SUMMER TERM 2024
DEPARTMENTALLY-MANAGED REMOTE ONLINE EXAMINATION
ECON0019: QUANTITATIVE ECONOMICS AND ECONOMETRICS

Assessment Component: 80% Remote Online Controlled Condition Examination

Time Allowance: You have 3 hours to complete this examination, plus an additional collation time of 20 minutes and an Upload Window of 20 minutes. The additional collation time has been provided to cover any additional tasks that may be required when collating documents for upload, and the Upload Window is for uploading and correcting any minor mistakes. The additional collation time and Upload Window time allowance should not be used for additional writing time.

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

If you miss the submission deadline during the 40-minute Late Submission Period and do not receive approved mitigation for the circumstances relating to your late submission, a Late Submission Penalty will be applied by the Module Administrator. At the end of the Late Submission Period, you will not be able to submit your work via Moodle and you will not be permitted to submit work via email or any other channel.

All work must be submitted anonymously in a single PDF file. Do not write your name and student number in either the file or the file name. The file name must be in the following format: Module Code-80%Exam i.e. ECON0019-80%Exam.

Page Limit: 8 pages. Your answers 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 penalised mark will not be reduced below the pass mark and marks already at or below the pass mark will not be reduced.

Academic Misconduct: 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.

Number of Questions Answered Policy: 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.

QUESTIONS:

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 mark. Tables for the normal and F-distribution are at the end of the examination paper.***

PART A

Answer all questions from this section.

A.1 Consider a given population that we are following over time. You have collected data on n randomly chosen individuals from the population over two time periods. For individual i $(= 1, ..., n)$, you observe (y_{it}, x_{it}), $t = 1, 2$. Suppose that y_{it}, is related to x_{it}, as follows:

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}, \tag{1}$$

where u_{it} captures other unobserved factors influencing y_{it}, $t = 1, 2, 3, ...$ In the following, assume that $\mathbb{E}[u_{it}] = 0$, $\text{Var}(x_{it}) > 0$ and $\mathbb{E}\left[(x_{it+1} - x_{it})^2\right] > 0$, $t = 1, 2, ...$

(a) You decide to estimate β_0 and β_1 in (**??**) by the solution to the following least–squares problem,

$$\min_{b_0, b_1} \sum_{i=1}^{n} \sum_{t=1}^{2} (y_{it} - b_0 - b_1 x_{it})^2 .$$

Suppose that

$$\mathbb{E}[u_{it} x_{it}] = 0, \quad t = 1, 2. \tag{2}$$

Show that your chosen estimator of β_1 is consistent.

Answer: *The estimator takes the form*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \sum_{t=1}^{2} (x_{it} - \bar{x}) y_{it}}{\sum_{i=1}^{n} \sum_{t=1}^{2} (x_{it} - \bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^{n} \sum_{t=1}^{2} (x_{it} - \bar{x}) u_{it}}{\sum_{i=1}^{n} \sum_{t=1}^{2} (x_{it} - \bar{x})^2}.$$

By the LLN together with (??), $E[u_t] = 0$ and $Var(x_t) > 0$ as assumed

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x})^2 \quad \to \quad {}^{p}\sum_{t=1}^{2}\mathbb{E}\left[(x_t-\mu_x)^2\right] = \sum_{t=1}^{2} Var(x_t) > 0,$$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x})\,u_{it} \quad \to \quad {}^{p}\sum_{t=1}^{2}E\left[(x_t-\mu_x)\,u_t\right] = \mathbb{E}\left[x_t u_t\right] + \mu_x E\left[u_t\right] = 0$$

Combining above two displays yields $\hat{\beta}_1 \to^p \beta_1$.

(b) A colleague of yours thinks that you should rather estimate β_1 by the solution to the following alternative least–squares problem,

$$\min_{b_1}\sum_{i=1}^{n}\left(\Delta y_{i2} - b_1\Delta x_{i2}\right)^2,$$

where $\Delta y_{i2} = y_{i2} - y_{i1}$ and $\Delta x_{i2} = x_{i2} - x_{i1}$. Suppose that

$$E\left[\Delta u_{i2}\Delta x_{i2}\right] = 0. \tag{3}$$

Show that your colleague's alternative estimator of β_1 is consistent.

Answer: *The new estimator takes the form*

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}\Delta x_{i2}\Delta y_{i2}}{\sum_{i=1}^{n}\left(\Delta x_{i2}\right)^2} = \beta_1 + \frac{\sum_{i=1}^{n}\Delta x_{i2}\Delta u_{i2}}{\sum_{i=1}^{n}\left(\Delta x_{i2}\right)^2}.$$

By the LLN together with (??) and $\mathbb{E}\left[(x_2 - x1)^2\right] > 0$ as assumed,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\Delta x_{i2}\right)^2 \quad \to \quad {}^{p}\mathbb{E}\left[(x_2-x_1)^2\right] > 0$$

$$\frac{1}{n}\sum_{i=1}^{n}\Delta x_{i2}\Delta u_{i2} \quad \to \quad {}^{p}\mathbb{E}\left[\Delta x_2\Delta u_2\right] = 0.$$

Combining above two displays yields $\hat{\beta}_1 \to^p \beta_1$.

