STA502: Math & Stat for MBA I
Final Exam S2022

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:

$$P = \alpha_1 + \alpha_2 Q + \alpha_3 Z_1 + u \qquad (1)$$
$$Q = \beta_1 + \beta_2 P + v \qquad (2)$$

where Z_1 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 P. (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 $\alpha_2\beta_2 = 1$? (5 marks)

3. Demonstrate that the instrumental variables (IV) estimator of β_2, using Z_1 as an instrument for P, 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 P:

$$\hat{P} = m_1 + m_2 Z_1 \qquad (3)$$

   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 Z_2 for equation (2), as is Z_1 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)

Bonus. Your boss is an engineer maniac, but also a rational person. He understood your points. But, he still be-
lieves machines, with much more computing power, can solve the simultaneity problem by itself. Provide
your rebuttal. (5 marks)

Question 2. SIAI, the most famous online university across the world, has branch schools in 124 cities. Stu-
dents can physically attend classes in their own city, and classrooms are connected through global satellite
network, called SIAI-link. Such simultaneity in attendance is also extended to virtual classrooms given to stu-
dents living in suburban areas where connection to city classrooms are limited.

Amid growing complaints from suburban students that limited access harness them with less opportunity
in communication, SIAI launched a research project to prove whether virtual reality classrooms affect students
performance. Researcher A has obtained data for the students from all cities and would like to study the effect
of watching lectures on VR classrooms on the students' exam scores.

During the data collection, she has come across an behavior that a number of students create study groups
and they also use the virtual classrooms for study group meeting. Initially she hypothesized that only subur-
ban students would do the group meetings in the VR classroom, but it turned out even the in-city students
frequently use VR classrooms.

Given this behavior, for each student, she has the exam score (0 - 100, 70+ is an A, 40 and below a fail),
which she uses as her left hand side variable, the fraction of lectures watched online, and how many times the
student visited the VR classrooms for study group meeting. Researcher A obtains the following regression results:

Regressor	(1) - OLS	(2) - IV	(3) - IV	(4) - IV	(5) - IV
Fraction online	-4.91	-2.08	-0.56	-0.75	6.09
	(0.11)	(0.22)	(0.40)	(0.90)	(2.22)
VR study group	–	–	0.91	–	2.09
			(0.08)		(0.43)
Sample	All	All	All	VR rental	VR rental

All regressions also contain a constant term and a dummy variable for each city branch school. Robust
standard errors are reported in parentheses.

1. What is the interpretation of the coefficient in column (1). If this were a causal effect, would it be big or
   small? Explain whether this estimate is likely to have a causal interpretation. (5 marks)

2. Researcher B observes that some suburban students have slower internet speed, which disables them to
   have easier connection to VR classrooms. He suggests to use the internet speed as an instrumental variable
   for the fraction of lectures watched online. Is this a valid argument for IV? (5 marks)

3. Results for this IV regression are displayed in column (2). Explain why instrumental variables may pro-
   duce a better estimate of the causal effect of watching lectures online, and which assumptions need to be
   satisfied for this to be the case. Discuss the validity of the assumptions in this case. Can you ascertain
   whether any of these assumptions are true from the results in the table above? (5 marks)

4. Researcher C points out to B that given the cost of SIAI's tuition, buying a premium internet speed
   service is not really a decisive concern. Instead, he believes which VR set a student uses can be a much
   more critical factor. A good VR machine costs \$\$\$, but he believes dedicated students should buy more
   expensive VR machine sets. How does this information affect your assessment of the IV strategy? (5
   marks)

5. Researcher D realizes that the data also include a variable for the number of times a student has used VR
   study group meetings. He suggests to rerun the instrumental variables regression adding this variable as
   a control. Results for this regression are displayed in column (3). Assess D's strategy. (5 marks)

6. Researcher E notices that there are five city branches which offers top-quality VR rental service. She suggests to run the IV model from columns (2) and (3) for the subsample of students from these branches only. Results are displayed in columns (4) and (5). Assess E's regressions. (10 marks)

7. Drawing on the results in the table above, what have you learned from this exercise about the causal effect of watching lectures online on students' exam results? (5 marks)

8. Thanks to a tech giant which wants to lead the VR industry, high-end VR machines become incredibly cheap. What will be the consequence to above estimation model? (5 marks)

9. In the presence of such cheap and high quality VR, the board of SIAI has come to a conclusion that they no longer need to provide physical classrooms. Students are encouraged to study at their home and connect to SIAI-link. If regression model (5) has high enough R^2, in terms of students performance in exams, what do you expect in the following year? If you were the chairman of SIAI, what would you do? (10 marks)