Pricing Analytics Project 1:

Estimating Demand for Cars

Takeaki Sunada

Updated Jan 19 (An extra tip for IV added)

1 General notes

Please send your answers by Friday, Jan 31 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 down your write-up as texts in the report, and not as comments 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".

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 you are addressing. I will let our graders subtract points if formatting/labeling issues cause them hardship to grade. Concise write-ups are greatly appreciated (though of course we understand sometimes it just takes space to justify your specifications).

The use of generative AI is allowed throughout the project. However, for coding purposes, I'd recommend making use of the template code and manually giving a couple of necessary edits. If any parts of the project make use of generative AIs, please properly acknowledge it in the report (i.e., explain where and how they are used).

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 - (1) estimate demand for cars in Europe, and (2) assess the competitiveness of the car market there. 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. Hence, you need to accomplish your tasks solely with the data of aggregate sales and prices.

3 Data set

The dataset consists of "cars.csv" and "iron_ore.csv". "cars.csv" includes quantity sold, 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 the sales quantity and the price of the vehicle in a given market and year: qu measures the number of new car registrations, 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 may 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 primary input for car production. "production" is the variable that tracks the production level, and "unit_value_98" is the variable that tracks the commodity price of iron ore, inflation-adjusted to 1998 dollars. 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.

Throughout the project, please use "felm" function in "lfe" package. A template R code is available on Blackboard. The template should cover most of the specifications you may want to try and, hence, I expect you to mostly copy-paste the necessary lines and manipulate the variables to get the results.

4 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. β_1 estimate from this regression is likely not causal. Explain why.

5 Adding control variables

For reasons you stated in the previous question, your colleague's demand estimate is not causal and hence not usable for pricing purposes. Your task is to improve the model and obtain a better causal estimate of β_1. Let's start from adding some of the demand-side variables as a control variable (X) and fixed effects to 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 (coefficients and standard errors). Please make use of the template code, Section 1.

2. Justify your specification choice. Why did you choose that set of variables over others?

    - Tip: As we discussed in the class, there's no way to formally test the best specification. Similarly, there's no formal way to test the best instruments. Thus, when we ask you to defend your specification, we are expecting some verbal reasoning from you. Intuitively, why do you think adding that specific variable helps reduce correlation between P and ϵ (and hence helps establish causality)? Note that there's no unique "right" answer to this question, because, again, there is no formal test.

6 Instrumental variables

You still have a concern that X terms and fixed effects you added in the previous section might not be sufficient to fully eliminate the confounds. Thus, you decided to also try out the instrumental variable approach.

Questions:

1. Building on the specification from the previous section, consider running an IV regression. 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? Please feel free to make use of the template code, Section 2.

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

- Tip 1: If you simply add an IV to the specification from the previous section and it works out, great! Sometimes you aren't so lucky and may end up with weird estimates. In that case, try adding or dropping a couple of X variables or play with the specification of your IV.
- Tip 2: If your initially chosen IV does not work, recall the assumption - the amount of iron ore used to produce each car model is proportional to the weight of that model.

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 the prices of rival cars. Use the log of "avgeurprrival" (average of the prices of rival cars in that market in that period) as the measure of rival car price.

Note: if you have already included it in the previous sections, no need to delete it - please provide your answers to the following questions based on that specification.

Questions:

1. Conceptually, what does the coefficient of log(avgeurprrival) represent? (Note that the average rival price is the sum of rival prices divided by the number of rival car models.)
2. Conceptually, what is the most plausible range for the coefficient of log(avgeurprrival) (e.g., should it be positive or negative, etc.)? Does your estimate fit within that range? If not, is there anything you can do about it?
3. What does the estimated coefficient tell you about the competitiveness of the market?

8 Recovering costs

You now have a much better demand model - it offers a better causal estimate, and also a measure of cross-elasticity! There's a lot you can write to your boss with it.

However, note that you are also requested to provide your assessment of market competitiveness. Cross-elasticity you estimated above is only part of the whole picture here - to add to your report, it's probably a good idea to also estimate the production cost of your rivals. Rival costs are

obviously an important input to measuring market competitiveness (e.g., if you have a lower cost than your rivals, you obviously have more competitive advantage).

Obviously you don't know your rival companies' cost structure. However, it is known that if the market exhibits a log-log demand system and the rivals are setting an optimal price against it, you can estimate their cost by the following formula:

$$\text{Cost} = \text{Price} \times \frac{1 + \beta_1}{\beta_1}.$$

Use this relationship to estimate the rivals' costs and address the following questions (the rationale behind this expression is available in the Appendix).

Questions

1. First, use $\beta_1 = -0.2925$ (your colleague's estimate) to recover the costs. What do you find? What does the result imply about the validity of demand estimates your colleague had?

2. Next, use β_1 you obtained in the previous section to recover the costs. How does your cost estimate improve over your colleague's? What does the result imply about the validity of your demand estimates?

    - Tip: because we are inferring rival costs based off of our demand estimates (and the assumption that the rival firm is optimizing prices against that demand), if our cost estimates are strange, that means our demand estimates aren't quite right.

9 Appendix: The rationale behind the cost expression

The formula we used to recover costs comes from the assumption that the price of each car model is set to maximize profits based on the log-log demand we estimate. Consider the profit maximization problem of a firm facing a log-log demand function:

$$\pi = \max_P \exp(\beta_0 + \beta_1 \log(P) + \beta_2 \log(P_2) + X + \text{Fixed effects})(P - C),$$

where π stands for profit and C stands for unit cost (hence $P - C$ is unit margin). Note the "exp": because we measure demand in $\log(Q)$, we need to make it back to Q by exponentiate it.

Maximizing profit is equivalent to taking the first-order-condition as follows.

$$\begin{aligned}
\frac{\partial \pi}{\partial P} =& \exp(\beta_0 + \beta_1 \log(P) + \beta_2 \log(P_2) + X + \text{Fixed effects}) \\
&+ \exp(\beta_0 + \beta_1 \log(P) + \beta_2 \log(P_2) + X + \text{Fixed effects}) \frac{\beta_1 (P - C)}{P} = 0, \\
\Rightarrow & 1 + \frac{\beta_1 (P - C)}{P} = 0.
\end{aligned}$$

Note that the exp term cancels out in the first-order condition. Rearranging, we have

$$C = P \frac{1 + \beta_1}{\beta_1}$$

and we have the formula we used.