(c) Suppose that (??) holds but $\mathrm{Cov}(x_{i,s}, u_{i,t}) > 0$, $s \neq t$. Does (??) hold in this case? Explain. Suppose instead that $u_{it} = \alpha_i + v_{it}$, where $\mathbb{E}\left[v_{it}|x_{i1}, x_{i2}\right] = 0$ while $\mathbb{E}\left[\alpha_i x_{it}\right] \neq 0$. Does (??) hold in this case? Does (??) hold? Explain.

Answer: *$Cov(x_{i,s}, u_{i,t}) > 0$, $s \neq t$ implies $\mathbb{E}\left[x_2 u_1\right] + \mathbb{E}\left[x_1 u_2\right] > 0$ since $\mathbb{E}\left[u_t\right] = 0$. This in turn implies that (??) does not hold,*

$$\mathbb{E}\left[\Delta x_2\Delta u_2\right] = \sum_{t=1}^{2}\mathbb{E}\left[u_t x_t\right] + \mathbb{E}\left[x_2 u_1\right] + \mathbb{E}\left[x_1 u_2\right] = \mathbb{E}\left[x_2 u_1\right] + \mathbb{E}\left[x_1 u_2\right] > 0.$$

If $u_t = \alpha + v_t$, where $\mathbb{E}[v_t|x_1, x_2] = 0$ while $\mathbb{E}[\alpha x_t] \neq 0$. Then (??) holds,

$$\mathbb{E}[\Delta x_t \Delta u_t] = \mathbb{E}[\Delta v_t \Delta x_t] = \mathbb{E}[E[v_2|x_1, x_2]\Delta x_t] - \mathbb{E}[\mathbb{E}[v_1|x_1, x_2]\Delta x_t] = 0,$$

but ??) does not,

$$\mathbb{E}[u_t x_t] = \mathbb{E}[\alpha x_t] + \mathbb{E}[\mathbb{E}[v_t|x_1, x_2] x_t] = \mathbb{E}[\alpha x_t] \neq 0.$$

(d) Suppose that

$$\mathbb{E}[u_{it}|x_{i1}, x_{i2}] = 0 \qquad t = 1, 2. \tag{4}$$

Are the estimators of β_1 in (a) and (c) unbiased? Explain.

Answer: *Yes, they are both unbiased: Conditional on $\mathcal{X}_n = \{x_{i1}, x_{i2}\}_{i=1}^n$,*

$$\mathbb{E}\left[\hat{\beta}_1|\mathcal{X}_n\right] = \beta_1 + \frac{\sum_{i=1}^n \sum_{t=1}^2 (x_{it} - \bar{x})\,\mathbb{E}[u_{it}|\mathcal{X}_n]}{\sum_{i=1}^n \sum_{t=1}^2 (x_{it} - \bar{x})^2},$$

where, using that individuals are randomly sampled and then (??),

$$\mathbb{E}[u_{it}|\,\mathcal{X}_n] = \mathbb{E}[u_{it}|\,x_{i1}, x_{i2}] = 0,$$

Combining the above two displays, we find that the $\hat{\beta}_1$ is unbiased. Similarly,

$$\mathbb{E}\left[\tilde{\beta}_1|\mathcal{X}_n\right] = \beta_1 + \frac{\sum_{i=1}^n \Delta x_{i2}^2 \mathbb{E}[\Delta u_{i2}|\mathcal{X}_n]}{\sum_{i=1}^n \Delta x_{i2}^2},$$

where, again using random sampling and then (??),

$$\mathbb{E}[\Delta u_{i2}|\mathcal{X}_n] = \mathbb{E}[u_{i2}|\,x_{i1}, x_{i2}] - \mathbb{E}[u_{i1}|\,x_{i1}, x_{i2}] = 0.$$

(e) Maintain the assumption (??) and suppose furthermore that $\mathrm{Cov}(u_{i1}, u_{i2}|x_{i1}, x_{i2}) = 0$ and $\sigma^2 = \mathrm{Var}(u_{it}|x_{it})$, $t = 1, 2$. Show that

$$\mathrm{Var}(\hat{\beta}_1|\mathcal{X}_n) = \frac{\sigma^2}{\sum_{i=1}^n \sum_{t=1}^2 (x_{it} - \bar{x})^2}.$$

Answer:

$$\begin{aligned} Var(\hat{\beta}_1|\mathcal{X}_n) &= \frac{\sum_{i=1}^n Var\left(\sum_{t=1}^2 (x_{it} - \bar{x})\,u_{it}|\mathcal{X}_n\right)}{\left[\sum_{i=1}^n \sum_{t=1}^2 (x_{it} - \bar{x})^2\right]^2} = \frac{\sum_{i=1}^n \mathbb{E}\left[\left(\sum_{t=1}^2 (x_{it} - \bar{x})\,u_{it}\right)^2|\mathcal{X}_n\right]}{\left[\sum_{i=1}^n \sum_{t=1}^2 (x_{it} - \bar{x})^2\right]^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^n \sum_{t=1}^2 (x_{it} - \bar{x})^2}, \end{aligned}$$

