STA501: Data-based Decision Making

Final Exam F2022

Question 1. You are a data science team leader of a unicorn start-up. Though your team is regarded as the most qualified data science team in the market, recently after launching a new project, you have lost 3 out of 8 team members. They claimed that they did not like to do 'ordinary' data science stuff. They wanted to do 'top-notch' stuff. You are well aware that they are more than qualified data scientists, but as a team leader, you find it more productive to give those tasks to other members. Now with insufficient manpower, the other 5 members given 'ordinary' tasks also complain. Given the situation, as a team member, you can see through that your team needs 'weaker' data scientists whose jobs do not have to be 'top-notch'. Your boss rejected your request, because he believes your team is the icon of the unicorn start-up, and it must preserve 'purity' of 'brain'.

To persuade the boss, you have created a labor composition model with Cobb-Douglas specification, assuming that your boss's economics training from his undergrad is still effective. The model of which is composed of final production, Y, low-skilled labor's and high-skilled labor's contributions, L, H, respectively. You have collected your market's data science firms' stats:

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

where Y_i is a measure of output of each data science 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 $(\beta_0, \beta_L, \beta_H)$.

1. Interpret β_0, β_L and β_H. (5 marks)

2. How would you theoretically back-up your decision to assign 'ordinary' tasks to certain members of the team? Assume that labor law that has imposed maximum 52-hours of work per week is abolished. Then how's your logic change? (5 marks)

3. Assume that you are unsure of other companies' high-skilled labor. They might not be high-skilled, given that the market has shown long practice of inability to differentiate the quality of brain. Explain why OLS would not provide consistent estimates for $(\beta_0, \beta_L, \beta_H)$. Would it over- or under-estimate β_H on average? Clearly explain your answer. (5 marks)

4. As a data scientist with growing abomination to fake 'high-end data scientists', you would like to argue that β_H is under-estimated due to endogeneity of the model. What is your strategy? Provide an argument with data scientific background. (5 marks)

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

6. 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 $(\beta_0, \beta_L, \beta_H)$ and what would you need to assume about both low- and high-ends labor? (5 marks)

7. You have so far assumed that labor market is efficient that wages are determined by the true ability, which is reflected in your specification of L and H. 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 (6)? Explain your answer. (5 marks)

8. 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)

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

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

Bonus. If you are one of those data scientists assigned 'ordinary' tasks, would you also leave the firm? (10 marks)

Question 2. Suppose the causal relationship between an outcome variable Y and the true values of two regressors X_1^* and X_2^* 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_1^* is measured without error but the researcher only has an error-ridden measure of X_2^* available. The observed value is X_2 and is related to the true value by the relationship:

$$X_{2i} = X_{2i}^* + u_i$$

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

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

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_2? 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? (5 marks)

2. What is the likely consequences of the measurement error for the estimated coefficient of X_1? (5 marks)

3. A researcher interested only in the coefficient on X_1 suggests that it might be better to omit X_2 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. (5 marks)

4. Suppose that X_2^* 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_2 (though this remains binary). Explain why this measurement error cannot be of the form described earlier in the question. (5 marks)

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? (5 marks)

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? (5 marks)

7. Since overall scores for X_2 were distinctively lower than X_1, the professor has decided to add bonus scores. He is wondering between moving mean up while keeping variances and re-grading with easily guideline. In each case, how does this affect measurement error problem discussed in 1)-4)? (5 marks)

8. It turned out that after the first exam (X_1), many incompetent students left MBA AI/BigData program and switch to much easily ones. They no longer have exam scores on the 2nd basic math and stat course. How would you manage the missing data in your regression? (5 marks)

9. A boss of yours, a deep-learning maniac with zero statistical (in fact, any scientific) training, claims that a model that requires 'instrumental variable' is affected by 'human', thus worse than 'AI' outcome. 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? (10 marks)