Econ 8210: Problem Set #1

Peter W Newberry

DUE: October 13, 2025

INSTRUCTIONS: Feel free to work together but turn in your own write up. You can use any programming language you wish, but turn in a pdf of the code as well as a write-up of your solutions (tables, estimates, explanations, etc).

1 Data Import and Descriptive Statistics

There is a dataset on eLC called "IRI.csv". This data set contains product level sales/prices of cereals from 2008 to 2011 across a number of markets (stores) by week. That is, each observation is a product-store-week combination. Below is a table of each variable and a description of the variable. We are going to assume a market is made up of the cereals (and outside option) in a given store in a given week. Note that the week_id variable uniquely identifies the time period (i.e., it is not the 'week of the year'), so no need to use the month and year to define a market. In other words, the 'market' is defined by the store_id and week_id variables.

Table 1: Variables

Variable	Description
Name	desc
year	Year of data
month	Month of data
store_id	Store (use this as the market in estimation)
week_id	Week of the data (uniquely identifies the time period)
market_name	City where the store is located (do not use this as the market estimation)
state	State where the store is located
brand	Brand name of cereal
parent	Manufacturer of cereal
flavored	Dummy variable if it is 'flavored'
fortified	Dummy variable if it is 'fortified'
fiber	Grams of fiber in each serving
sugars	Grams of sugar in each serving
price	Price per ounce
quantity	Quantity of servings sold (in lbs)
puma	Identity code for the public use micro-data area (puma)
sugar_price	Price of raw sugar
M	Market size (including outside option)
segment	Segment the cereal belongs to.

1. Load the data. Create a table that has descriptive stats about the market. Specifically, what is the average/standard deviation/median/max number of cereal brands in a market? What is average/standard deviation/median/max number of cereal manufacturers in the market (hint: this will have little variation)? What is the average/standard deviation/median/max total sales and brand level sales in a market? What is the average/standard deviation/median/max price? What is the average/standard

deviation/median/max of the characteristics? Describe any interesting patterns in the data. Feel free to make the table however you wish, but i recommend that you put the variable on the left of the table and then the descriptive stats in columns going from left to right. NOTE: again we are considering the market as a store/week, NOT the city. There are multiple stores within a city but we are considering each store as independent (consumers aren't shopping across stores).

2. Create two time series charts (or one chart with two vertical axes): one charting the average price of cereal over time and one charting the average sales of cereal over time. That is, the average should be the average price/quantity across all cereals and across all markets. Describe any interesting patterns in the data.

3. Calculate the HHI in each market (store/week), first by brand. Chart the average HHI (average across markets) over time. Do the same but calculate the HHI by manufacturer. Describe any interesting patterns. Is this market concentrated? Why or why not? (recall the HHI is $HHI = \sum_j s_j^2$ where s_j is the market share (in sales) of brand (or manufacturer) j.

2 Multinomial Logit and Nested Logit

We are now going to estimate two basic demand models, the Multinomial Logit and the Nested Logit. Start by assuming that the utility of consumer i of consuming cereal j in market m in time period t is given by:

$$u_{ijmt} = X_{jtm}\beta - \alpha p_{jtm} + \xi_{jtm} + \epsilon_{ijtm}$$

where X_{jtm} includes product characteristics, as well as a market fixed effect, a manufacturer fixed effect, and a time period fixed effect. The price is p_{jmt} and ξ_{jmt} is the unobserved quality of the cereal. We assume that ϵ is a type I EV shock.

1. Construct the data needed to estimate the models. First calculate the market share of each brand in each market (you already calculated this for the first questions actually) and calculate the market share of the outside option. What is the estimating equation?

2. Run OLS using the estimating equation. Put your estimates in a table and describe your results and whether or not they make sense. (Do not add the estimated fixed effects to the table, just put the estimates of the price coefficient and the preferences for the observed characteristics). Construct a histogram of the own price elasticities.

3. Now run 2SLS by instrumenting for price. Construct a price instrument using the sugar price times the sugar content of the cereal.. Add a column to your table for these estimates. Construct a histogram of the own price elasticities.

4. Construct a elasticity matrix for the top 5 brands in terms of sales (own and cross price elasticities). Do the substitution patterns make sense? Note: create the average matrix by taking the average of the market-level elasticity matrices.

5. Calculate average mark up for each manufacturer and plot them in a bar graph (now include all brands when calculating the average). Recall, we can calculate mark-ups just from the demand side (look at the FOC to get an expression for $p - c$ and then and $\mu = \frac{p-c}{p}$). Assume firms are setting the price of a brand independently (i.e., ownership matrix is just the identity matrix.

