ECON0019 Past Paper 2024

Question A1

1. We minimize the sum of squared residual for both parameters

$$\frac{\partial SSR}{\partial b_0} = -2\sum_{i=1}^{n}\sum_{t=1}^{2}(y_{it} - b_0 - b_1 x_{it}) = 0$$

$$\frac{\partial SSR}{\partial b_1} = -2\sum_{i=1}^{n}\sum_{t=1}^{2}(y_{it} - b_0 - b_1 x_{it})x_{it} = 0$$

For the OLS intercept parameter:

$$\sum_{i=1}^{n}\sum_{t=1}^{2} y_{it} - 2n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n}\sum_{t=1}^{2} x_{it} = 0$$

Let $N = 2n$,

$$\hat{\beta}_0 = \underbrace{\frac{1}{N}\sum_{i=1}^{n}\sum_{t=1}^{2} y_{it}}_{\bar{y}} - \hat{\beta}_1 \underbrace{\frac{1}{N}\sum_{i=1}^{n}\sum_{t=1}^{2} x_{it}}_{\bar{x}}$$

For the OLS slope parameter:

$$\sum_{i=1}^{n}\sum_{t=1}^{2} x_{it}y_{it} - \hat{\beta}_0 \sum_{i=1}^{n}\sum_{t=1}^{2} x_{it} - \hat{\beta}_1 \sum_{i=1}^{n}\sum_{t=1}^{2} x_{it}^2 = 0$$

$$\sum_{i=1}^{n}\sum_{t=1}^{2} x_{it}y_{it} - \left(\bar{y} - \hat{\beta}_1\bar{x}\right)N\bar{x} - \hat{\beta}_1 \sum_{i=1}^{n}\sum_{t=1}^{2} x_{it}^2 = 0$$

$$\sum_{i=1}^{n}\sum_{t=1}^{2} x_{it}y_{it} - N\bar{x}\bar{y} + N\bar{x}^2\hat{\beta}_1 - \hat{\beta}_1 \sum_{i=1}^{n}\sum_{t=1}^{2} x_{it}^2 = 0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{2} x_{it}y_{it} - N\bar{x}\bar{y}}{\sum_{i=1}^{n}\sum_{t=1}^{2} x_{it}^2 - 2n\bar{x}^2}$$

$$= \frac{\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it} - \bar{x})(y_{it} - \bar{y})}{\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it} - \bar{x})^2}$$

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

By the law of large number (LLN), when sample size and time period approach infinity

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x})u_{it} \xrightarrow{p} \mathbb{E}[(x_{it}-\mu_x)u_{it}] = \underbrace{\mathbb{E}[x_{it}u_{it}]}_{=0} - \mu_x \underbrace{\mathbb{E}[u_{it}]}_{=0} = 0$$

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x})^2 \xrightarrow{p} \mathbb{E}[(x_{it}-\mu_x)^2] = \mathrm{Var}(x_{it}) > 0$$

Given that the numerator of the sampling error is naught while we assume the variance of independent variable (the denominator) is non-zero, this OLS slope estimator is consistent $\hat{\beta}_1 \xrightarrow{p} \beta_1$.

2. Using similar method of minimizing the object function, the first difference estimator will take the form of the OLS estimator without an intercept

$$\hat{\beta}_1 = \frac{\sum_i(\Delta y_{i2})(\Delta x_{i2})}{\sum_i(\Delta x_{i2})^2} = \beta_1 + \frac{\sum_i(\Delta x_{i2})(\Delta u_{i2})}{\sum_i(\Delta x_{i2})^2}$$

By the law of large number (LLN),

$$\frac{1}{n}\sum_{i}(\Delta x_{i2})(\Delta u_{i2}) \xrightarrow{p} \mathbb{E}[(\Delta x_2)(\Delta u_2)] = 0$$

