CODING ASSIGNMENT
SIEPR PREDOCTORAL FELLOWSHIP APPLICATION 2022-23
JOSÉ IGNACIO CUESTA

This assignment has the goal of assessing the skills, thinking and creativity of applicants for a predoctoral fellowship position with Professor José Ignacio Cuesta for the 2022-23 academic year.[1] We are looking for work that carefully executes the assignment with clear documentation of how decisions were made. If you can't figure something out, that is perfectly okay. Document your confusion, make a decision on how to proceed, and move on. Be resourceful: look at the hints in the prompts, think about what the question is asking, use google, etc.

Please submit your answers in a single compressed file. Collect all written answers in a single nicely formatted pdf document generated in LaTeX. For code, we should be able to execute them changing no more than the name of the current directory. Please comment your code and be sure to note things that you were not sure about, as well as the solution that you wound up choosing. Please submit your results to jicuesta@stanford.edu by November 30, 2021 or before.

1 Maximum likelihood

The logit model is one of the most commonly used non-linear models in applied work. In this question, you have to use a logit model to simulate data from it and then estimate its parameters. The model is given by the following equation:

$$y_i^* = \alpha + \beta x_i + \varepsilon_i$$

where $y*_i$ is a latent variable, $x_i \sim N(0, 2)$, and ε_i follows a type 1 extreme value distribution. The parameters of the model are $(\alpha, \beta) = (1, 0.5)$. The model implies that the probabilities that y_i takes the values 0 and 1 are:

$$\begin{aligned} P(y_i = 0) &= \frac{1}{1 + \exp(\alpha + \beta x_i)} \\ P(y_i = 1) &= \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \end{aligned}$$

a. Simulate a sample of $N = 1,000$ realizations of y_i from the model, using the parameter values provided above for (α, β).

b. Write down the log-likelihood function of the data.

c. Estimate the parameters of the model (α, β) by maximum likelihood.

d. How long does the estimation routine takes? What could you do to make it faster? No need to do it here, but suggestions would be helpful.

[1] This is the same task that Professor Shoshana Vasserman is asking for in her application.

e. Provide standard errors for your estimates.

f. Using your estimates, compute predicted values of the probability $P(y_i = 1)$. Plot a histogram of predicted values of $P(y_i = 1)$. What is the average predicted probability?

g. Plot predicted values of the probability $P(y_i = 1)$ over the domain of x_i, for values $x_i \in [-2, 2]$. Add the values of the probability $P(y_i = 1)$ under the true parameter values in the same plot. Your plot should have a domain of $[-2, 2]$ for x_i and of $[0, 1]$ for $P(y_i = 1)$.

2 Working with data

We are interested in studying the effect on an incentivized training program for small businesses that began on $t_0 =$ January 1st, 2013. The program was aimed at increasing productivity. After many interviews with administrators in the Ministry of Economics, we learned the following facts about how the program was run:

- The program consisted of a set of training sessions to firm managers, which had the goal of delivering management counseling sessions and increasing productivity.
- Only firms with 100 or fewer employees were eligible to participate in the program.

Attached you will find three datasets for a broad set of firms:

1. `firm_information.csv` - Basic firm information
2. `aggregate_firm_sales.csv` - Firm sales by month
3. `monthly_data/YYYY-M.csv` - Firm auxiliary data by month.

Part 1. Merge the 3 types of datasets into a single dataset using firm IDs.

Note: This process may or may not be straightforward. In the latter case, discuss the issues you faced and how you dealt with them.

Part 2. We want to understand the differences between groups of firms according to their eligibility for the program, and whether they adopted it. Using graphs and/or tables:

- Compare eligible vs non-eligible firms when the program begins
- Compare adopting vs non-adopting firms when the program begins
- Compare adopters vs non-adopters among eligible firms when the program begins.

Note: There are many ways to characterize groups. Limit yourself to no more than 4 figures/tables. Please explain your reasoning behind each output: what characteristics did you choose to examine and why? What are you trying to learn with each comparison? What time periods are important to look at if we want to analyze the impact of the training program?

Part 3. To estimate the impact of the program, it is natural to use the eligibility cutoff at 100 employees. Please focus on the following dependent variables: (i) A dummy for being eligible, (ii) a dummy for adopting the program, (iii) log(sales), (iv) log(revenue), (v) log(employment), (vi) log(wage bill).

a. This research design is called a regression discontinuity design (RDD). Briefly explain the idea behind this strategy.

b. Begin by simply plotting the data. In particular, produce plots that demonstrate the relationship between the 6 outcomes and eligibility in the program. To make the figure easier to interpret, consider plotting the average of the outcome over fixed bins of employment. This is often referred to as a bin-scatter—see here for an implementation in R/Python/Stata, though you are welcome to write your own as well. Add a vertical line at 100, and add linear fits of the discontinuity. Discuss the results briefly.

   Note: There are a number of ways to implement the specification above—you have to choose which parts of the panel to plot, what size of bins to use, whether or how to color or group observations, etc. Feel free to be creative in making figures that inform our understanding of the program's impacts and explain your decisions.

c. To estimate the impacts of the program, you want to estimate the following regression:

$$Y_{it} = \tau D_{it} + \beta X_{it} + \alpha_i + \eta_t + \varepsilon_{it}$$

   where where $D_{it} = 1[employment_{it_0} >= 100, t > t_0]$, X is employment in $t-1$, α_i are firm fixed effects, and η_t are time fixed effects.

   Interpret this equation. Why do you think we structured it this way? What is τ? What is β? Are there any details that we forgot to mention for implementing this design?

d. Report the results from part (c) in a table with 6 columns. Interpret the results: what can we says about the impact of this program?

e. Pause and consider: what kinds of (conceptual) errors/issues might we have missed? One or two examples are sufficient.

3 Coding Sample

We would like to see a coding sample of which you are proud. If you have any such sample, please send it along with the respective output. Links to GitHub repositories are encouraged if available. The output could be a paper, a figure that you put together, a model that you estimated, the output from a machine learning algorithm, among others. Note there is no need for editing at all here. Briefly describe why the problem was challenging and how you dealt with it.