Now let's turn to a nested logit model. Let's assume that cereals are nested by their segment, so consumers first choose a segment, then they choose a cereal. The utility of a consumer choosing cereal j that belongs to segment g is given by:

$$u_{ijmt} = X_{jtm}\beta - \alpha p_{jtm} + \xi_{jtm} + \zeta_{jmt} + (1 - \sigma_g)\epsilon_{ijtm}$$

We assume that $\zeta_{jmt} + (1 - \sigma_g)\epsilon_{ijtm}$ is distributed GEV.

1. Construct the data needed to estimate the models. Calculate the 'conditional share' for each product (share of each segment) in each market. What is the estimating equation?

2. Run 2SLS by instrumenting for price AND the conditional share. Construct a second instrument for the conditional share that is the number of products in the segment. You can estimate the model assuming that $\sigma_g = \sigma$ for all g. Present your results in a table. The table will look similar to the table above, but now it will have the σ term. Construct a histogram of the own price elasticities.

3. Construct a elasticity matrix for the top 5 brands in terms of sales (own and cross price elasticities). Do the substitution patterns make sense? Note: create the average matrix by taking the average of the market-level elasticity matrices.

4. Calculate average mark ups for each manufacturer and plot them in a bar graph (now include all brands when calculating the average). Recall, we can calculate mark-ups just from the demand side (look at the FOC to get an expression for $p-c$ and then and $\mu = \frac{p-c}{p}$). Assume multi-product firms internalize the price effects. Hint: think about the ownership matrix.

5. What is the average diversion ratio between Cheerios and Corn Flakes (across markets)? How about Cheerios and Fruit Loops. Do these make sense?

6. What would happen to prices if General Mills and Kelloggs merged (calculate average of the mark-up changes)?

3 Mixed Logit

Now we will turn to the mixed logit. There are two files which give draws for consumer demographics (income and number of kids) (simulated_agents_income and simulated_agents_nkids). The following table gives the variable names and definitions for both files. The files is connected to the market data through the 'puma' and 'year' variables.

Table 2: Variables

Variable	Description
Name	desc
year	Year of data
puma	Identity code for the public use micro-data area (puma)
income#	Income draw # for that puma/year
nchild#	Number of children draw # for that puma/year

The utility of the consumer is now given by:

$$u_{ijmt} = X_{jtm}\beta_i - \alpha_i p_{jtm} + \xi_{jtm} + \epsilon_{ijtm}$$

where

$$\alpha_i = \frac{\alpha}{y_i}$$

and

$$\beta_i^x = \bar{\beta}^x + \beta^x z_i + \sigma\nu_i^x$$

Income is denoted y_i and general demographics (which are income and number of children) are z. ν_i is unobserved heterogeneity in tastes which we assume are distributed normal with mean 0 and variance σ^x. Estimate heterogeneity in preferences only for sugar content and fiber content, so you should have $2 \times 3 + 1$ non-linear parameters to estimate...σ for fiber and sugars, as well as β^{sugar} for income and number of children and β^{fiber} for income and number of children. The additional non-linear parameter is α because we have essentially interacted it with income.

1. Estimate the full Mixed logit model. Hint: you will need to draw consumers and calculate shares on each iteration of the contraction mapping. Keep the draw of consumers the same! Put your estimates in a table. Construct a histogram of the own price elasticities. Hint: use BLP instruments for the random coefficient estimates. That is, the instruments for the non-linear parameters should be $[SugarPrice, BLP^{sugar}, BLP^{fiber}, Fiber \times Inc, Sugar \times Inc, Fiber \times NChild, Sugar \times NChild]$, where:
$$BLP^x_{jmt} = (\bar{X}_{-jmt} - X_{jmt})^2$$
or in words: the average value of X for my competitors in that market and that year ($-j$ means not j, aka competitors) minus my value of X squared.

2. Construct a elasticity matrix for the top 5 brands in terms of sales (own and cross price elasticities). Do the substitution patterns make sense? Note: create the average matrix by taking the average of the market-level elasticity matrices.

3. Calculate average mark ups for each manufacturer and plot them in a bar graph (now include all brands when calculating the average). Recall, we can calculate mark-ups just from the demand side (look at the FOC to get an expression for $p-c$ and then and $\mu = \frac{p-c}{p}$). Assume multi-product firms internalize the price effects. Hint: think about the ownership matrix.