Problem Set 2
GR6254

Prof. Pietro Tebaldi

March 29, 2022

The problem set is due on **April 29, 2022**. You are allowed to work in groups of 2 or 3, but everyone must hand in their **own writeup of answers and code files**. Your code submissions should come with a **readme file** guiding through replication steps of each exhibit or number you are referring to in the text. Note that some of the results might not make economic sense. If this happens, you are expected to *briefly* discuss why this might be the case. In part, this is intentional and will make you realize the importance of the assumptions we make and their potential implications (the other part is the attempt to keep the specification simple and the data sets small).

I suggest you read the entire problem set in advance, to manage your time and organize your code. Importantly, before you begin answering (or coding for) a specific question, remember that parts of that answer or code might be useful later. Using flexible functions and parameters, rather than "hard coding" any specific part of the problem set will make your life easier in the future, although a little harder in the short horizon.

This dataset is designed to resemble a "real" empirical application, using individual level data, and employing a few of the methods that we have studied in class. There are no right or wrong answers, and you can be quite creative with empirical specifications, use of individual characteristics, richness of unobserved heterogeneity, etc. I am leaving things open-ended intentionally, specify your equations and go through the steps of a real empirical investigation.

NOTE: You can treat all variables as exogenous for the purpose of the problem set. This is a simplification, and you are welcome to think and discuss how you would deal with endogeneity. But this is not needed this time around.

The Data

The dataset Shining32 contains toothpaste purchase records for a sample of 10,000 New York City residents. For each individual, the data collects a number of demographic characteristics, as well as indicators for residential location. The labels should be self explanatory, and you should behave like a "real researcher" in dealing with this amount of information. When you write up your answer, specify how you make your decisions.

The data lists, for each individual, a set of 10 products, some characteristics, and the price of a package as listed at the day of purchase. Lastly, the dataset indicates the product purchased by the individual using the variable *Chosen*. Purchases are a combination of online and in store. For a subsample of individuals, the surveying company collected information on what would have been the hypothetical second choice of the individual if the first product was not available.

Research Questions

Demand Analysis I

i. Do different consumers respond differently to product characteristics when selecting their toothpaste? What are the relevant determinants of their choice?

ii. What is the (price-)elasticity matrix for the ten products? Does this differ if the researcher uses a simple logit model, a nested logit model (one nest for each brand), or a mixed logit model? How important are observable characteristics in explaining substitution patterns between toothpaste? How useful are the second choice observations in estimating unobserved heterogeneity? [hint: you can create a number of "fake" individuals after removing the first choice, and model their second choice as the maximum over the now limited choice set. Importantly, the draw of random coefficients must be the same for the real and the fake individual. This will help you pin down the correlation matrix between random parameters. You are encouraged to experiment with the number of random parameters and assumptions on independence.]

iii. **HARD; extra points** What is the correlation between brand fixed effects and price sensitivity?

iv. Do preferences vary across location?

v. What is the average Lerner index for each of the ten products if the researcher does not consider that brands offer multiple products?

vi. What is the average Lerner index for each of the ten products if the researcher considers that brands offer multiple products?

vii. What would be the effect of a merger between Crest and Colgate on profit margins? (You can assume that marginal costs remain fixed).

viii. What would be the effect of removing Crest from the market on consumer surplus, holding prices fixed?

Demand Analysis II

ix. One review article suggests that a number of individuals picks the cheapest product when shopping in store. At the same time, the article suggests that the online platform used by consumers at first shows only the N cheapest products, and consumers must click "See more" to observe the rest of the choice set. Using only raw data, can you infer the probability that the in store shoppers pick the cheapest product without doing any comparison? Can you infer what is N? Does your answer vary if the article measures "cheapest" in \$/oz, or \$/package?

x. Using your initial investigation of the raw data, can you write a choice model that describes demand for toothpaste in this context? Note: it is possible that online shoppers also choose only the cheapest product with some fixed probability.

xi. Estimate the model you specified in x. How do preference parameters change adjusting your analysis for consideration sets?

xii. Repeat part ii. adjusting your analysis for consideration sets. How does the elasticity matrix differ?

Note: if you wish to push yourself to become as good as possible in empirical IO, try to do all models (simple, nested, and mixed logits, with various combinations of observables and correlation structure of unobservables). However, this might be challenging and time consuming. To the very least, I expect you to be able to set up and estimate a simple logit model with rich observed heterogeneity. If you engage with unobserved heterogeneity, keep your number of draws small, in the interest of a short runtime.

xiii. Do you find evidence of preferences varying across location? How do you explain differences with your findings in part iv.?

xiv. Does the consumer surplus effect of removing Crest from the market differ after adjusting your analysis for consideration sets?

xv. Would expanding the first screen to all ten products improve competition? What about randomizing products into the first screen (holding its size to N products) with probability π? Can you plot the effect of π on consumer surplus and average Lerner indices? Would you recommend regulators to require online platforms to randomize products in the first screen?