ECON 370: Assignment 3

Drew Van Kuiken

Due Friday, December 8, 2024

This assignment has 2 parts. The first part includes some more advanced simulation review. This will be good practice for those of you considering running a Monte Carlo simulation for a final project. The second part asks you to start working on your final project. Note: **AI is not allowed on the first part. It is allowed on the second part.** Consider using StackOverflow before asking AI for a solution!

Part One: Firm Product Introduction (Total Points: 25)

Suppose that a firm can release one of three products or nothing at all. A firm will release a product that makes the most profit. Suppose, for simplicity, the revenue of each product j is simply $r_j = j$.[1] So the revenue of product 1 is simply 1, the revenue of product 2 is 2, the revenue of product 3 is 3, and the revenue of not releasing anything ("product 0") is 0. However, suppose that the cost of production is $c_j \sim \text{Exp}(1/2)$. Note, each product has it's own cost c_j. The cost of not releasing anything is 0. Profits are then

$$\pi_j = r_j - c_j$$

for products $j = 1, 2, 3$, and $\pi_0 = 0$ for not releasing anything. The reason for this (simplified but motivated) setup is that from an econometric perspective, we can usually observe revenues but costs are trickier, so we are assuming they are random.[2]

Question: Estimate the probability that the firm will release each product (including nothing at all). *Again, a product is released if it results in the most profit among the four options* (the three products and not releasing anything).

Hint: Form a profit matrix that is $N_{sim} \times 4$ where each row is a simulation and each column is the profit of a product. That is, each row should have the following entries in the columns from left to right: 0, $1 - c_1$, $2 - c_2$, and $3 - c_3$ where c_1, c_2, c_3 are three *different*

[1] Typically, one would have actual data on revenues.

[2] While we might be able to get accounting data, the way costs are accounted for and the actual "economic costs" are usually different. Plus, if you remember from Econ 101, implied costs (i.e. opportunity costs) are obviously not in accounting data.

draws from Exp(1/2). Then, for each row, see which column (or product) has the highest amount of profit. Lastly, use these results to estimate the probability.

Note: This problem is different (and easier) than an example in the lecture slides. So if you strictly copy that code, ***it will be incorrect.*** That said, feel free to use the slides as inspiration.

Final Project Prep (Total Points: 75)

As a reminder, you can review expectations and details for the final project in course materials\final_project on Canvas (or on GitHub).

Question 1: What is your research question? Is your question descriptive or causal? Why? (5 points)

- What results do you expect to find?

Question 2: Describe the data generating process for your research question (5 points)

Question 3: Navigate to the sharepoint folder located here: `https://adminliveunc-my.sharepoint.com/:f:/r/personal/dvankuik_ad_unc_edu/Documents/econ370_data_samples?csf=1&web=1&e=548tYq`. Create a subfolder labeled [first]_[last] (I created an example subfolder with the solar panel data already). Within your subfolder, store copies of the datasets you plan to use for this project. (Note: if you plan to run Monte Carlo simulations for this project, set a seed in your script and save a script with your data generating process (i.e., the process you use to simulate the data, not a description of your DGP in words) to the sharepoint site.) (5 points)

Question 4: What is your dependent variable? What are your independent variables? What are your controls? (5 points)

- Is there bias in your regression? What data would you have in an ideal world?
- Note: you can think of the distinction between independent variables and controls as follows: independent variables capture an effect that you're interested in understanding. Control variables are things that would lead to omitted variable bias if you didn't include them, but which are not particularly interesting to understand in the context of your research question. For example: in our gapminder regression, we controlled for being a defender in a war. Whether or not we cared about this effect would depend on our research question. If we wanted to understand the relationship between GDP per capita and Life Expectancy in Asia, we wouldn't really care about the size of the coefficient on defender in a war. In this case, we'd still have omitted variable bias if we didn't include it in our regression, but it would be a control because we wouldn't per se care about the size of the coefficient. If, instead, our research question was about the effect of wars at home on life expectancy, we might

think of GDP per capita as a control, and the size of our "defender in a war" dummy would be our main coefficient of interest.

Question 5: Using the data cleaning package of your choice, join your datasets together in a way that is relevant to your project (10 points)

- What keys are you merging on in this case?
- If you only have 1 dataset, for this question you can describe precisely how you are planning on matching on observables or otherwise accounting for selection bias in this project. If you are only using 1 dataset and you are not accounting for selection bias (or, more broadly, you're planning on pursuing a more descriptive question), please select a different research question.

Question 6: For the next set of questions, please provide answers using **both** the tidyverse (25 pts) and data.table (25 pts). Complete the following steps for your dependent and primary independent (i.e., don't worry aobut doing these steps for your control variables for the moment):

- Clean the data sufficiently that you you have intelligible densities for independent and dependent variables (i.e., no values that are NA, or at least all NAs are intentional and important for your analysis; you've cleaned up outliers; you've subsetted the data to include only relevant observations, etc.) (10 points)
- Plot the density of each variable using ggplot (5 points)
- Summarize the mean, median, minimum, maximum, and standard deviation of each variable (5 points)
- Summarize the mean, median, minimum, maximum, and standard deviation of your **dependent** variable for different levels of your independent variable(s). (5 points)

Note: If your independent variables are discrete, this should be relatively straightforward. If they are continuous, summarize your main dependent variable for each quantile of the independent variable(s). If any of these words are confusing, look them up! Or ask me about them at office hours.