GR6255: Industrial Organization III
Problem Set 1

Andrey Simonov

Spring 2025

Due at **11:59pm ET on March 14**. You may submit solutions in groups of two or three. Please submit the write-up electronically to *as5443@gsb.columbia.edu*, including a succinct Latex (.pdf) write-up and a well-commented executable code implemented in the program of your choice (e.g. in R/MATLAB/Python/Julia/other).

The goal of this exercise is to provide you with hands-on experience in coding and estimating demand, while deepening your understanding of how these models work. Given this goal, you can use any tools that are helpful to you. In particular, you may use LLMs when writing your code, but you must disclose which models you have used, for which parts of your response, and include prompts that you have used (if there are a lot of prompts, you can include informative examples). Regardless of what you have used, you are fully responsible for the write-up and code that you submit, in terms of it correctness, clarity, and efficiency.

1 Demand Estimation with Individual-Level Data

Part a

Suppose consumers indexed by $h = 1, ..., H$ make discrete purchase decisions j from a choice set $\mathcal{A} = \{0, 1, \ldots, J\}$ during each of the time periods $t = 1, ..., T$. A consumer can select not to purchase any products, which we index by choice $j = 0$.

Each period a consumer faces latent, choice-specific indirect utilities:

$$u_j(x_{jt}^h, s_t^h, \boldsymbol{\epsilon}_t^h; \Theta^h) = \begin{cases} \beta_j^h + \eta^h x_{jt}^h + \epsilon_{jt}^h, & j = 1, ..., J \\ \\ \epsilon_{0t}^h, & j = 0 \end{cases}$$

where $\boldsymbol{\Theta}^h = \left(\beta_1^h, ..., \beta_J^h, \eta^h\right)'$ are the consumer's taste parameters, x_{jt}^h is the price of the product j in period t and $\epsilon_{jt}^h \sim i.i.d.$ EV$(0, 1)$ are random utilities. The consumer makes a discrete choice

$$y_{jt}^h = \mathbf{1}\left(u_j(x_{jt}^h, \boldsymbol{\epsilon}_t^h; \Theta^h) \geq u_k(x_{jt}^h, \boldsymbol{\epsilon}_t^h; \Theta^h),\ \forall k \neq j\right),\ j = 0, ..., J.$$

i)

Write down the individual-level likelihood and score functions for Θ^h.

ii)

Simulate choice histories for $N = 1,000$ households with $T = 50$. Assume that $J = 4$ and that consumers' preferences are distributed as a multivariate normal:

$$\begin{bmatrix} \beta_1^h \\ \beta_2^h \\ \beta_3^h \\ \beta_4^h \\ \eta^h \end{bmatrix} \sim N\left(\begin{bmatrix} -1.71 \\ 0.44 \\ -1.37 \\ -0.91 \\ -1.23 \end{bmatrix}, \begin{bmatrix} 3.22 & 0 & 0 & 0 & 0 \\ 0 & 3.24 & 0 & 0 & 0 \\ 0 & 0 & 2.87 & 0 & 0 \\ 0 & 0 & 0 & 4.15 & 0 \\ 0 & 0 & 0 & 0 & 1.38 \end{bmatrix}\right). \tag{1}$$

Assume that prices evolve as an exogenous Markov process with 100 discrete states (joint distribution for all four products). You can find the state space (`price_transtion_states.csv`) and the transition probability matrix (`transition_prob_matrix.csv`) under the `problem sets/pset1` folder on courseworks.

For the simulated data, estimate Θ^h using the following specifications:

- MLE assuming homogenous Θ^h;
- MLE assuming discrete-class Θ^h with $K = 2$ classes;
- MLE assuming that Θ^h has a normal distribution across the households.

For each estimator, code up the likelihood and gradient and use them to find the MLE estimates using your favorite optimization routine. Report the point estimates and standard errors, as well as the implied average own- and cross-price elasticities. Compare the estimates and elasticities between the specifications and to the parameters and elasticities from the true data-generating process. Make sure to include an economic interpretation of the magnitudes of the biases in the estimates that you get.

