Pricing Analytics Project 1:

Estimating demand for cars

Takeaki Sunada

1 General notes

Please send your answers by Friday, Jan 27 by the end of day (11:59pm), to "mkt440submission@gmail.com". The submission should take either of the following forms.

- Option 1 (3 files to attach): A write-up in either pdf or html format (one per team), your .R code and an HTML file compiled by Rstudio. The HTML file is generated by selecting "files - compile report" in Rstudio console. It then automatically runs your R code and report all results side by side with your code. In this case, your write-up should be in a separate file and not mixed up in the code.

- Option 2 (1 file to attach): An R-markdown report, either in pdf or in html format, that puts together your code, results and the write-up. Please put your write-up as a text in the report, and not as a comment in the code.

Either case, please use as the file name "first names of the team members, in alphabetical order, separated by underscore (_)". No need to say "Project 1" (it's obvious from the submission date).

In your write-up, please clearly indicate the first and last names and student ID of all team members. Also please clearly label each of your answers (for example, "answer to question 4-1"), so that we know which question your answer is addressing. I will let our graders subtract points if formatting/labeling issues cause them hardship to grade.

If you can keep your write-up somewhat concise, our graders would really appreciate it (of course, we understand it may take some space to fully defend your specification, but no need to defend every single detail - just get what you believe to be the key points right).

Sections 2 and 3 below are the description of the problem and the data. The questions start from Section 4.

2 Environment

It is in early 2000s. You are hired as a senior analyst by a new U.S. car manufacturer, T (as a shorthand), which is considering launching operations in Europe. In order to make the important managerial decision, your team is assigned a task - estimating demand for cars and competitiveness of the car market in Europe. As you are new to the market, the only data set available to you is the one that is publicly available - the sales record of each car model produced by existing car manufacturers in each European country between 1970 and 1999. You don't really know the cost structures of the rival firms, neither do you have access to their customer records. You need to provide demand estimates from the data of aggregate sales and prices.

3 Data set

The dataset consists of "cars.csv" and "iron_ore.csv". "cars.csv" includes sales, prices and observed attributes of cars. Description of all variables is available in "variables.xlsx". Some notes are in order.

- Variables ye, ma, and co are the keys of the dataset, representing year, market (country), and the car model, respectively. A unique combination of the three variables defines a unique observation of the data.
- Variables qu and pr represent sales quantity and price of the vehicle in the given market and year: qu measures the number of new car registered in the market in the year, and pr is denoted by the domestic currency. Because each country uses a different currency, it is difficult to compare between countries. As such, you can use the variable eurpr – which is price converted to Euros – to represent price.
- Variables cy, hp, we, le, wi, he, li, sp, ac, and some others, measure attributes of the car model. Note that not all variables may be relevant for demand.
- Other variables may include possible factors that partly determine production costs, such as exchange rates and consumer price indices.

"iron_ore.csv" records production level and prices of iron ore, which is the main input for car production. During the exercise, you may assume that the amount of iron ore used to produce each car model is proportional to the weight ("we" in cars.csv) of that model. "production" is the variable that stores production level, and "unit_value_98" is the variable that stores the commodity price of iron ore, inflation-adjusted to 1998 dollars.

A template R code for the project is available on Blackboard for your convenience.

4 Control variables

4.1 Interpreting a log-log regression

Your colleague starts the analysis by estimating a log-log regression, trying to get a sense of the price elasticities. In particular, denoting j as a car model (co), m as a market (ma), and t as a year (ye), she estimates

$$\log(qu_{jmt}) = \beta_0 + \beta_1 \log(eurpr_{jmt}) + \epsilon_{jmt}.$$

She gets the following results:

```
Call:
   felm(formula = log(qu) ~ log(eurpr), data = data)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.2904     0.2046   55.19   <2e-16 ***
log(eurpr)   -0.2925     0.0231  -12.66   <2e-16 ***
```

Questions:

1. What is the interpretation of the regression result (e.g. intercept and coefficient)?
2. Does the shape of the estimated demand look reasonable? Why do you think so (not so)?
3. Is there any way to justify your statement?