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

By similar argument, we have $\hat{\beta}_1 \xrightarrow{p} \beta_1$

3. We expand the covariance term if $s \neq t$

$$\mathrm{Cov}(x_{i1}, u_{i2}) = \mathbb{E}[x_{i1}u_{i2}] - \mathbb{E}[x_{i1}]\underbrace{\mathbb{E}[u_{i2}]}_{=0} = \mathbb{E}[x_{i2}u_{i2}] > 0$$

$$\mathrm{Cov}(x_{i2}, u_{i1}) = \mathbb{E}[x_{i2}u_{i1}] > 0$$

Then we first expand equation (3)

$$\begin{aligned}\mathbb{E}[\Delta u_{i2}\Delta x_{i2}] &= \mathbb{E}[(x_{i2}-x_{i1})(u_{i2}-u_{i1})] \\ &= \underbrace{\mathbb{E}[x_{i2}u_{i2}]}_{=0} - \underbrace{\mathbb{E}[x_{i2}u_{i1}]}_{>0} - \underbrace{\mathbb{E}[x_{i1}u_{i2}]}_{>0} + \underbrace{\mathbb{E}[x_{i1}u_{i1}]}_{=0} \\ &= -(\mathbb{E}[x_{i2}u_{i1}] + \mathbb{E}[x_{i1}u_{i2}]) < 0\end{aligned}$$

We see that if only contemporaneous exogeneity holds, which does not guarantee cross-period orthogonality, equation (3) no longer holds and first difference estimator will no longer be consistent.

When $u_{it} = \alpha_i + v_{it}$ and v_t is strictly exogeneous of x_t, Equation (3) holds

$$\begin{aligned}\mathbb{E}[\Delta u_{i2}\Delta x_{i2}] &= \mathbb{E}[(x_{i2} - x_{i1})(u_{i2} - u_{i1})] \\ &= \mathbb{E}[x_{i2}\alpha_i] - \mathbb{E}[x_{i2}\alpha_i] - \mathbb{E}[x_{i1}\alpha_i] + \mathbb{E}[x_{i1}\alpha_i] \\ &= 0\end{aligned}$$

However, Equation (2) does not hold. For example, $t = 1$

$$\mathbb{E}[u_{i1}x_{i1}] = \mathbb{E}[\alpha_i x_{i1} + v_{i1}x_{i1}] = \mathbb{E}[\alpha_i x_{i1}] \neq 0$$

4. The sampling error form of pooled regression estimator in (a) was

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

By driving its conditional mean, we see that estimator in (a) is unbiased.

$$\begin{aligned}\mathbb{E}\left[\hat{\beta}_1 \middle| x_{i1}, x_{i2}\right] &= \beta_1 + \mathbb{E}\left[\frac{\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it} - \bar{x})u_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it} - \bar{x})^2} \middle| x_{i1}, x_{i2}\right] \\ &= \beta_1 + \frac{1}{\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it} - \bar{x})^2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it} - \bar{x})u_{it} \middle| x_{i1}, x_{i2}\right] \\ &= \beta_1 + \frac{1}{\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it} - \bar{x})^2}\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it} - \bar{x})\underbrace{\mathbb{E}[u_{it}|x_{i1}, x_{i2}]}_{=0} \\ &= \beta_1 \blacksquare\end{aligned}$$

The sampling error form of the first differencing estimator in (c) was

$$\hat{\beta}_1 = \frac{\sum_i(\Delta y_{i2})(\Delta x_{i2})}{\sum_i(\Delta x_{i2})^2} = \beta_1 + \frac{\sum_i(\Delta x_{i2})(\Delta u_{i2})}{\sum_i(\Delta x_{i2})^2}$$

Using similar derivation method, we can also show that the first differencing estimator is unbiased as well

$$\begin{aligned}\mathbb{E}\left[\hat{\beta}_1 \middle| x_{i1}, x_{i2}\right] &= \beta_1 + \mathbb{E}\left[\frac{\sum_i(\Delta x_{i2})(\Delta u_{i2})}{\sum_i(\Delta x_{i2})^2} \middle| x_{i1}, x_{i2}\right] \\ &= \beta_1 + \frac{1}{\sum_i(\Delta x_{i2})^2}\sum_i(\Delta x_{i2})\mathbb{E}[\Delta u_{i2}|x_{i1}, x_{i2}]\end{aligned}$$

$$\because \mathbb{E}[\Delta u_{i2}|x_{i1}, x_{i2}] = \mathbb{E}[u_{i2}|x_{i1}, x_{i2}] - \mathbb{E}[u_{i1}|x_{i1}, x_{i2}] = 0$$

$$\therefore \left[\hat{\beta}_1 \middle| x_{i1}, x_{i2}\right] = \beta_1 \blacksquare$$

