STA501: Data-based Decision Making

Lecture Note 8

Question 1. This question looks at a population of applicants to the Doctor of Medicine (M.D.) degree in 2085 who had similarly marginal credentials, which are followed up over time. Suppose that admission was offered randomly to half of the population in question. Consider the following data scientific framework:

$$\begin{aligned} Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\ P_i &= D_i P_{1i} + (1 - D_i) P_{0i}, \end{aligned} \tag{1}$$

where D_i is an indicator for person i having been admitted in 2085 to a Doctor of Medicine degree; Y_i is log annual earnings of person i in 2110; Y_{1i} is log annual earnings of person i in 2110 if they had been admitted in 2085 to a doctor of Medicine degree; Y_{0i} is log annual earnings of person i in 2110 if they had not been admitted in 2085 to a Doctor of Medicine degree; P_i is an indicator for person i working as a physician in 2110, P_{1i} is an indicator for person i working as a physician in 2110 if they had been admitted in 2085 to a Doctor of Medicine degree; and P_{0i} is an indicator for person i working as a physician in 2110 if they had not been admitted in 2085 to a Doctor of Medicine degree.

1) Could you estimate the causal effect of admission to the Doctor of Medicine degree in 2085 on log annual earnings in 2110 for person i in this population? Explain your answer briefly.

2) Could you estimate an average causal effect of admission to the Doctor of Medicine degree in 2085 on log annual earnings in 2110? Explain your answer briefly.

3) Now suppose that we regress Y_i on D_i for the people for whom $P_i = 1$. Write the probability limit of the estimated coefficient on D_i and its relation to the effect $\mathbb{E}[Y_{1i} - Y_{0i} | P_{1i} = 1]$. Explain your answer.

4) Explain intuitively (without equations) why conditioning a regression of Y_i on D_i for the people for whom $P_i = 1$ is fundamentally different from conditioning a regression of Y_i on D_i on an indicator for being born after 2062. Which approach makes more sense and why?

Suppose that the country's president wanted to limit the boundaries of medical business between doctors and pharmacists, but the union of doctors have achieved a spectacular win against the idealistic president and, since 2100, they could claim a majority of profit in medical business. College admission to medical school became more tough since 2100, according to experts in higher education.

5) Do you find any needs to split your data set before and after 2100? How can you test your claim?

6) Your colleague, a deep-learning maniac, wonders if a simple deep learning library application can create any issues. Provide your rebuttal.

Question 2. The Zurich Canton government wants to estimate the impact on future earnings of a job training program that it operated in 2100 and 2101. Access to the program is governed by an eligibility rule: only individuals whose income in the prior tax year was less than CHF 12,000 can participate.

1) Explain how you could use this eligibility rule to estimate the causal effect of the program. Describe any data you would need, write down the regression equation(s) you would estimate, define all variables precisely, and explain how you would interpret the regression results. If you make any additional assumptions, state them clearly.

2) A colleague worries that because the eligibility rule was public, your estimate of the program's causal effect may be biased. Why might this pose a problem for identification? How could you use the data to assess the validity of this concern?

The city of Bern, a neighboring Canton's capital city (with most population), has a similar program that was set for CHF 11,000 while the Gross Regional Domestic Product (GRDP) per person is slightly lower than Zurich. After a successful negotiation with the local government, you have obtained the same set of regressor variable data as you did for Zurich.

3) How would this additional data set can help you? Your boss questions you if you absolutely need this data. How would you like to answer your boss?

4) It turned out that two Cantons have quite heterogeneous business environments in terms of capital availability and international access by transportation. For example, due to the proximity to Germany, citizens of Zurich enjoy cheaper shipping costs for online purchases, and German corporations tend to set up a Swiss local office in Zurich. In addition to that, as a transportation hub of Central Europe, people in Zurich have higher mobility in jobs across central European countries. How would these conditions can affect comparative advantage of the Zurich against Bern? Can you leverage your argument to explain average wage gap between two Cantons?

5) If two Cantons have heterogeneous comparative advantage, what is your estimation strategy, assuming that you have required data? In your answer, provide relevant regressors and back up arguments.

6) Now that due to Corona-2101 outbreak, most citizens of Swiss have to be locked in their hometown. In other words, the citizens of Zurich are unable to enjoy the city's international connectivity. How would this restriction can affect the comparative advantage? How would you test the changes in comparative advantage?