where we have used that

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{t=1}^{2}(x_{it}-\bar{x})\,u_{it}\right)^{2}|\mathcal{X}_n\right] &= \sum_{t=1}^{2}(x_{it}-\bar{x})^{2}\,\mathbb{E}\left[u_{it}^{2}|\mathcal{X}_n\right]+2\,(x_{i1}-\bar{x})\,(x_{i2}-\bar{x})\,\mathbb{E}\left[u_{i1}u_{i2}|\mathcal{X}_n\right]\\ &= \sum_{t=1}^{2}(x_{it}-\bar{x})^{2}\,\mathbb{E}\left[u_{it}^{2}|x_{i1},x_{i2}\right]+2\,(x_{i1}-\bar{x})\,(x_{i2}-\bar{x})\,\mathbb{E}\left[u_{i1}u_{i2}|x_{i1},x_{i2}\right]\\ &= \sum_{t=1}^{2}(x_{it}-\bar{x})^{2}\,\sigma^{2}.\end{aligned}$$

(f) Maintain the assumptions stated in question (e). Derive an expression of the variance of the estimator in question (c) conditional on the regressors. Based on your derivations in questions (d)-(e) and here, which of the two estimators would you recommend using? Explain.

***Answer**: We have*

$$\begin{aligned}Var(\hat{\beta}_1|\mathcal{X}_n) &= \frac{\sum_{i=1}^{n}Var(\Delta x_{i2}\Delta u_{i2}|\mathcal{X}_n)}{\left[\sum_{i=1}^{n}(\Delta x_{i2})^{2}\right]^{2}}=\frac{\sum_{i=1}^{n}\mathbb{E}\left[(\Delta x_{i2})^{2}(\Delta u_{i2})^{2}|\mathcal{X}_n\right]}{\left[\sum_{i=1}^{n}(\Delta x_{i2})^{2}\right]^{2}}\\ &= \frac{\sum_{i=1}^{n}(\Delta x_{i2})^{2}\,\mathbb{E}\left[(\Delta u_{i2})^{2}|\mathcal{X}_n\right]}{\left[\sum_{i=1}^{n}(\Delta x_{i2})^{2}\right]^{2}}\\ &= \frac{2\sigma^{2}}{\sum_{i=1}^{n}(\Delta x_{i2})^{2}},\end{aligned}$$

where we have used that

$$\begin{aligned}\mathbb{E}\left[(\Delta u_{i2})^{2}|\mathcal{X}_n\right] &= \mathbb{E}\left[\Delta u_{i2}^{2}|x_{i1},x_{i2}\right]=\mathbb{E}\left[u_{i1}^{2}|x_{i1},x_{i2}\right]+\mathbb{E}\left[u_{i2}^{2}|x_{i1},x_{i2}\right]-2\mathbb{E}\left[u_{i1}u_{i2}|x_{i1},x_{i2}\right]\\ &= 2\sigma^{2}\end{aligned}$$

Both estimators are unbiased under the stated assumptions so we will prefer the estimator with the smaller variance. Comparing the variance expressions in (e) and (f) we see that $Var(\hat{\beta}_1|\mathcal{X}_n)\leq Var(\tilde{\beta}_1|\mathcal{X}_n)$ *if and only if*

$$2\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x})^{2}\geq\sum_{i=1}^{n}(\Delta x_{i2})^{2}.$$

But above inequality will always hold and so $\hat{\beta}_1$ is the preferred estimator: Writing

$$\begin{aligned}\sum_{i=1}^{n}(\Delta x_{i2})^2 &= \sum_{i=1}^{n}((x_{i2}-\bar{x})-(x_{i1}-\bar{x}))^2 \\ &= \sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x})^2 - 2\sum_{i=1}^{n}(x_{i2}-\bar{x})(x_{i1}-\bar{x}),\end{aligned}$$

substituting the final right-hand side expression into the inequality, and rearranging yields

$$\sum_{t=1}^{2}\sum_{i=1}^{n}(x_{it}-\bar{x})^2 + 2\sum_{i=1}^{n}(x_{i2}-\bar{x})(x_{i1}-\bar{x}) \geq 0.$$

Dividing through with n, we recognise the left-hand side as $\widehat{Var}(x_1)+\widehat{Var}(x_2)+2\widehat{Cov}(x_1,x_2) = \widehat{Var}(x_1+x_2) \geq 0$.

A.2 Using data from a survey of 456 male college students' own drinking habits together with their randomly assigned room mates' drinking habits, we obtain the following regression estimates,

$$\widehat{gpa} = \underset{(0.891)}{2.411} + \underset{(0.043)}{0.112}hsgpa + \underset{(0.082)}{0.442}sat - \underset{(0.150)}{0.109}frd - \underset{(0.119)}{0.028}ocd - \underset{(0.128)}{0.282}rfrd - \underset{(0.101)}{0.263}rocd,$$

where heteroskedasticity robust standard errors are report in parentheses and

- gpa – student's college grade point average (0–4).
- $hsgpa$ – student's high school grade point average (0–4).
- sat - student's SAT score divided by 100 (0–1.2).
- $frd = 1$ if student drank frequently in high shoool; 0, otherwise.
- $ocd = 1$ if student drank occasionally in high shoool; 0, otherwise.
- $rfrd = 1$ if student's room mate drank frequently in high shoool; 0, otherwise.
- $rocd = 1$ if student's room mate drank occasionally in high shoool; 0, otherwise.

