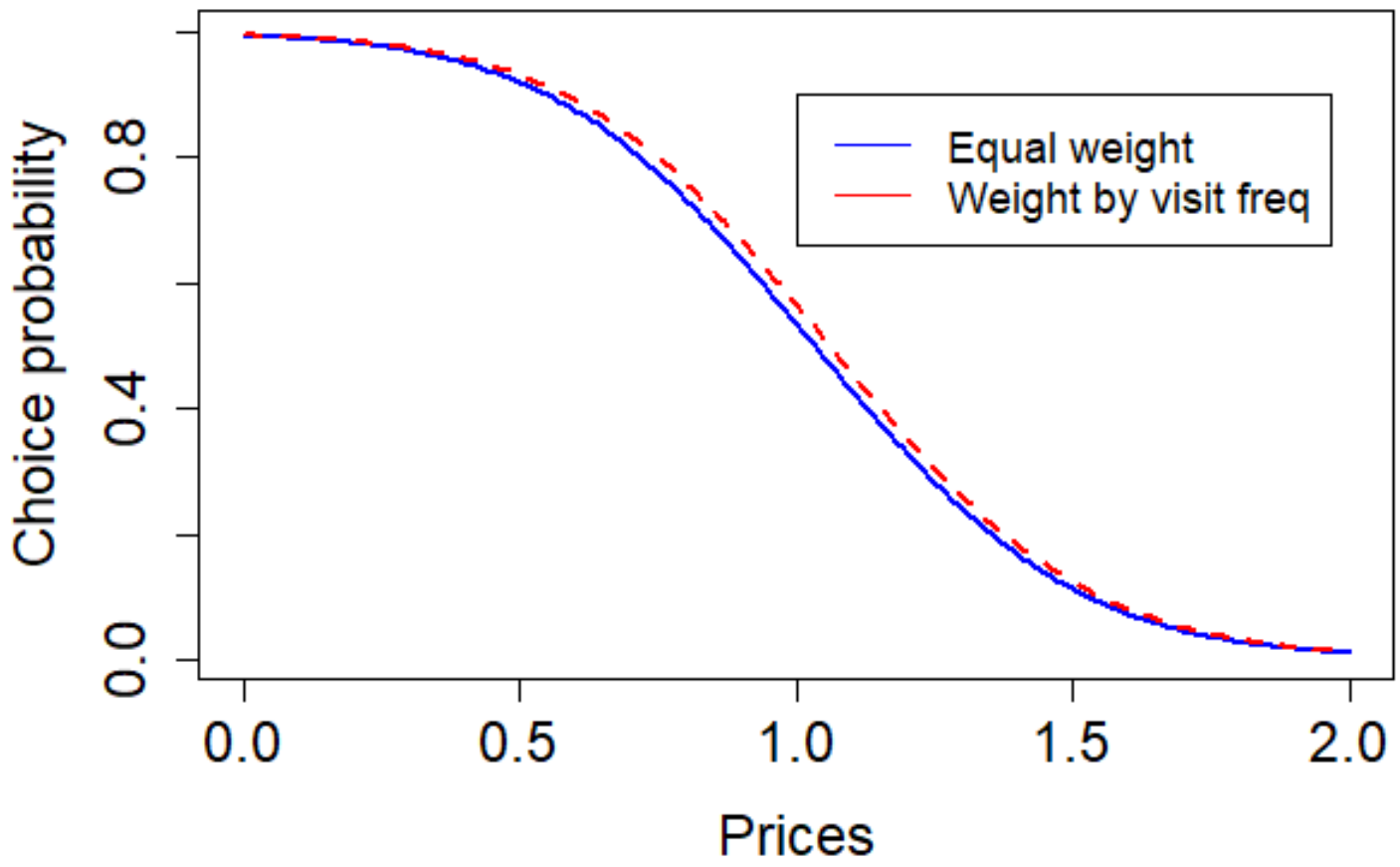
Aggregate choice probability at the store

- We need a weighted average of choice probabilities, where the weight equals the frequency of each consumer's visit. i.e.

