SUMMER TERM 2022
CENTRALLY-MANAGED ONLINE EXAMINATION
ECON0019: QUANTITATIVE ECONOMICS AND ECONOMETRICS

Time allowance
You have 3 hours to complete this examination, plus an Upload Window of 20 minutes. The Upload Window is for uploading, completing the Cover Sheet and correcting any minor mistakes and should not be used for additional writing time.

If you have been granted SoRA extra time and/ or rest breaks, your individual examination duration will be extended pro-rata and you will also have the 20-minute Upload Window added to your individual duration.

All work must be submitted anonymously in a PDF file and you should follow the instructions for submitting an online examination in the AssessmentUCL Guidance for Students.

If you miss the submission deadline, you will not be able to submit your work via AssessmentUCL and you will not be permitted to submit the work via email or any other channel. If you are unable to submit your work due to technical difficulties which are substantial and beyond your control, you should apply for a Deferral via the AssessmentUCL Query Form.

Page limit: 8 pages.
Your answers, excluding the Cover Sheet, should not exceed this page limit. Please note that a page is one side of an A4 sheet with a minimum margin of 2 cm from the top, bottom, left and right borders of the page. The submission can be handwritten or typed, but the font size should be no smaller than the equivalent to an 11pt font size. This page limit is generous to accommodate students with large handwriting. We expect most of the submissions to be significantly shorter than the set page limit. If you exceed the maximum number of pages, the mark will be reduced by 10 percentage points, but the penalized mark will not be reduced below the pass mark and marks already at or below the pass mark will not be reduced.

Answer ALL TWO questions from Part A and answer ONE question from Part B.

Questions in Part A carry 60 ***per cent of the total mark and questions in Part B carry*** 40 ***per cent of the total. Tables for the normal and F-distribution are at the end of the examination paper.***

In cases where a student answers more questions than requested by the examination rubric, the policy of the Economics Department is that the student's first set of answers up to the required number will be the ones that count (not the best answers). All remaining answers will be ignored.

If you have a query about the examination paper, instructions or rubric, you should complete an AssessmentUCL Query Form. Please note that you will not receive a response during your examination.

By submitting this assessment, you are confirming that you have not violated UCL's Assessment Regulations relating to Academic Misconduct contained in Section 9 of Chapter 6 of the Academic Manual.

PART A

Answer all questions from this section.

A.1 You wish to measure the effect of hiring more teachers on student performance. To this end you randomly select 300 schools; for each school you collect data on the number of students per teacher (*str*) and the average exam score (*score*) for final year students in 2021. Suppose that in the population the following equation holds:

$$score = \beta_0 + \beta_1 str + \beta_2 ability + \beta_3 \left(str \times ability \right) + u, \tag{1}$$

where *ability* is the average level of student ability in a given school. This equation statisfies MLR.3–MLR.4 in Wooldridge's text book.

(a) You do not have data on *ability* and so decide to estimate β_1 by regressing *score* on *str*. Derive the probability limit of the estimator.

(b) A colleague of yours conjectures that $\mathrm{Cov}\left(str, ability\right) < 0$, $\mathrm{Cov}\left(str, str \times ability\right) > 0$, $\beta_2 > 0$ and $\beta_3 < 0$. Do these sign restrictions seem plausible to you? Explain. Supposing they hold, is it possible to determine the sign of the asymptotic bias of the estimator in (a)?

(c) You speculate that the joint population distribution of *ability* and *str* is such that

$$\mathbb{E}\left[ability | str\right] = \mathbb{E}\left[ability\right]. \tag{2}$$

Interpret the restriction in (4). Do you think it is likely to hold?

(d) Assuming that (4) holds, demonstrate that the probability limit of the OLS estimator in (a) equals $\beta_1 + \beta_3 \mathbb{E}\left[ability\right]$. Interpret the probability limit. In particular, is it a meaningful measure of the effect of class size on student performance?

(e) You decide to collect additional data on the average mark that the final–year students earned in their first year at each school. With *mark* denoting this new variable, you hypothesise that, for some unknown coefficients θ, the following two conditions hold:

$$\mathbb{E}\left[score | str, ability, mark\right] = \mathbb{E}\left[score | str, ability\right], \tag{3}$$

$$\mathbb{E}\left[ability|str, mark\right] = \mathbb{E}\left[ability|mark\right] = \theta mark. \quad (4)$$

Interpret the two conditions and compare them to (4). Do they seem reasonable?

(f) You run the following regression,

$$\widehat{score} = \hat{\beta}_0 + \hat{\beta}_1 str + \hat{\beta}_2 mark + \hat{\beta}_3 \left(str \times mark\right)$$

What is the probability limit of $\hat{\beta}_1$ under (5)–(6)? Explain.

