MSc Data Science Math & Stat Preparation Examination - Spring 2022

Instructions to candidates
Time allowed: 2 hours
This exam contains 2 questions. Each one carries 50% of total marks.
Calculators are **NOT** allowed (nor needed) in this examination.

Question 1. Let $ds10$ denote the percentage of students at SIAI receiving passing score on data science. We are interested in estimating the effect of per student spending on data science performance. A simple model is

$$ds10_i = \beta_0 + \beta_1 \log(expend_i) + \beta_2 \log(enroll_i) + \beta_3 math_i + u_i$$

where $math_i$ is the percentage of students have taken linear algebra. There are some transfer students from less sophisticated schools where quality of math education is questionable. In other words, you are faced with the fact that data is partly unavailable or measured with error on a key variable: $math$. You do have information available on a closely related variable: the students' SAT score for math, sat_i.

1. In this question we want to use sat_i as a proxy for $math_i$ that we consider running the regression

$$\text{Model 0: } ds10_i = \gamma_0 + \gamma_1 \log(expend_i) + \gamma_2 \log(enroll_i) + \gamma_3 sat_i + \eta_i$$
$$\text{Model 1: } ds10_i = \gamma_0 + \gamma_1 \log(expend_i) + \gamma_2 \log(enroll_i) + \gamma_4 \log sat_i + \zeta_i$$

where both η_i and ζ_i follow i.i.d. $N(0,1)$, $\gamma_3, \gamma_4 \in \mathbb{R}$, and we assume that the following relationship exists

$$math_i = \alpha_0 + \alpha_1 sat_i + v_i$$

where α_1 has some positive number.

1) Briefly discuss why sat_i is a sensible proxy variable for the variable in question, $math_i$. [5 marks]

2) Discuss the assumptions you need to make that enables consistent parameter estimation on β_1 and β_2 using your estimable equation for $ds10_i$. Will your estimates of β_1 and β_2 be unbiased as well? [5 marks]

3) How would you choose a model between $M = 0$ and $M = 1$? [5 marks]

4) The OLS results with and without sat_i as an explanatory variable are given by (standard errors in parentheses):

$$\widehat{ds10}_i = \underset{(26.72)}{-69.24} + \underset{(3.30)}{11.13}\log(expend_i) + \underset{(0.615)}{0.022}\log(enroll_i),$$
$$N = 428, \quad R^2 = 0.0297$$
$$\widehat{ds10}_i = \underset{(24.99)}{-23.14} + \underset{(3.04)}{7.75}\log(expend_i) - \underset{(0.58)}{1.26}\log(enroll_i) - \underset{(0.036)}{0.324}sat_i,$$
$$N = 428, \quad R^2 = 0.1893$$

Explain why the effect of expenditures on $ds10_i$ is lower in the regression where sat_i is included than where it is excluded. [5 marks]

5) If your choice model was in favor of Model 1 in 3), what would be the γ_4's contribution to R^2 in 4)? [5 marks]

6) Your research assistant for data collection confessed that the entries for $enroll_i$ are partly made up number due to unavailability. How does this affect your estimation of γ_2? Does it affect other estimates? [5 marks]

7) A neighboring school with questionable quality teaching in data science approaches to SIAI that they have 10 times more students for the same data set. A researcher from that school claims that $N = 428$ is not a big data, but with more than 4,000 students, the bigger data can significantly raise R^2. Provide your rebuttal with pertinent differentiation between larger set of data and "BigData". [5 marks]

8) If the other school's data is given, in what way will you exploit the information? Can your exploitation potentially increase R^2? What are the necessary conditions? If not, why? [5 marks]

9) Given the low R^2 got improved by an additional factor of sat_i, one researcher presents an idea that more data that reflects high school or even earlier education may help removing omitted variable bias. Do you agree with the researcher? Discuss. [5 marks]

10) Can deep-learning like non-parametric method can be any help for this case? Do you expect R^2 will be significantly different? Provide your logic. [5 marks]

Question 2. A researcher from SIAI using cross-sectional data hypothesizes that two variables Y and X are jointly determined by a simultaneous equations model consisting of the following two relationships:

$$Y = \beta_1 + \beta_2 X + \beta_3 Z + u \qquad (1)$$
$$X = \alpha_1 + \alpha_2 Y + v \qquad (2)$$

where Z may be assumed to be an exogenous variables and u and v are identically and independently distributed disturbance terms with zero means. The observations for Z are drawn from a fixed population with finite mean and variance.

1) Derive the reduced form equation for Y [2 marks]

2) Demonstrate that the OLS estimator of α_2 is, in general, inconsistent. How is your conclusion affected in the special case $\beta_2 = 0$? How is your conclusion affected in the special case $\alpha_2\beta_2 = 1$? [6 marks]

3) Demonstrate that the instrumental variables (IV) estimator of α_2, using Z as an instrument for Y, is consistent. Why do you need an IV estimator? [6 marks]

4) Instead of using IV, the researcher decides to use 2-Stage-Least-Square (2SLS) in the expectation of obtaining a more efficient estimator of α_2. He fits the reduced form equation for Y:

$$\hat{Y} = h_1 + h_2 Z \qquad (3)$$

saves the fitted values, and uses them as an instrument for Y in equation (2). Demonstrate that the 2SLS estimator is consistent. [6 marks]

5) Determine whether the researcher is correct in believing that the 2SLS estimator is more efficient than the IV estimator. [5 marks]

6) How do you prove that IV (or 2SLS) estimation is superior to OLS? [5 marks]

7) Now that another researcher from an engineering department claims that adding a quadratic term, instead of IV or 2SLS, can be a better estimation strategy, because he believes there is a hidden non-linearity in the model. Build a statistical test and provide your argument based on your projection. [10 marks]

8) Now that a software engineer claims that running a deep-learning model, instead of whatever statistical model, is far more superior calculation strategy, because he believes non-parametric estimation by computers are better than human logic, as was witnessed by Alpha-Go. Provide your rebuttal. [10 marks]