5. Using homoskedasticity and strict exogeneity assumption, the conditional variance is

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1|\chi_n) &= \mathrm{Var}\left(\beta_1 + \frac{\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x})u_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x})^2}\middle|\chi_n\right) \\
&= \frac{1}{[\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x})^2]^2}\mathrm{Var}(\textstyle\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x})u_{it}\,|\chi_n) \\
&= \frac{1}{[\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x})^2]^2}\sum_{i=1}^{N}\sum_{t=1}^{2}\mathrm{Var}\big((x_{it}-\bar{x})u_{it}|\chi_n\big) \\
&= \frac{1}{[\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x})^2]^2}\sum_{i=1}^{N}\sum_{t=1}^{2}(x_{it}-\bar{x})^2\underbrace{\mathrm{Var}(u_{it}|\chi_n)}_{\sigma^2} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{2}(x_{it}-\bar{x})^2}{[\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x})^2]^2}\sigma^2 \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x})^2}\blacksquare
\end{aligned}$$

6. Using similar derivation steps

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_{FD}|\chi_n) &= \mathrm{Var}\left(\beta_1 + \frac{\sum_i(\Delta x_{i2})(\Delta u_{i2})}{\sum_i(\Delta x_{i2})^2}\middle|\chi_n\right) \\
&= \frac{1}{[\sum_i(\Delta x_{i2})^2]^2}\mathrm{Var}(\textstyle\sum_i(\Delta x_{i2})(\Delta u_{i2})\,|\chi_n) \\
&= \frac{1}{[\sum_i(\Delta x_{i2})^2]^2}\sum_i(\Delta x_{i2})^2\mathrm{Var}(\Delta u_{i2}|\chi_n)
\end{aligned}$$

Given that $\Delta u_{i2} = u_{i2} - u_{i1}$, then

$$\begin{aligned}
\mathrm{Var}(\Delta u_{i2}|\chi_n) &= \mathrm{Var}(u_{i2}-u_{i1}|\chi_n) \\
&= \underbrace{\mathrm{Var}(u_{i2}|\chi_n)}_{\sigma^2} + \underbrace{\mathrm{Var}(u_{i1}|\chi_n)}_{\sigma^2} - 2\underbrace{\mathrm{Cov}(u_{i2},u_{i1})}_{0} \\
&= 2\sigma^2
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_{FD}|\chi_n) &= \frac{\sum_i(\Delta x_{i2})^2}{[\sum_i(\Delta x_{i2})^2]^2}\cdot 2\sigma^2 \\
&= \frac{2\sigma^2}{\sum_i(\Delta x_{i2})^2}
\end{aligned}$$

We have concluded that both estimators are unbiased and consistent under strict exogeneity. However, the pooled OLS estimator is more efficient under homoskedasticity, so we prefer the pooled OLS estimator. However, given that the fixed effect estimator is

more robust to time-invariant biases at the cost of higher sampling variance, it is preferred to pooled OLS estimator in real world setting where you cannot carry out a perfectly clean experiment.

Question A2

1. We calculate the absolute *t*-statistics of β_{frd} and β_{ocd} with null hypothesis $\mathcal{H}_0: \beta_{frd} = 0$ and $\mathcal{H}_0: \beta_{ocd} = 0$. Since sample size is well above 30, we may use the critical value of the standard normal distribution directly.

$$|t_{frd}| = 0.727, \qquad |t_{ocd}| = 0.235$$

Note that both null hypotheses are failed to be rejected under 10% level of significance, meaning that the drinking habit of the student himself has no significant effect on his college GPA on an individual level.

Additionally, we test $\mathcal{H}_0: \beta_{rfrd} = 0$ and $\mathcal{H}_0: \beta_{rocd} = 0$

$$|t_{rfrd}| = 2.203, \qquad |t_{rocd}| = 2.604$$

Note that both null hypotheses are rejected at the 5% level of confidence while *rocd* is rejected at the 1% level of significance, meaning that roommate's drinking habit has a statistically significant impact on the student's performance in college.

Note that we cannot test for joint significance due to lack of information regarding the sum of squared residuals of the restricted model and the full model.

