STA501: Data-based Decision Making

Final Exam S2022

Question 1. A data scientist is interested in estimating the production function for fake AI business in comparision with real AI business. Both types are vaguely integrated as "High-end technology" service in most countries, but the quality of services are drastically different between real and fake AI. Though there is widespread bias that well-paid employees tend to have better knowledge in AI, due to mathematical requirements in high-end technologies and public ignorance in engineers' limited understanding in mathematics, it is not clear whether high salary is a good indicator. Instead of high salary itself, the data scientist would like to rely on final production (Y), labor's and capital's contribution (L, K) by a Cobb-Douglas specification:

$$Y_i = exp(\beta_0) \cdot L_i^{\beta_L} \cdot K_i^{\beta_K} \cdot exp(u_i),$$

where Y_i is a measure of output of each AI firm i, and u_i is an unobserved term that captures technological or managerial efficiency and other external factors (e.g., weather). The parameters to be estimated are $(\beta_0, \beta_L, \beta_K)$.

(1) Interpret β_0, β_L and β_K for fake and real AI businesses. (5 marks)

(2) Assume that you have a cross-section of independent firms and that real AI firms hire less workers (labor). Explain why OLS would not provide consistent estimates for $(\beta_0, \beta_L, \beta_K)$. Would it over- or under-estimate β_L on average? Clearly explain your answer. (5 marks)

(3) As a data scientist with growing abomination to fake AI, you would like to argue that β_L is underestimated due to endogeneity of the model. What is your strategy? Provide an argument with data scientific background. (5 marks)

(4) Given (3), name any possible instrumental variable, and back up your argument. (5 marks)

(5) Describe in detail how you would estimate the parameters of the production function using Two Stage Least Squares (2SLS). What restrictions would be necessary for this researcher to successfully use this instrumental variable in the estimation of the parameters $(\beta_0, \beta_L, \beta_K)$ and what would you need to assume about capital stock? (5 marks)

(6) If average wages per firm do not vary much by firm (potentially because of inability to discern difference in true analytical skill in data science), how would this affect the properties of the estimation procedure suggested in (5)? Explain your answer. (5 marks)

(7) Assume that there is a litmus paper like test that shows true quality of data scientists. Does the new information, assuming the companies accept the signal, of the analysis can help removing necessity of 2SLS? (5 marks)

(8) Assume that the signal is not accepted to company officials. How does this affect your regression analysis? (5 marks)

(9) How would you statistically test the difference in estimators, if there is any? Be more specific about the test steps. (5 marks)

(10) A boss of yours, a deep-learning maniac with zero statistical (in fact, any scientific) training, claims that a model with L and K is only imaginary, thus pointless, the very example of "fake AI". He continues that it is better to simply dump all your data into a computer-based model (any deep-learning model, for example), a coding library that he keeps calling "Artificial Intelligence". As a trained data **scientist**, how would you respond? (5 marks)

Question 2. This question looks at a population of applicants to the MBA in AI/BigData degree at SIAI 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, but note that the population itself is biased to a group of potential students who are dedicated to learn "real" data science. 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 MBA 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 MBA; Y_{0i} is log annual earnings of person i in 2110 if they had not been admitted in 2085 to MBA; P_i is an indicator for person i working as a data scientist in 2110, P_{1i} is an indicator for person i working as a data scientist in 2110 if they had been admitted in 2085 to MBA; and P_{0i} is an indicator for person i working as a data scientist in 2110 if they had not been admitted in 2085 to MBA.

1. Could you estimate the causal effect of admission to the MBA in 2085 on log annual earnings in 2110 for person i in this population? Explain your answer briefly. (5 marks)

2. Could you estimate an average causal effect of admission to the MBA in 2085 on log annual earnings in 2110? Explain your answer briefly. (5 marks)

3. Note that your population could be exposed to selection bias. How does this affect your average treatment effect? (5 marks)

4. 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. How does the selection bias affect your story? (5 marks)

5. 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? (5 marks)

Suppose that the country's left-wing government that took the office in 2100, without detailed blueprint, had provided heavy subsidy to computer programming private education sector for 5 years, which they claimed as "investment in Artificial Intelligence". Despite public condemnation by experts, the public funding seeded the coding business, and raised coders' salary for some time.

6. Do you find any needs to split your data set before and after 2100? If so, how can you test your claim? Re-write the entire test setting given the new information. (10 marks)

After 5 years, a right-wing party took over the office and eliminated all the inefficient subsidy done by the earlier administration. Coding private academies, without public funding, gradually disappeared.

7. How does this affect your analysis in 6)? Do you find any needs to split your data set before and after 2100 and 2105, respectively? If so, re-write the entire test setting given the new information.(5 marks)

8. How can you model the transition period when the previous government's budgeting still provides funding, and brain-washed students come for coding study for a few following years, until they completely disappear? (10 marks)

Bonus. Why is simple computational approach does not work on this case? Note that the least computational approach is simple OLS and one of the heaviest is Deep Neural Network (DNN), but the consequence in this case is not different. (Upto 5 marks)