Part b

Suppose the consumer utility exhibits first-order Markov state dependence:

$$u_j(x_{jt}^h, s_t^h, \boldsymbol{\epsilon}_t^h; \Theta^h) = \begin{cases} \beta_j^h + \eta^h x_{jt}^h + \gamma^h \mathbb{I}\{s_t^h = j\} + \epsilon_{jt}^h, & j = 1, ..., J \\ \\ \epsilon_{0t}^h, & j = 0 \end{cases}$$

where now $\boldsymbol{\Theta}^h = \left(\beta_1^h, ..., \beta_J^h, \eta^h, \gamma^h\right)'$. The state variables $s_t^h \in \{0, \dots, J\}$ indicate the consumer's current loyalty state (where $s_t^h = 0$ stands for having no loyalty state, e.g. before the first purchase). In the model, the previous period's choice determines the current period's loyalty state. More explicitly, $s_{t+1}^h = \sum_{j=1}^J j \cdot I(y_{jt}^h = 1) + s_t^h \cdot I(y_{0t}^h = 1)$, meaning that the state s_{t+1}^h stays the same ($s_{t+1}^h = s_t^h$) if consumer h chose an outside option in period t, and changes to the purchased product if there was a purchase in period t.

Notice that to simulate the states, you can run the simulation sequentially starting in period $t = 1$, since the choice and state in period $t = 1$ determine the state in period $t = 2$.

The specification above implies that the loyalty state follows a Markov Process, $s_{t+1}^h = g\left(s_t^h, \Theta^h\right)$, where the transition probabilities consist of the conditional choice probabilities. The transition probability of the loyalty state for a consumer h from j to k is

$$\Pr\left\{s_{t+1}^h = k|s_t^h = j, \boldsymbol{\Theta}^h\right\} = \begin{cases} \Pr\left\{y_{jt} = j|s_t^h, \boldsymbol{\Theta}^h\right\} + \Pr\left\{y_{jt} = 0|s_t^h, \boldsymbol{\Theta}^h\right\}, & k = j \\ \Pr\left\{y_{jt} = k|s_t^h, \boldsymbol{\Theta}^h\right\}, & k \neq j. \end{cases} \tag{2}$$

i)

Once again, simulate the choice histories for $N = 1,000$ households, now with $T = 150$. Assume that the initial state is $s_1^h = 0$ and that consumers' preferences for β and η are distributed as in part a; the distribution of the state dependence is $\gamma^h \sim N(1,1)$, independent of β and η. Use the same price process.

Drop the first 100 choices for each consumer in the simulated data. Assume that the state you observe in period $t = 101$ (first estimation period) is exogenous. For the rest of the 50 choices, estimate Θ^h using the same specifications as before:

- MLE assuming homogenous Θ^h;
- MLE assuming discrete-class Θ^h with $K = 2$ classes;
- MLE assuming that Θ^h has a normal distribution across the households.

Report the point estimates and standard errors. Comment on the γ's mean and standard deviation estimates, comparing them between the specification and to the true DGP. Is there a bias in the γ estimates when you correctly specify the parametric form of heterogeneity (normal distribution)? What is the intuition behind this? If there is bias, how would you interpret its magnitude (e.g. what does it imply about own-price elasticities when the state is equal to 0 vs. 1)?

ii)

For the first simulated choice (out of 50, after removing the first 100 choices), set the initial state $s_1^h = 0$ for each household (make sure to adjust the initial states until the household make the first purchase in the 50 observed choices). Re-estimate the distribution of Θ^h under the correctly specified parametric distribution of heterogeneity. Comment on the estimates of γ; what is the intuition behind this result?

iii)

Given your estimates of γ for a specification with a normal distribution from parts i an ii, comment on how you should change the estimation procedure to eliminate the bias coming from the specifications of s_1^h above. Write out the likelihood corresponding to your proposed approach and discuss the ways to find the MLE of Θ^h or to take draws from the implied posterior distribution.

iv) (optional)

Re-do the estimation of Θ^h with a normal distribution across the households using Bayesian inference. For this, in R, use the *rhierMnlRwMixture* function in the *bayesm* package to make MCMC draws from the posterior distribution. Notice that *rhierMnlRwMixture* is a routine that allows you to specify the first-stage prior as a mixture of normal distribution; you can also use this function for a normally distributed first-stage prior (setting the number of normal components equal to 1). Run the MCMC chain for 20,000 draws. Plot the likelihood draws as well as the draws of mean Θ. Report the posterior point estimates of the mean and variance of Θ^h and the 95% credibility interval around them.

2 Demand Estimation with Market-Level Data

Note: This question is an adjusted version of the BLP demand estimation exercise from Chris Conlon that uses his pyBLP package in the last (optional) section.