(a) Based on above estimates, do a student's own drinking habits appear to have a significant effect on his performance? Do his peer's habits appear to have a significant effect? Explain.

***Answer**: The student's own past drinking habits are individually not statistically significant at a 10% level since*

$$|t_{frd}| = \frac{0.109}{0.150} = 0.727, \quad |t_{ocd}| = \frac{0.028}{0.119} = 0.236$$

which both are smaller than the 10% critical value $cv_{10\%} = 1.64$. In fact, the corresponding p-values are, with $Z \sim N(0,1)$,

$$\begin{aligned} p_{frd} &= \Pr(|Z| > 0.727) = 2\Pr(Z < -0.727) = 46.72\%, \\ p_{ocd} &= \Pr(|Z| > 0.236) = 2\Pr(Z < -0.236) = 81.34\%. \end{aligned}$$

So we find very little empirical support of an effect from own frequent or occasional drinking in the past. In contrast, the two indicators of room mate's past drinking behaviour are individually significant since

$$|t_{rfrd}| = \frac{0.282}{0.128} = 2.203, \quad |t_{rocd}| = \frac{0.263}{0.101} = 2.603$$

with the first being greater than the 5% critical value $cv_{5\%} = 1.96$ and the second greater than the 1% critical value $cv_{1\%} = 2.58$. The corresponding p-values are

$$\begin{aligned} p_{frd} &= \Pr(|Z| > 2.203) = 2\Pr(Z < -2.203) = 2.76\%, \\ p_{ocd} &= \Pr(|Z| > 2.603) = 2\Pr(Z < -2.603) = 0.92\%. \end{aligned}$$

Based on these individual tests, we conclude that the individual effects of frequent and moderate drinking, respectively, of the room mate are statistically significant. However, above does not allow us to conclude that the over-all effect of student's own past drinking is statistically insignificant since this would involve testing the joint hypothesis $\mathcal{H}_0 : \beta_{frd} = \beta_{ocd} = 0$. Similarly, we cannot make any statements above the over-all peer effect since this would involve testing $\mathcal{H}_0 : \beta_{rfrd} = \beta_{rocd} = 0$.

(b) Consider a student who drank frequently in high school. How much higher a high school SAT would the student have needed to achieve in order to offset the effect of his past drinking habits on college GPA?

***Answer**: To off-set the effect of frequent drinking in high school on his college GPA, the student's highschool SAT score had to grow by $\hat{\Delta}_{sat}$ solving*

$$0.442\hat{\Delta}_{sat} - 0.109 = 0 \Leftrightarrow \hat{\Delta}_{sat} = \frac{0.109}{0.442} = 0.247.$$

(c) Suppose that the estimators of the coefficients of $rfrd$ and $rocd$ are uncorrelated with each other. Test whether the expected effects on college GPA of the room mate drinking occasionally or frequently, respectively, in high school are identical at a 5% level. Conclude.

***Answer**: We are asked to test the hypothesis $\mathcal{H}_0 : \beta_{rfrd} = \beta_{rocd}$. This can be done using a t-test,*

$$t = \frac{\hat{\beta}_{rfrd} - \hat{\beta}_{rocd}}{\hat{\sigma}_{\hat{\beta}_{rfrd} - \hat{\beta}_{rocd}}},$$

where $\hat{\sigma}_{\hat{\beta}_{rfrd}-\hat{\beta}_{rocd}}$ *is a consistent estimator of* $std(\hat{\beta}_{rfrd}-\hat{\beta}_{rocd})$*. If indeed* $\hat{\beta}_{rfrd}$ *are uncorrelated* $\hat{\beta}_{rocd}$ *then* $std(\hat{\beta}_{rfrd}-\hat{\beta}_{rocd})=\sqrt{Var(\hat{\beta}_{rfrd})+Var(\hat{\beta}_{rocd})}$ *which is consistently estimated by*

$$\hat{\sigma}_{\hat{\beta}_{rfrd}-\hat{\beta}_{rocd}}=\sqrt{0.128^2+0.101^2}=0.163. \tag{5}$$

All together,

$$t=\frac{\hat{\beta}_{rfrd}-\hat{\beta}_{rocd}}{\hat{\sigma}_{\hat{\beta}_{rfrd}-\hat{\beta}_{rocd}}}=\frac{-0.282+0.263}{0.163}=-0.117$$

which is smaller than the 10% critical value. In fact, the test's p-value is $2\Pr(Z<-0.117)=90.72\%$ *and so we find find strong empirical support of the hypothesis. We conclude that it appears irrelevant whether the room mate was a heavy or a light drinker. Either type has the same peer effect.*

(d) Suppose that the estimators of the coefficients of $rfrd$ and $rocd$ are in fact positively correlated in the sample. Does this change the results and conclusions of your answer to question (c)? Explain.