2. In order to offset the negative impact of frequent drinking, the student need $\Delta SAT \times 100$ more points in SAT

$$0.442 \cdot \Delta SAT = 0.109$$
$$\Delta SAT = 0.2466$$
$$\Delta SAT \times 100 = 246.6$$

3. We wish to test the following null hypothesis

$$\mathcal{H}_0: \beta_{rfrd} = \beta_{rocd}$$

The student's *t*-statistics is

$$t = \left|\frac{\hat{\beta}_{rfrd} - \hat{\beta}_{rocd}}{se(\hat{\beta}_{rfrd} - \hat{\beta}_{rocd})}\right| = \left|\frac{\hat{\beta}_{rfrd} - \hat{\beta}_{rocd}}{\sqrt{\text{Var}(\hat{\beta}_{rfrd}) + \text{Var}(\hat{\beta}_{rocd}) - 2\text{Cov}(\hat{\beta}_{rfrd}, \hat{\beta}_{rocd})}}\right|$$

$$t = \left|\frac{\hat{\beta}_{rfrd} - \hat{\beta}_{rocd}}{\sqrt{\text{Var}(\hat{\beta}_{rfrd}) + \text{Var}(\hat{\beta}_{rocd})}}\right| = \left|\frac{-0.282 + 0.263}{\sqrt{0.128^2 + 0.101^2}}\right| = 0.1165$$

Therefore, we failed to reject the null hypothesis at any conventional level of significance and we may conclude that the effect on GPA is constant regardless whether the roommate drinks frequently or occasionally.

4. If $\hat{\beta}_{rfrd}$ and $\hat{\beta}_{rocd}$ are positively correlated, then variance $\text{Var}(\hat{\beta}_{rfrd}) + \text{Var}(\hat{\beta}_{rocd}) - 2\text{Cov}(\hat{\beta}_{rfrd}, \hat{\beta}_{rocd})$ is strictly smaller than the zero-covariance approximation, so the denominator falls and *t*-statistics must increase. However, if the covariance term is large enough, previous conclusion may be overruled as the test statistics approaches the critical values.

5. We may modify this regression with interaction terms between the student's drinking habit and his roommate's drinking habit

$$\widehat{gpa} = \gamma_0 + \gamma_1 hsgpa + \gamma_2 sat + \gamma_3 frd + \gamma_4 ocd + \gamma_5 rfrd + \gamma_6 rocd + \gamma_7 frd \times rfrd + \gamma_8 frd \times rocd + \gamma_9 ocd \times rfrd + \gamma_{10} ocd \times rocd + u$$

Coefficients $\gamma_7, \dots, \gamma_{10}$ will capture the conjectural effect mentioned in the question.

6. As innate capability constitutes an omitted variable, it is proxied in the model using *hsgpa* and *sat*. Both are intuitively reasonable proxies but they cannot perfectly predict the effect of innate capability. Therefore, the coefficients of regressors of interest may still be biased, but the magnitude of bias shall be smaller than if we do not control for *sat* and *hsgpa*.

Question B.1

1. [OLS Assumptions & OVB] It's reasonable and intuitive to postulate that wealth is negatively correlated with mortality while positively correlated with water quality, omitted this relevant variable contributed to omitted variable bias (OVB) and undermine the exogeneity assumption (SLR.4)

2. [2SLS Procedure] When using instrumental variable, we shall adopt the two-stage least-square estimation protocol (2SLS). First, we regress the number of water source on elevation and an intercept, estimate the fitted value of the first-stage outcome variable. Then, we regress the outcome variable (i.e. mortality) on the fitted value of number of water source. The OLS estimate of the second-stage regression is the 2SLS estimator.

   [IV Assumptions] Elevation may be a relevant instrument for number of water sources because water tend to originate from upstream of river and other streams. However, elevation may not be a exogenous & valid instrument because elevation may correlate with average wealth in the parish, which is omitted and a part of the error term.

   [Over identification test] Exogeneity of instrument is not testable if the endogenous regressor is just identified. We have the following procedure of a Sargan-Hansen Overidentification Test:

   1) Record the residual

      $$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

      Where $\hat{\beta}_0$ and $\hat{\beta}_1$ are TSLS estimators.

   2) Regress $\hat{u}_i$ on all exogenous variables. When there are one endogenous variable instrumented by one exogenous variables, the regression is

      $$\hat{u}_i = \gamma_0 + \gamma_1 z_1 + v$$

   3) Under the null hypothesis that all instruments are valid (uncorrelated with the error term), the test statistics is

      $$nR^2 \sim \chi^2_{M-1}$$

When the endogenous regressor is just identified. For this SLR model in Step (2), the R-squared is defined is identical to the square of the Pearson correlation coefficient between the regressor and the outcome variable.

$$R^2 = \frac{[\text{Cov}(z, \hat{u}_1)]^2}{\sigma_z^2 \sigma_{\hat{u}_1}^2}$$

Given that under the null hypothesis $\text{Cov}(z, \hat{u}_1) = 0$ for this just identified model, the test statistics is always naught, and this validity test cannot be carried out.