Setting

You will estimate demand in a stylized model of the market for pay-TV services. There are T markets, each with four inside goods $j \in \{1, 2, 3, 4\}$ and an outside option. Goods 1 and 2 are satellite television services (e.g., DirecTV and Dish); goods 3 and 4 are wired television services (e.g., Frontier and Comcast). The conditional indirect utility of consumer i for good j in market t is given by

$$\begin{aligned} u_{ijt} &= \beta^{(1)} x_{jt} + \beta_i^{(2)} satellite_{jt} + \beta_i^{(3)} wired_{jt} + \alpha p_{jt} + \xi_{jt} + \epsilon_{ijt} \qquad j > 0 \\ u_{i0t} &= \epsilon_{i0t}, \end{aligned}$$

where x_{jt} is a measure of good j's quality, p_{jt} is its price, $satellite_{jt}$ is an indicator equal to 1 for the two satellite services, and $wired_{jt}$ is an indicator equal to 1 for the two wired services. The remaining notation is as usual in the class notes, including the i.i.d. type-1 extreme value ϵ_{ijt}. Each consumer purchases the good giving them the highest conditional indirect utility.

Goods are produced by single-product firms. Firm j's (log) marginal cost in market t is

$$\ln mc_{jt} = \gamma^0 + \mathrm{w}_{jt}\gamma^1 + \omega_{jt}/8,$$

where w_{jt} is an observed cost shifter. Firms compete by simultaneously choosing prices in each market under complete information. Firm j has profit

$$\pi_{jt} = \max_{p_{jt}} M_t(p_{jt} - mc_{jt})s_{jt}(p_t).$$

You can find simulated data for this market under (`market_demand_simulated_data.csv`) under the `problem sets/pset1` folder on courseworks. The outside option $j = 0$ is omitted from the data. This data contains the following variables:

- `market`: market ID (t)
- `product`: product ID (j)
- `s`: simulated market share of j in t

- `p`: simulated price of j in t
- `x`: quality of j in t
- `w`: cost shifter of j in t
- `mc`: marginal cost of j in t
- `satellite`: a dummy variable for j being a satellite product
- `wired`: a dummy variable for j being a wired product

This data is simulated for $T = 600$ markets. Product characteristics and cost shifters x_{jt} and w_{jt} are simulated as absolute values of iid standard normal draws. Demand and cost unobservables are simulated assuming that

$$\begin{pmatrix} \xi_{jt} \\ \omega_{jt} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix}\right) \text{ iid across } j, t.$$

Preference estimates are distributed as:

$$\begin{aligned} \beta^{(1)} &= 1,\ \beta_i^{(k)} \sim \text{iid } N(4,1) \text{ for } k = 2,3 \\ \alpha &= -2 \\ \gamma^{(0)} &= 1/2,\ \gamma^{(1)} = 1/4. \end{aligned}$$

Prices correspond to the equilibrium prices for each good in each market assuming that j belong to single-product firms that maximize static profits. Finally, observed shares are simulated given the draws of product characteristics, prices, preferences, and unobservables.

i)

Estimate the plain multinomial logit model of demand by OLS (ignoring the endogeneity of prices).

ii)

Re-estimate the multinomial logit model of demand by two-stage least squares, instrumenting for prices with the exogenous demand shifters x and excluded cost shifters w. Discuss how the results differ from those obtained by OLS.

iii)

Now estimate a nested logit model by two-stage least squares, treating "satellite" and "wired" as the two nests for the inside goods. Note that you should allow a different nesting parameter for each nest. Consider various instruments that we have discussed in the class and argue for your preferred specification.

iv)

Using estimates from your various specifications above, provide a table comparing the estimated own-price elasticities to the true own-price elasticities. Discuss the results.

v) (optional)

Estimate a random coefficient model using:

- Your code that uses nested fixed point algorithm from BLP (1995)
- Your code that uses MPEC (Judd Su 2012)
- `pyBLP` package from Conlon and Mortimer (2020): see https://pyblp.readthedocs.io/en/stable/ to get started

For each of these estimations, you can use only demand moments or demand and supply moments, and compare the estimates. Using your preferred estimates, provide a table comparing the estimated own-price elasticities to the true own-price elasticities.

Note: I expect part **v)** to be challenging, especially when using your own estimation routine. I include it in case you are interested to practice full BLP estimation; we will return to the nested fixed point and MPEC estimators in problem set 2.