Answer*: If in fact the estimated coefficients of* $rfrd$ *and* $rocd$ *are positively correlated then the standard errors in (??) will be bigger than the actual standard error of* $\hat{\beta}_{rfrd}-\hat{\beta}_{rocd}$*:*

$$std(\hat{\beta}_{rfrd}-\hat{\beta}_{rocd})=\sqrt{Var(\hat{\beta}_{rfrd})+Var(\hat{\beta}_{rocd})-2Cov(\hat{\beta}_{rfrd},\hat{\beta}_{rocd})}$$

This in turn implies

$$|t|=\frac{|\hat{\beta}_{rfrd}-\hat{\beta}_{rocd}|}{\hat{\sigma}_{\hat{\beta}_{rfrd}-\hat{\beta}_{rocd}}}<\frac{|\hat{\beta}_{rfrd}-\hat{\beta}_{rocd}|}{\tilde{\sigma}_{\hat{\beta}_{rfrd}-\hat{\beta}_{rocd}}}=|t^*|$$

where $\tilde{\sigma}_{\hat{\beta}_{rfrd}-\hat{\beta}_{rocd}}$ *is the correct standard error taking into account the positive correlation and* t^* *being the corresponding t statistic. If* $Cov(\hat{\beta}_{rfrd},\hat{\beta}_{rocd})$ *is big enough then* t^* *may exceed the critical value, but we don't know. Thus, we cannot make any draw any firm conclusions under this scenario.*

(e) You conjecture that if a student drank in high school then the effect of his room mate being a drinker will be stronger (more negative) compared to the case where the student did not drink in high school in the first place. Explain how you would modify above regression to empirically examine whether this conjecture is true.

Answer*: We would augment the regression with interaction terms of the form* $frd\times rfrd$*,* $frd\times rocd$*,* $ocd\times rfrd$ *and* $ocd\times rfrd$*. The coefficients of these additional regressors will capture above conjectured effects.*

(f) Innate ability is an often cited factor driving student performance. How is this factor controlled for in above regression? Discuss potential remaining biases in the OLS estimates

reported above due to innate ability not being directly observed and included as a covariate in the estimated regression model.

***Answer**: hsgpa and sat can be seen as proxies for innate ability. If these are valid proxies in the sense of Wooldridge, Section 9.2 then the OLS estimates of the coefficients of frd, ocd, rfrd and rocd are unbiased and consistent. Moreover, the OLS estimates of the coefficients of hsgpa and sat can be interpreted as the ceteris paribus effects on college GPA of these two variables. However, it seems unlikely that hsgpa and sat are valid proxies. Specifically, once we control for sat and hsgpa, the drinking dummies should not covary with ability. This does not seem plausible. As such above the OLS estimates reported above are probably biased. But we expect these biases to be smaller than if we had not included hsgpa and sat.*

PART B

Answer ONE question from this section.

B.1 In "For Want of a Cup: The Rise of Tea in England and the Impact of Water Quality on Mortality" (Review of Economics and Statistics, 2023), Francisca Antman studies the impact of water quality on mortality in 18th century England. To do so, the author examines the unintentional dissemination of boiled water consumption as tea drinking became widespread in England after the Tea and Windows Act of 1784 reduced the tea tax from 119 to 12.5 percent at one stroke. In this question, we will not focus on this particular policy but use the setting to examine several topics studied in class. Let $\texttt{Deaths}_{it}$ be the (log) number of deaths in parish i in year t and where water quality is measured in the study according to the number of water sources within 3 km of the parish or its elevation. Elevation is believed to be positively correlated with water quality since parishes at higher elevation would have been less likely to be subjected to water contamination from surrounding areas. (The article also controls for various parish specific variables, but we will abstract from those here for simplicity.) For the items below, we consider one particular year and thus abstracting from the subscript t for simplicity.

(a) To gauge the causal relationship between the availability of good quality water on mortality, one could focus on the following regression model:

$$\texttt{Deaths}_i = \alpha_0 + \alpha_1 WQ_i + \epsilon_i, \tag{6}$$

where WQ_i is a (not necessarily binary) measure of water quality in parish i and ϵ_i are additional determinants for $\texttt{Deaths}_i$ that are unobserved by the researcher. An important determinant of mortality which is omitted from the regression above is the average wealth in the parish. If wealth influences mortality and is also related to water quality, how would this affect the properties of the conventional OLS estimator for the regression above?
ANSWER: *This would lead the error term to be correlated with WQ_i thus compromising econometric exogeneity.*

(b) As noted above, there are actually two variables that proxy for water quality in the study: the number of water sources within 3km and its elevation.

i. A friend suggests using the latter as an instrumental variable for the former in estimating the regression in (**??**). Describe in detail the estimation procedure to implement this suggestion within the context of this exercise.
ANSWER: *Describe TSLS.*

ii. Do you think elevation would be a reasonable instrumental variable for the number of water sources near the parish? Explain your answer. (Remember that an instrumental

variable should be both relevant and valid.)
ANSWER: *The two variables are likely to be related so the IV would be relevant. They are both likely to be related to average parish wealth so the IV would not be valid.*

iii. Explain mathematically why it would not be possible to test whether the instrumental variable suggested is valid.
ANSWER: *See Problem Set.*

