STA501: Data-based Decision Making

Final Exam

Question 1. A data scientist is interested in estimating the production function for Wordpress-based webpages, one of the IT jobs that requires minimal computer programming skills, which is postulated to follow 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 a multi-nationally funded company's webpage production for firm i, L_i is labor in the country, K_i is capital stock and u_i is an unobserved term that captures technological or managerial efficiency and other external factors (e.g., weather). Note that our company universe consists of web publishing companies with multi-national operation. The parameters to be estimated are $(\beta_0, \beta_L, \beta_K)$.

(1) Interpret β_L and β_K for such a low skilled business in IT. Between developed and developing countries, on average, how would β_0, β_L, and β_K be different? (5 marks)

(2) Assume that you have a cross-section of independent firms and that more productive 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 representing labor union, you would like to argue that β_L is under-estimated 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 unionization or high mobility of the labor force), how would this affect the properties of the estimation procedure suggested in (5)? Explain your answer. (5 marks)

(7) Another data scientist representing the company argues that it would have been more profitable to set up an off-shore office with a cheaper labor in AI production. Given that, the data scientist claims that the labor union asks too much raise in wage upto the level that gives the executives an incentive to seriously consider cross-border operation. How would you form your counterargument? Does the new formation of the analysis can help removing necessity of 2SLS? (5 marks)

(8) For the past three years, due to Corona pandemic, mobility of the labor force has been significantly affected across the world. What will be the impact to the regression coefficients, assuming that you have found a way to overcome endogeneity? (5 marks)

(9) How would you statistically test the difference in estimators, if there is any? (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. 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) A cynical mocking will be rewarded extra points. (Upto 5 marks)

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. (5 marks)

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? (5 marks)

The city of Genève, a neighboring Canton's capital city (with most population), has a similar program that was set for CHF 13,000 while the Gross Regional Domestic Product (GRDP) per person is slightly higher 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? (5 marks)

4) It turned out that two Cantons have quite heterogeneous business environments in terms of capital availability and international access by transportation. For example, as a transportation hub of Central Europe, people in Zurich have higher mobility in jobs across central European countries, whereas, Genève, as a financial centre of central Europe provides easier access to capital. How would these conditions can affect comparative advantage of the Zurich against Genève? Can you leverage your argument to explain average wage gap between two Cantons? (5 marks)

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. (5 marks)

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? (5 marks)

7) How would average wage between two Cantons change, may there be changes in comparative advantage? How would you test it, as a data scientist? (5 marks)

8) In year 2102, there has been a global financial crisis that impacted Swiss banking system first time in the history of Switzerland. The surprise puts a force among international investors who begin questioning the stability of Swiss CHF. How would such a change in financial environment affect the comparative advantage? How would you test the change in the year of 2112? (5 marks)

9) Instead of measuring the impact on future earnings, you are given to test the validity of the job training in terms of unemployment rate. How do you change the model? (5 marks)

10) In your model in 9), how would one city's data help measuring the other, assuming the quality of job training program is nearly identical? Should it be? What if it is not? (5 marks)