A.2 You are interested in the relationship between sales, profits and research & development (R&D). For that purpose you obtain the following regression based on data collected from a sample of 45 firms in the UK concrete industry in 2016,

$$\widehat{rd} = \underset{(1.369)}{.42} + \underset{(.116)}{.21} \log\left(sales\right) + \underset{(.046)}{.07}\, profit, \quad \bar{R}^2 = .079, \quad (7) \quad (5)$$

where rd is expenditures on R&D of a firm as percentage of its annual sales, $sales$ is the firm's annual sales (in millions GBP) and $profit$ is its annual profits as percentage of sales. Robust standard errors are reported in parentheses.

(a) Interpret the coefficient on $\log\left(sales\right)$. If $sales$ increases by 10% what is the exact estimated percentage point change in rd? Is this an economically large effect?

(b) Test the hypothesis that rd does not change with $sales$ against the alternative that it does increase with sales. Perform the test at the 5% and 10% level. What is the p-value of the test? Conclude.

(c) You compute the F-test statistic of the hypothesis that $sales$ and $profit$ are jointly insignificant and obtain $F = 4.12$. Do you accept or reject the null at the 5% level? Explain.

(d) Do you trust the critical values that you used in (b) and (c) and the the p-value that you computed in (b)? Are they valid? What do you conclude about the reported test results?

(e) You estimate the following alternative regression model for rd,

$$\begin{aligned} \widehat{rd} \quad &= \quad \underset{(1.245)}{.35} + \underset{(.014)}{.030} sales - \underset{(.00000038)}{.0000070} sales^2 + \underset{(.047)}{0.048} profit, \quad \bar{R}^2 = .099. \quad (9) \quad (6) \end{aligned}$$

At what point does the estimated marginal effect of $sales$ on rd become negative in this model?

(f) Write up a composite model that would allow you to test (7) and (9), respectively, against the composite model. Would the outcomes of these two tests be able to determine which of the two models, (7) and (9), is the preferred one? Explain.

(g) You collect data on annual R&D, sales and profits of the same 45 firms in 2017 and re-estimate (9) by running a pooled regression across the two years, 2016 and 2017. A colleague tells you that you should rather estimate the model using the first-difference estimator. Is your colleague right? Explain.

PART B

Answer ONE question from this section.

B.1 In "Rainfall and Conflict: A Cautionary Tale" (Journal of Development Economics, 2015), Heather Sarsons studies whether lower income can lead to more violent conflict among religious groups in India. She studies a sample of 142 districts in the country's 28 states. Simplifying things a bit, the baseline equation of interest is:

$$C_i = \beta Y_i + \sum_{s=1}^{28} \gamma_s \mathbf{1}[S_i = s] + \varepsilon_i,$$

where C_i ("conflict") is the number of riots in district i in a particular year, Y_i ("income") is income per capita, S_i is the state in which district i is located, $\mathbf{1}[S_i = s]$ indicates a dummy variable which takes the value of one when $S_i = s$, and ε_i is the error term.

(a) Why may the OLS estimate of β be inconsistent? Provide at least one economic justification.

income is endogenous

Sarsons proceeds to use an instrumental variable strategy: she instruments income with two measures of rainfall in the district. The first one, R_{1i}, is the amount of rainfall in district i in the year of study *minus* its typical value (across many years) for the district. The second measure, R_{2i}, is a dummy variable that R_{1i} is below its 20th percentile. The idea is that agricultural production is a key source of income in much of India and it relies on sufficient rainfall.

(b) Explain in detail (step by step) how her instrumental variable estimate $\hat{\beta}$ is constructed from data on $(C_i, Y_i, R_{1i}, R_{2i}, S_i)$. Then write down the formal conditions under which $\hat{\beta}$ is consistent for β. Which of them can be tested? For those which can, describe the testing procedure. For those which cannot, explain why not.

(c) Why does Sarsons subtract the typical rainfall in the district when constructing R_{1i}? Which condition or conditions from part (b) would be more likely violated if she did not do this? Give an economic justification.

For simplicity, drop R_{2i} and keep R_{1i} as the single instrument in the final parts of the question. Sarsons observes that in some Indian districts there are dams on local rivers, and thus reservoirs of water which do not dry out even in low-rainfall years. This should make income less dependent on weather. Another issue is that the causal effects of income on conflict can be heterogeneous across regions.

LATE

(d) With these complications, can the instrumental variable estimate $\hat{\beta}$ still be interpreted as some average of causal effects of income on conflict and, if so, what kind of average? Which condition or conditions would have to be added, compared to part (b), for such an interpretation to be valid? Is it (or are they) plausible?

(e) Sarsons finds that in the districts with a dam the reduced-form coefficient is significantly different from zero, but the first-stage coefficient is not. She concludes that rainfall may *not* be an exogenous instrument. Explain intuitively and formally how she makes this conclusion.