(c) As noted in the article, "[p]revailing views on the causes of mortality crises focused on miasmas, clouds of noxious gases that moved indiscriminately across the population spreading illness and death. It was not until the 1840s that William Budd and John Snow argued that typhoid and cholera were spread through contaminated water." To investigate the issue, Snow [1855] discovered that during the 1853-1854 London cholera epidemic, households were supplied with water by two independent suppliers: the Lambeth water company, which in 1849 had moved its water intake to a point in the Thames above the main sewage discharge (thus less amenable to contamination), and the Southwark and Vauxhall company, whose intake remained below the discharge. (thus more amenable to contamination). Consider then the following (very!) stylised representation for how WQ_i is affected by (say, cholera or typhoid) deaths in the mid-19th century:

$$WQ_i = \gamma_0 + \gamma_1 \texttt{Deaths}_i + \gamma_2 \texttt{Supplier}_i + \eta_i, \tag{7}$$

where $\texttt{Supplier}_i$ is the proportion of houses supplied by Lambeth water company in parish i. The variable η_i represents additional factors contributing to water quality above an beyond the number of deaths and the water supplier for parish i.

i. Snow collected data on the addresses of cholera victims and found that there were 8.5 times as many deaths per thousand among households supplied by the Southward and Vauxhall company than those supplied by the Lambeth company! To emulate his exercise, suppose you run a regression of $\texttt{Deaths}_i$ on $\texttt{Supplier}_i$:

$$\texttt{Deaths}_i = \varphi_0 + \varphi_1 \texttt{Supplier}_i + \omega_i. \tag{8}$$

Express the estimands (i.e., φ_0 and φ_1) for this regression in terms of the parameters in equations (**??**) and (**??**).
ANSWER: *Plugging in equation (**??**) into equation (**??**) one obtains:*

$$\texttt{Deaths}_i = \frac{\alpha_0 + \alpha_1\gamma_0}{1 - \alpha_1\gamma_1} + \frac{\alpha_1\gamma_2}{1 - \alpha_1\gamma_1}\texttt{Supplier}_i + error$$

ii. Would you be able to estimate the parameters in equation (**??**)? Explain your answer and provide any additional conditions you might need.
ANSWER: *Yes: the equation is exactly identified and one can estimate the parameters if the errors are unrelated to* $\texttt{Supplier}_i$.

iii. Would you be able to estimate all the parameters in equation (**??**)? Explain your answer and provide any additional conditions you might need.
ANSWER: *No: the equation is under-identified and one would not be able to estimate the parameters even if the errors are unrelated to* $\texttt{Supplier}_i$.

(d) Suppose that historical death records are unavailable for certain parishes. Instead assume that whether historical records are available for parish i ($S_i = 1$) or not ($S_i = 0$) can be modelled as:

$$S_i = \mathbf{1}(\pi_0 + \pi_1 \texttt{Wealth}_i + \xi_i \geq 0),$$

where $\texttt{Wealth}_i$ is a variable measuring how wealthy a parish is and ξ_i follows a standard normal distribution independent of other observable variables.

i. Under what conditions would an OLS estimator for equation (**??**) using the available data be consistent?
ANSWER: *Independence between* ω_i *and* ξ_i.

ii. If those conditions do not necessarily hold, how would you estimate equation (**??**) instead? Explain your answer.
ANSWER: *Estimate a Heckman selection model either in two stages or via MLE.*

iii. How could you test whether the conditions for consistency of OLS hold above? Explain your answer.
ANSWER: *Test whether the coefficient on the inverse Mills' ratio is different from zero.*

(e) Suppose the variable D_i is equal to 1 if parish i is marked as having access to good quality water and is equal to 0, otherwise. One way to investigate the relationship between water quality and mortality is to estimate the following model:

$$D_i = \mathbf{1}(\beta_0 + \beta_1 \texttt{Deaths}_i + \nu_i \geq 0).$$

where ν_i follows a standard logistic distribution.

i. Write down the log-likelihood function for this estimation problem when you have N observations (i.e., parishes).

ANSWER: *The log-likelihood function is:*

$$\sum_{j=1}^{n}\{-b_0-b_1\texttt{Deaths}_i-\ln[1+\exp(-b_0-b_1\texttt{Deaths}_i)]\}\times(1-D_i)-\ln[1+\exp(-b_0-b_1\texttt{Deaths}_i)]\times D_i$$

or equivalently

$$\sum_{j=1}^{n}\{b_0+b_1\texttt{Deaths}_i-\ln[1+\exp(b_0+b_1\texttt{Deaths}_i)]\}\times D_i-\ln[1+\exp(b_0+b_1\texttt{Deaths}_i)]\times(1-D_i).$$

ii. Given estimates for β_0 and β_1 how would you estimate the Average Partial Effect (APE) of `Deaths` on the probability of a parish having good quality water? How would you interpret this estimate? Please explain your answer.
ANSWER: *Once estimates are obtained, the APE can be estimated as*

$$\sum_{i=1}^{N}\frac{1}{N}b_1\frac{\exp(b_0+b_1\texttt{Deaths}_i)}{[1-\exp(b_0+b_1\texttt{Deaths}_i)]^2}.$$

The APE corresponds to the estimate of the average increment in the probability of parish i having good quality water for a small change in the (log) number of deaths. Notice that this is not a causal statement, but a correlational one.

iii. Suggest one measure of goodness-of-fit for the model above. Please explain your answer.
ANSWER: *Explain Efron or McFadden psedo-R^2.*