4.2 Adding control variables

You spotted that without controlling for demand-side factors, the estimated demand may not represent the right causal relationship. As such, you decide to include some of the demand-side

variables as an X in the regression.

Questions

1. From the data, pick control variables and fixed effects to add to your regression, find your preferred specification and report results from that specification.

2. Justify your specification choice.

5 Instrumental variables

You still have a concern that available X terms and fixed effects in the data may not be sufficient to recover the true causal relationship. You hence decide to try an instrumental variable approach.

Questions

1. From the data, use a variable (or variables) that you find appropriate as an IV for the price and report results from that IV regression. How does the result change with and without IV?

2. Justify your choice - why do you think it's a good IV?

6 Recovering costs

You have now obtained an estimate of the demand structure. However, in order to assess the potential competitiveness of the market, you may also want to know the production cost of your rivals.

Here's how you may be able to get an estimate of rivals' unit cost. Note that once we know the demand, we can calculate the optimal price of a firm with a given unit cost (if not sure how, review asynchronous video 2). Let's reverse-engineer this exercise. Assume that rivals set the optimal price in the exact same way as you do - given the demand (that you estimated) and their own cost, your rivals set the price to maximize their profit. You then know both the estimated demand that your rivals face and how rivals set prices optimally given their cost. You should be able to calculate the rival cost that makes the rival set the price that you see in the data.

Questions

1. Based on the insight above, estimate the unit cost of cars in the data.

Note: Technically, you can recover the cost for each car in each country in each year. Recovering costs for all observations may be computationally burdensome. If you find it overwhelming, feel free to pick a few car model, year and country you like and only estimate the cost for them.

2. Are the estimated costs reasonable? Why (why not)?

   A possible approach to discuss this: Do you see any specific patterns in how the estimated costs are distributed across products? What does the pattern tell you about the validity of your demand estimates?

7 Cross-elasticities and competitive effects

You realize that the log-log regression you have estimated can only give you an estimate of the own-price elasticities, and not the cross-price elasticities. With this thought, you decided to also include prices of rival cars. Use "avgeurprrival" (average of the prices of the rival cars in that market in that period) as a measure of rival's car price.

1. What does the coefficient of the log(avgeurprrival) represent? (Note that average rival price is the sum of rival prices divided by the number of rival car models.)[1]

2. What is the value of the coefficient you expect to get? Does your estimate fulfill your expectation?

3. What does the estimated coefficient tell you about the competitiveness of the market?

8 Miscellaneous

Conceptual hints:

- Throughout the exercise, focus on establishing the causal β_1. i.e. the purpose of adding X and/or finding an IV is NOT, for example, to improve model fit. Of course in reality we also care about the overall model fit, but let's focus on what we learned in the class for this exercise.

[1] For those of you with math background - for this question, you may ignore the nonlinearity of log transformation.

- In terms of what to add to the regression, think about more fixed effects or product attributes, or both. Note that this is a panel data with three dimensions - multiple geographical markets, multiple time periods and multiple products.

- As we discussed in the class, there's no way to formally test which variable we should add as X in the regression. Similarly, no formal way to test which one is the right IV. Thus, your decisions of what variable to use solely rely on *your guesswork*. In that regard, there's no unique "right" answer to this question and different teams likely end up with different specifications, and that's fine.

- Given that no right answer exists, what we want to see is your reasoning. Why do you believe that you should use the variable you use? Tell us a story. Telling a story that is right in light of what we discussed in the class will positively impact your score. The story may come from your commonsense, may be based on the lecture slides, or there may be a way to infer something from your data - for example, as you add a new variable as an X to the regression, what you expect to see if the variable does the job you expect it to do?

- As we discussed, in general it's a good idea to include many variables in X when in doubt. But for this exercise, do not follow this strategy - we want to see you *constructively* come up with a specification with a reasoning that convince us. As such, we will subtract points if we see a lot of variables in X without much reasoning.

Practical tips in R:

- "factor(variable name)" is an easy way to create the set of dummy variables that serves as a fixed effect.

- I strongly recommend using "felm" function included in "lfe" package, instead of standard "lm", to run a regression. The example syntax is available in the template code file.