7) How would average wage between two Cantons change, be there changes in comparative advantage? How would you test it, as a data scientist?

Solution.

1. Arbitrary assignment rules tend to provide excellent natural experiments, but they rarely come with flashing signs (like the word "arbitrary" to indicate their presence.) If you can catch this, you are very well on your way to being a "Real" data scientist. The truth is, the arbitrary rule isn't so much a random as it affects people's behavior in one way or another.
   The 12,000 income cutoff is ideal for considering a regression discontinuity design. We could approach this in two ways:

   - Use the full sample and run the following regression

     $$y_i = \alpha + \beta \times 1(x_i < 12,000) + f(x_i) + \epsilon_i$$

     where x_i is individual i's income in the prior tax year, the "running variable", $1(x_i < 12,000)$ is an indicator equal to 1 if this income was less than the eligibility cutoff (which perfectly determines whether the individual was eligible for the program or not); and f(.) is a function of prior income, e.g., we could model a linear, quadratic, cubic, etc. relationship between future and prior income.

   - Alternatively, we could focus on observations for individuals close to the eligibility cutoff. With this range, the functional form has little effect, so we would run either

     $$y_i = \alpha + \beta \times 1(x_i < 12,000) + \epsilon_i$$

     which would compare mean outcomes within the window on either side of the cutoff, or

     $$y_i = \alpha + \beta \times 1(x_i < 12,000) + \gamma x_i + \epsilon_i$$

     which models the relationship between the running variable and the outcome linearly.

   The key assumption for this approach to work is that the relationship between the running variable and the outcome is smooth ("there are no jumps but out jump at the jump". We could test this by looking for continuity in other chracteristics at the cutoff. It is reasonable to discuss this as saying that for those near the threshold, being slightly above/below the threshold must be random and that this implies no important differences on other characteristics on both sides of the threshold. Another assumption is no manipulation of the running variable; in this line, one could discuss that the distribution of individuals around the threshold should be smooth)

2. Because the eligibility rule was publicly known, we may be concerned that individuals who wanted to participate in the program manipulated their prior incomes to be below 12,000 either by reporting a value less than their true earnings or actually earning less than the cutoff amount in order to be eligible. If this occurred, individuals just above the cutoff would not be a valid counterfactual for those below. Those individuals who were willing to manipulate their earnings would not appear in the control group (above the cutoff) and would appear in the treatment (below). We could use the data on prior earning and other individual characteristics to see if they varied smoothly across the cutoff. E.g., if we say a discrete jump at the cutoff in the share of individuals who hand not completed secondary school, we could be concerned about out identifying assumption.

Question 3. Suppose the causal relationship between an outcome variable Y and the true values of two regressors X_1^* is given by :

$$Y_i = \beta_0 + \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + \epsilon_i$$

You may assume that ϵ_i is indepedent of X_{1i}^* and X_{2i}^*. Assume X_2^* is measured without error but the researcher only has an error-ridden measure of X_1^* available. The observed value is X_1^* and is related to the true value by the relationship:

$$X_i = X_{1i}^* + u_i$$

where u_i is independent of $(X_{1i}^*, X_{2i}^*, \epsilon_i)$. The researcher estimates by OLS the regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}^* + \nu_i$$

In answering the following questions you should try to be as formal as possible.

1) What is the likely consequence of the measurement error for the estimated coefficient on X_1? What determines the likely size of the bias? Why is the correlation between X_{1i}^* and X_{2i}^* important in determining the size of the bias?

2) What is the likely consequences of the measurement error for the estimated coefficient of X_2?

3) A researcher interested only in the coefficient on X_2 suggests it might be better to omit X_1 from the regression altogether because 'the bias for omitted variables might be less than the bias induced by including an error-ridden regressor'. Evaluate this argument.

4) Suppose that X_1^* is a binary variable (i.e. can only take the values zero or one). However, because of mis-classifications we have measurement error in our observed value X_1 (though this remains binary). Explain why this measurement error cannot be of the form described earlier in the question.

It turned out that X_1 and X_2 are final exam scores of SIAI's two basic math and stat courses given to MBA students, and Y is the pass or fail grade of graduating dissertation.

5) Rewrite the above regression in a form of linear probability model. What is the advantage of logistic regression in this case?

6) A friend of yours claims that since both regressors are highly correlated with each other, the regression may suffer from multi-collinearity. Instead of running the model with both variables, she suggests to use one of them as an instrumental variable to the other. What is your position to her argument?