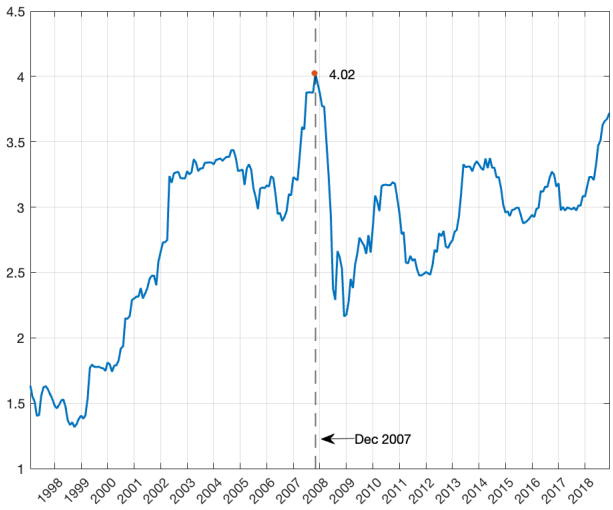
B.2 This question concerns the problem of forecasting inflation, using either autoregressive models or autoregressive distributed lag models with UK macroeconomic data. In particular, we will aim to forecast inflation for October 2023. The variables are defined in Table **??** below, and Tables **??** and **??** present regression results. Figure **??** will help with part (d). For the purposes of this question, **do not worry about the fact that inflation may display a unit root**, or possible non-stationarity as a result.

Table 1: Variable Definitions

Unit of Observation: Monthly, April 1992 – September 2023 ($T = 378$)

Variable name	Variable definition	Mean	Std. Dev.
Inf	Monthly percentage change in the UK Consumer Price Index (CPI), annualised	2.43	1.60
U	UK Unemployment rate (16 and over, seasonally adjusted)	6.16	1.84
IP	Monthly percentage change in the UK industrial production index, seasonally adjusted annualised	2.64	20.79
Pet	Monthly percentage change in Petrol and Oil Retail Price Index, annualised	8.97	33.81

Figure 1: Chow statistic for regression (VIII) for various dates



(a) Suppose you are asked to select one of models (I)-(V) in Table **??** to make your forecast. Which would you choose? What criterion do you use to make this decision?
The BIC is the lowest for model (IV) – therefore it is the preferred model.

Table 2: Autoregressive inflation forecasting models, monthly data, 1992-2023

	(I)	(II)	(III)	(IV)	(V)
Sample	9/92-9/23	9/92-9/23	9/92-9/23	9/92-9/23	9/92-9/23
Const	0.04	0.04	0.05	0.06*	0.07**
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Inf_{t-1}	0.99***	1.21***	1.18***	1.15***	1.13***
	(0.02)	(0.07)	(0.07)	(0.07)	(0.07)
Inf_{t-2}		-0.22***	-0.07	-0.09	-0.08
		(0.07)	(0.10)	(0.09)	(0.10)
Inf_{t-3}			-0.13*	0.15	0.14
			(0.06)	(0.12)	(0.12)
Inf_{t-4}				-0.23***	-0.09
				(0.08)	(0.10)
Inf_{t-5}					-0.12*
					(0.07)
SEs	HR	HR	HR	HR	HR
BIC	-2.52	-2.55	-2.55	-2.58	-2.57
R^2	0.97	0.97	0.97	0.97	0.97
Adj R^2	0.97	0.97	0.97	0.97	0.97
SER	0.28	0.27	0.27	0.26	0.26
N obs	374	374	374	374	374

Notes: All regressions are estimated by OLS. The type of standard errors (in parentheses below coefficients) is indicated: HR is heteroskedasticity robust, HAR is heteroskedasticity and autocorrelation robust. Statistical significance is indicated at the 10 (∗), 5 (∗∗) and 1% (∗ ∗ ∗) levels.

Table 3: Inflation forecasting models, 1992-2023

	(VI)	(VII)	(VIII)
Sample	8/92-12/15	5/92-12/15	6/92-12/15
Const	0.10**	0.10*	0.10*
	(0.04)	(0.05)	(0.05)
Inf_{t-1}	1.14***	0.96***	1.14***
	(0.07)	(0.02)	(0.07)
Inf_{t-2}	-0.14	0.00	-0.21***
	(0.09)	(0.00)	(0.07)
Inf_{t-3}	-0.01		
	(-0.11)		
Inf_{t-4}	-0.04		
	(-0.07)		
U_{t-1}		-0.00	0.16
		(0.01)	(0.15)
U_{t-2}			-0.15
			(0.15)
IP_{t-1}		-0.00	-0.00
		(0.00)	(0.00)
IP_{t-2}			-0.00
			(0.00)
Pet_{t-1}		0.00	0.00
		(0.00)	(0.00)
Pet_{t-2}			-0.00
			(0.00)
SEs	HR	HR	HR
BIC	-2.95	-2.98	-3.00
R^2	0.93	0.94	0.94
Adj R^2	0.93	0.94	0.94
SER	0.22	0.22	0.22
N obs	281	284	283
POOS RMSFE	0.14	0.15	0.15
POOS Sample	1/16-1/20	1/16-1/20	1/16-1/20

Notes: All regressions are estimated by OLS. The type of standard errors (in parentheses below coefficients) is indicated: HR is heteroskedasticity robust, HAR is heteroskedasticity and autocorrelation robust. Statistical significance is indicated at the 10 ($*$), 5 ($**$) and 1% ($***$) levels.