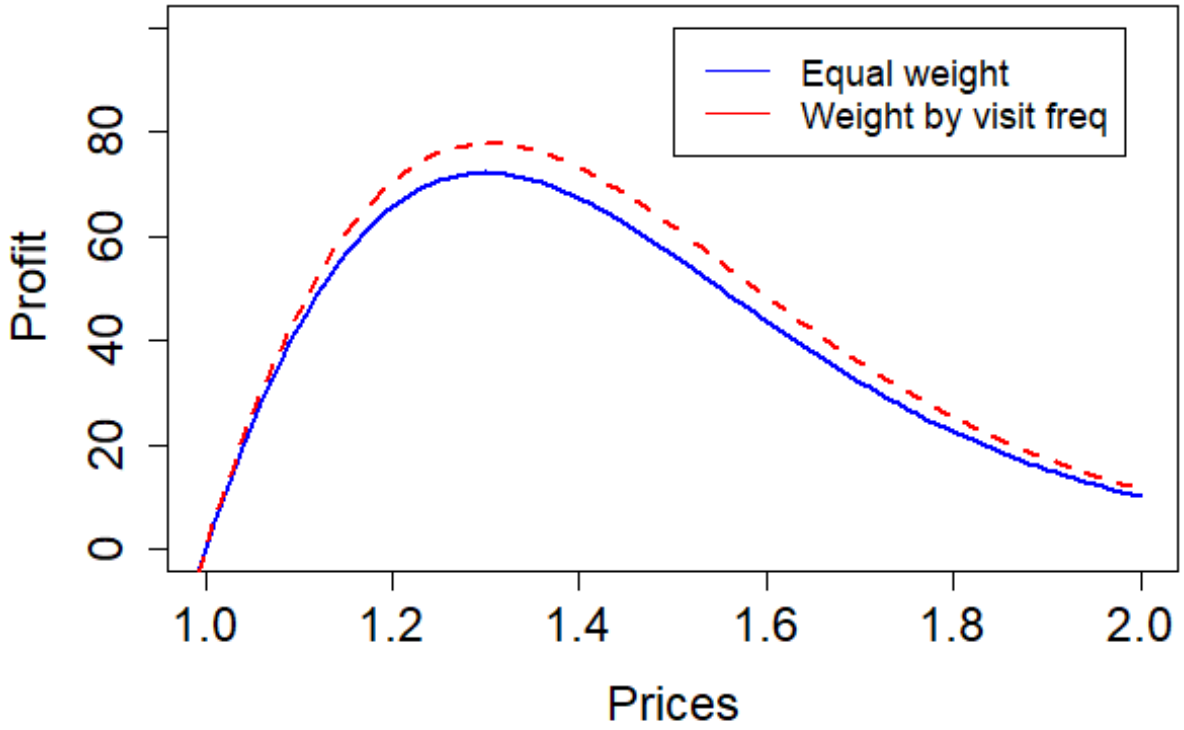
$$Pr(y = KB \mid P) = \frac{1}{\sum_{k=1}^{K} NT_k} \sum_{k=1}^{K} NT_k Pr_k(y = KB \mid P),$$

  where NT_k is the number of observations we have from consumer k in the data (= the number of k's visit to the store).
- This accounts for the fact that the population of consumers *we see in the store* is different from the population distribution of demographics we estimated.
- It is available as "seg.demo2" in the class R code.

Compare aggregate choice probabilities



Compare profit



- Profit is 77.8 if we set $P^{KB} = 1.3$.
- Price prediction is the same, but a reasonable difference in profit prediction.

Aggregate choice probability at national level

$$Pr(y = KB \mid P) = \frac{1}{K} \sum_{k=1}^{K} Pr_k(y = KB \mid P).$$

- This example is one illustration of a broader question.
- $Pr(y = KB \mid P)$ derived here represents the average preference of consumers *in the market from which we collected the data* (e.g. loyalty club card holder in Rochester, NY).
- If our objective is to launch the product at the national level, $Pr(y = KB \mid P)$ isn't the right metric - we need to account for the difference in demographic distribution between our data and our target market.

Aggregate choice probability at national level

- To account for such differences, we usually use the Census data.
- For each consumer in the data, check how many consumers with the same demographic realization exists - record the fraction as w_k^{data} (if three people share the same demographics, $w_k^{data} = 3/K$).
- From the Census, compute the fraction of people with such demographic background in the U.S - record it as $w_k^{national}$.
- Correct our estimate by taking the ratio of the two.

$$Pr(y = KB \mid P) = \frac{1}{K} \sum_{k=1}^{K} \frac{w_k^{national}}{w_k^{data}} Pr_k(y = KB \mid P),$$

Aggregate choice probability at national level

- The same caveat applies to K-mean clustering approach. There we need the proportion of consumers who belong to each segment to calculate aggregate demand.

```
> seg.share
   1    2    3    4    5
0.12 0.13 0.28 0.19 0.14 0.14
```

- If we predict the national-level demand, use the proportion computed from the Census, not from the data, as weights.

Summary

- We have covered two approaches to allow for heterogeneous demand in a multinomial logit framework.
- We can cluster observations based on the observed characteristics and estimate a multinomial logit model for each of them. Alternatively, we can include characteristics explicitly as a part of the model.
- These models are suitable for different purposes and jointly offer broad coverage of most real environments.
- In asynchronous video, I discuss some advanced materials about these models, including a way to use BIC for K-means, and to combine the two models to make a single, most flexible model.