B.2 In "Does Hospital Crowding Matter? Evidence from Trauma and Orthopedics in England" (American Economic Journal: Economic Policy, forthcoming), Thomas Hoe examines the impact of hospital crowding on medical treatment outcomes exploiting variation in emergency admissions. For simplicity, we abstract from some of the details examined in the paper. One possible outcome of interest is the length of the illness (measured in days) for a particular patient i. Let this be denoted by y_i. Suppose one focusses on the following model relating this outcome for an individual i admitted to a particular hospital:

$$\ln(y_i) = \beta_0 + \mathbf{x}_i'\beta_x + \epsilon_i, \tag{7}$$

where $\mathbf{x}_i$ comprises variables such as the individual's age, race and disease stage at the time of admission. As indicated in the article, other elements that might affect the outcome of interest relate to patient composition and hospital operation details (such as capacity constraints and utilisation) at the hospital where the individual is admitted, among other factors. Note that the unit of observation in items (a)-(c) below are the individual whereas in items (d)-(e) relates to a time period (day) for a particular hospital.

(a) Suppose that one is interested in estimating (11) with data from a particular hospital. Let $c_i > 0$ denote the number of days an individual is at the hospital. This variable and $\mathbf{x}_i$ is observed for every patient in the hospital. If individual i is discharged after recovering from the illness (i.e., $c_i \geq y_i$), y_i is known, but otherwise we only know c_i and that $y_i > c_i$. What additional assumptions would one need to estimate (11) by maximum likelihood? Write down the log-likelihood function for this regression and explain your answer.

(b) Suppose that individual i opts to go to a hospital according to the following choice model:

$$h_i = \mathbf{1}(\gamma_0 + \gamma_d d_i + v_i \geq 0) \tag{8}$$

where $h_i = 1$ if individual i goes to the hospital and $= 0$, otherwise. The variable d_i records i's distance to the closest hospital and v_i marks idiosyncratic unobservable factors informing this decision. Assume that v_i follows a standard normal distribution. One is interested in estimating (11) and information on y_i is only available for those who go to the hospital. Assume that OLS estimates are obtained for those observations. First, if ϵ_i and v_i are not necessarily independent, but $\gamma_d = 0$, would the OLS estimator above be consistent? Explain. What if γ_d is not necessarily zero, but d_i and $\mathbf{x}_i$ are independent? Explain.

(c) Suppose that going to the hospital depends not only on d_i but also on $\mathbf{x}_i$. In other words, consider now the following extended version of equation (12):

$$h_i = \mathbf{1}(\gamma_0 + \gamma_d d_i + \gamma_x' \mathbf{x}_i + v_i \geq 0) \tag{9}$$

How would you estimate the parameters in (11) consistently? Explain your answer.

(d) Suppose you have time series data on daily admissions to a particular hospital, denoted by q_t where t is a particular date, and consider for simplicity a linear regression of q_t on week-of-the-year dummy variables $\mathbf{s}_t$ recording which week of the year t pertains to. Using the number of admissions per hospital, Hoe notes that there is no evidence of serial correlation in the residuals of a regression of (emergency) admissions on seasonal dummies once one examines the estimated AR(1) coefficients for a regression of residuals on lagged residuals. Under what conditions does this residual regression offer a valid test for the absence of serial correlation? Describe the test in detail.

(e) For the regression above, under what conditions would the OLS estimator be consistent? Explain your answer.

(d) No original regressors in the auxiliary regression, so this is not the Durbin alternative test under no strict exogeneity
CANNOT be the DW test because DW requires no second-stage regression
so this test can only be the general test

5 % Critical values for the F_{ν_1,ν_2} distribution

$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	7	8	10	12	15	20	30	50	∞
1	161	199.	216.	225.	230.	234.	237.	239.	242.	244.	246.	248.	250.	252.	254.
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.70	8.66	8.62	8.58	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.86	5.80	5.75	5.70	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.62	4.56	4.50	4.44	4.36
10	4.96	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.38	2.31	2.23	2.16	2.07	2.00	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.35	2.28	2.20	2.12	2.04	1.97	1.84
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.16	2.09	2.01	1.93	1.84	1.76	1.62
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.99	1.92	1.84	1.75	1.65	1.56	1.39
80	3.97	3.11	2.72	2.49	2.33	2.21	2.13	2.06	1.95	1.88	1.79	1.70	1.60	1.51	1.32
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.93	1.85	1.77	1.68	1.57	1.48	1.28
120	3.91	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.91	1.83	1.75	1.66	1.55	1.46	1.25
∞	3.85	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.83	1.75	1.67	1.57	1.46	1.35	1.00

NORMAL CUMULATIVE DISTRIBUTION FUNCTION ($Prob(z < z_a)$ where $z \sim N(0,1)$)

z_a	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995