(b) Other variables, besides lags of inflation, may be helpful in forecasting inflation in October 2023. Regressions (VI)-(VIII) in Table **??** compare three forecasting models that include different regressors.

i. Which model would you choose to make your forecast? What criterion do you use to make this decision?
For forecasting recent data out-of-sample, regression (VI) is preferred since it has the lowest POOS RMSFE. This is our best assessment of forecasting performance on unseen data. In this case, additional regressors do not seem to help.

ii. Using the data below, compute a forecast for October 2023 using regression (IX).

	Inflation	Unemployment	Industrial production	Petrol and Oil inflation
9/23	6.30	4.20	1.27	72.87
8/23	6.30	4.20	-6.10	67.16

$0.10 + 1.14 \times 6.30 - 0.21 \times 6.30 + 0.16 \times 4.20 - 0.15 \times 4.20 - 0.00 \times 1.27 - 0.00 \times -6.10 + 0.00 \times 72.87 - 0.00 \times 67.16 = 6.00.$

iii. Compute a 95% forecast interval for your prediction, under the additional assumptions that the errors are normal and that the change in inflation is small.
With these assumptions, the approximate 95% forecast interval is approximated using the POOS RMSFE, $6.30 \pm 1.96 \times 0.15 = [5.71, 6.29]$.

(c) Figure **??** plots the Chow statistic for regression (XIII) for the central 70% of the sample.

i. What null hypothesis does the Chow statistic test? Describe the test.
The Chow test tests the null hypothesis of parameter stability. In particular, for a regression where each regressor is interacted with a dummy variable equal to zero before and one after a hypothetical break date, it is a joint F-test of the null hypothesis that all interaction coefficients are zero.

ii. The maximum value for the Chow statistic occurs in December 2007, with a value of 4.02. What is another name for this statistic? To what critical value must this statistic be compared for a 5% test?
The maximum value of the Chow statistic over some portion of the sample is called the QLR statistic. The critical values for this statistic are found in Table **??**. *There are 9 parameters in regression (VIII), so there are 9 degrees of freedom, and trimming is 15%, so the 5% critical value is 2.84.*

iii. What do you conclude about the stability of model parameters based on Figure **??** (at the 5% significance level)?
We reject the null hypothesis of parameter stability at the 5% level.

(d) Suppose you wanted to identify the effect of (lagged) industrial production (IP_{t-1}) on inflation. For this question, you can assume that the linear models considered are cor-

rectly specified, there is no collinearity, and the variables are jointly stationary and weakly dependent.

i. The coefficient on IP_{t-1} appears to be zero and statistically insignificant in the forecasting regressions (VII) and (VIII). Is this evidence that the causal effect of (lagged) industrial production on inflation is zero?
No; these are forecasting regressions and we have not argued that any exogeneity condition is satisfied, so we cannot give the coefficients causal interpretations.

ii. What additional assumption is required in order to interpret this coefficient as an unbiased estimate of the causal effect? What does this assumption state?
The strict exogeneity assumption is needed; it requires that the error, e_t, is conditionally mean zero given all past, present, and future values of IP_t.

iii. What additional assumption is required in order to interpret this coefficient as a consistent estimate of the causal effect? What does this assumption state?
The contemporaneous exogeneity assumption is needed; it requires that the error, e_t, is conditionally mean zero given IP_{t-1}.

(e) The FRED-MD database is a monthly database consisting of 127 macroeconomic timeseries. Suppose you had a similar database available for the UK of the sample considered above (1992-2023).

i. You add the one-period lagged values of each regressor to the regression in (VIII), and notice that the out-of-sample forecasting performance of the model gets worse, not better. What problem associated with many possibly irrelevant/redundant regressors could explain this?
The number of observations is very small relative to the number of variables. A regression including all of them would suffer from overfitting and imprecisely estimated coefficients.

ii. However, the in-sample performance (i.e., SER, R^2) is excellent. Is this surprising? Why?
No. This problem will not be apparent in-sample, since adding additional regressors can only improve the model fit, without suitable adjustment/penalisation.

iii. Describe one possible way to overcome the problem you identified in (i).
Possible answers are Principal Components, Ridge, and LASSO, with accurate descriptions. Principal Components "shrinks" the number of predictors down to a much smaller number of common components that explain as much of the variation in the original variables as possible. Ridge regression adds a penalty term to OLS to shrink the coefficients towards zero, accounting for the fact that in small samples many would receive non-zero coefficients due to random error. Lasso adds a penalty term to OLS to shrink large coefficients towards zero and small coefficients to exactly zero, dropping them from the regression entirely.

Table 4: Critical values of the QLR statistic with 15% trimming

Number of restrictions (q)	10%	5%	1%
1	7.12	8.68	12.16
2	5.00	5.86	7.78
3	4.09	4.71	6.02
4	3.59	4.09	5.12
5	3.26	3.66	4.53
6	3.02	3.37	4.12
7	2.84	3.15	3.82
8	2.69	2.98	3.57
9	2.58	2.84	3.38
10	2.48	2.71	3.23

Table 5: 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

Table 6: 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