STA502: Math & Stat for MBA I
Final Exam F2022

Question 1. As a recent graduate of SIAI's famous MBA in AI/BigData, you just got a job at an e-Commerce start-up. Amid heavy pressure and high expectation, your data science team leader asked you if you can prove his "unconventional" result by "machine learning" on demand and supply by price and quantity. As a trained engineer with so much condemnation on anything in "words", he does not believe anything written on economics textbook, the unconventional result of which seemingly confirms his condemnation. His demand curve has upward slope! (Note that, in economics, unless you are dealing with Giffen or luxury goods, demand curve always shows downward slope.)

With well-balanced training between science and real-world from SIAI, now it is your turn to tell him such joint estimation is exposed to simultaneity problem where a simple engineering approach based on fanatical belief on machine learning fails to show true nature of data generating process. Using cross-sectional data, you hypothesize that two variables P and Q are jointly determined by a simultaneous equations model consisting of the following two relationships:

$$Q = \alpha_1 + \alpha_2 P + \alpha_3 M_2 + u \qquad (1)$$
$$P = \beta_1 + \beta_2 Q + v \qquad (2)$$

where M_2 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 M are drawn from a fixed population with finite mean and variance.

1. Derive the reduced form equation for Q. (5 marks)

2. Demonstrate that the OLS estimator of β_2 is, in general, inconsistent. How is your conclusion affected in the special case $\alpha_2 = 0$? How is your conclusion affected in the special case $\alpha_2\beta_2 = 1$? What do these special case mean in words? (5 marks)

3. Demonstrate that the instrumental variables (IV) estimator of β_2, using M_2 as an instrument for Q, is consistent. Why do you need an IV estimator? (5 marks)

4. Instead of using IV estimation, the researcher decides to use 2-Stage-Least-Square (2SLS) in the expectation of obtaining a more efficient estimator of β_2. He fits the reduced form equation for Q:

$$\hat{Q} = k_1 + k_2 M_2 \qquad (3)$$

saves the fitted values, and uses them as an instrument for P in equation (2). Demonstrate that the 2SLS estimator is consistent. (5 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. If you have M_1 for equation (2), as is M_2 for equation (1), can you have any better result? If so, in what context? Can you argue more instruments promise better results? (5 marks)

8. Can you extend your logic in 6) to disprove a claim that machine learning model is superior to OLS? Assume that your model's errors, even after 1st-stage data pre-processing, follow Gaussian distribution jointly. What happens if non-Gaussian? (5 marks)

9. Having been benchpressed by your logic, your boss, with a firm belief on machine learning, claims that adding a quadratic term, instead of IV or 2SLS, is far more superior estimation strategy, because he believes non-linear & non-parametric estimation by computers are better than human's faulty logical thinking, as was witnessed by Alpha-Go and abundant achievements by "Artificial Intelligence". Provide your rebuttal. (10 marks)

Question 2. The lack of access to quality education on data science may be an important impediment for the growth of AI start-ups. Unless these ventures are lucky to have Silicon Valley's top-league data scientists, they may have to hire sub-tier employees and outsource training services such as SIAI's MBA AI/BigData, the most famous online AI/BigData program across the world.

The following regressions are for 9,125 AI start-ups which started operations in 2050. The data science strategy of the study is to compare various business outcomes (assets, sales, and number of employees) for SIAI-trained and no-SIAI-trained in 2055. The regressions also interact SIAI alumns' status with labor market's appreciation of the quality which is reflected in wage growth, W_g. Assume that the labor market of the country is efficient, that higher wage means higher productivity, at least in the field of data science. Wage growth is coded so that growth of 5% would be 0.05.

	Dependent variable		
	$ln(Assets)$	$ln(Sales)$	$ln(Employment)$
	(1)	(2)	(3)
D_h	0.089 (0.027)	−0.131 (0.026)	−0.108 (0.015)
$D_h \times W_g$	1.21 (0.18)	0.94 (0.17)	0.37 (0.11)
W_g	0.58 (0.28)	0.29 (0.26)	0.21 (0.22)

where D_h is for Dummy for SIAI alums, and W_g is for wage growth rate for SIAI alums. Standard errors are displayed in parentheses. All regressions also contain a constant term.

1. Explain why a simple regression of business outcomes on the SIAI-training alone may not answer the question data scientists are interested in. (5 marks)

2. Explain how the use of wage growth of SIAI alums may circumvent the problem you described in 1). What's the interaction term's function in words? (5 marks)

3. Explain verbally what the coefficient of 0.089 on the dummy for SIAI-alums in column (1) means. (5 marks)

4. If wage growth is 10 percentage points higher, how much higher are the sales of no-SIAI-trained companies in the sample on average? Explain whether this effect is statistically different from zero. (5 marks)

5. What do you conclude from the results in the table about the effect of SIAI training on AI start-ups' outcomes? (5 marks)

6. Suppose you also have data for assets, sales, and employment in these start-ups in 2060. Suppose you were to run analogous regressions with these dependent variables to the regressions in the table above. Explain how the new regressions would help you interpret the results above

7. As more and more SIAI graduates flow into the labor market, given the growing competition for top-minds, SIAI separated the training program to Standard and Advanced in 2055. Companies pay more to Advanced track students, thus the wage growth rates are now WA_g and WS_g for Advanced and Standard, respectively. How does this change affect your analysis in 6)? (5 marks)

8. Given the change of regime in 2055, you would like to see whether the split of the program helped companies. How do you formulate your data scientific test? (5 marks)

9. You have an engineering background boss whose understanding of data science is no better than collection of Github codes. He claims that deep-learning can solve every data science problems that no human logic is needed. He adds that your argument does not rely on 'the most recent and advanced deep-learning practices done by top-notch companies and researchers'. Provide your rebuttal. (10 marks)

Bonus. Assume that you are class of 2055-2056, but graduated Standard track of SIAI's MBA AI/BigData. You failed the admission exam, and while in school, you were not brave enough to challenge the Advanced track exams. Now given 8), you have a temptation to go back to school and re-try the upper track. If successful, you can enjoy higher wage and better appreciation of the market. Given your personal estimation of success rate, formulate your argument. (10 marks)