3. [SEM Reduced Form] i) Substitute (7) into (6) and rearrange

$$\text{Deaths}_i = \underbrace{\frac{\alpha_0 + \alpha_1\gamma_0}{1 - \alpha_1\gamma_1}}_{\varphi_0} + \underbrace{\frac{\alpha_1\gamma_2}{1 - \alpha_1\gamma_1}}_{\varphi_1}\text{Supplier}_i + \underbrace{\frac{\alpha_1\eta_i + \epsilon_i}{1 - \alpha_1\gamma_1}}_{\omega_i}$$

[Order condition] ii) We assume supplier is exogenous and uncorrelated with error terms. Notice that Equation (6) and Equation (7) constitutes a simultaneous equation model (SEM), where Equation (6) has one endogenous variable (WQ) and one excluded exogenous regressor (supplier), so that Equation (6) is just identified and could be estimated using 2SLS.

[Order condition] iii) Equation (7) is under identified as it fails the order condition because it has no excluded exogenous regressor in the SEM. Therefore, (7) could not be estimated using 2SLS.

4. [Incidental Truncation] Under an incidental truncation model, OLS on the truncated sample will be consistent iff ω and ξ are uncorrelated.

[Heckman Correction] If ω and ξ are correlated, we shall use the Heckman correction protocol. First, we run a probit regression to estimate the selection equation S_i and get $\hat{\pi}$. Second, we include $\lambda(\hat{\pi}^{\mathrm{T}}\mathrm{z})$ as an additional regressor in equation (8) and run OLS, where $\lambda(\cdot)$ is the inverse Mills' ratio. To see whether ω and ξ are correlated, test whether the coefficient of the inverse Mills' ratio is statistically different from naught.

5. [Logit Model & MLE] Let $\beta_0 + \beta_1 \text{Deaths}_i = \beta^{\mathrm{T}} x$. The log-likelihood function for each observation of this standard Logit model is

$$\begin{aligned}
\ell(\beta) &= D_i \ln[\Lambda(\beta^{\mathrm{T}}\mathrm{x})] + (1 - D_i) \ln[1 - \Lambda(\beta^{\mathrm{T}}\mathrm{x})] \\
&= D_i \ln\left[\frac{\exp \beta^{\mathrm{T}}\mathrm{x}}{1 + \exp \beta^{\mathrm{T}}\mathrm{x}}\right] + (1 - D_i) \ln\left[\frac{1}{1 + \exp \beta^{\mathrm{T}}\mathrm{x}}\right] \\
&= D_i[\beta^{\mathrm{T}}\mathrm{x} - \ln(1 + \exp \beta^{\mathrm{T}}\mathrm{x})] - (1 - D_i) \ln(1 + \exp \beta^{\mathrm{T}}\mathrm{x}) \\
&= D_i[\beta_0 + \beta_1 \text{Deaths}_i - \ln(1 + \exp(\beta_0 + \beta_1 \text{Deaths}_i))] \\
&\qquad - (1 - D_i) \ln(1 + \exp(\beta_0 + \beta_1 \text{Deaths}_i))
\end{aligned}$$

For a random sample with *N* observations, the aggregated log likelihood function becomes

$$\begin{aligned}
\mathcal{L}(\beta) &= \sum_{i=1}^{N} \ell(\beta) \\
&= \sum_{i=1}^{N} \{D_i[\beta_0 + \beta_1 \text{Deaths}_i - \ln(1 + \exp(\beta_0 + \beta_1 \text{Deaths}_i))] \\
&\quad - (1 - D_i) \ln(1 + \exp(\beta_0 + \beta_1 \text{Deaths}_i))\}
\end{aligned}$$

[Average Partial Effect] The average partial effect of

$$\mathbb{E}\left[\frac{\partial \Lambda(\hat{\beta}^{\mathrm{T}}\mathrm{x})}{\partial x_1}\right]$$

could estimated using its sample mean

$$\frac{\hat{\beta}_1}{N} \sum_{i=1}^{N} \Lambda'(\hat{\beta}^{\mathrm{T}}\mathrm{x}) = \frac{\hat{\beta}_1}{N} \sum_{i=1}^{N} \frac{e^{\beta_0 + \beta_1 \text{Deaths}_i}}{(1 + e^{\beta_0 + \beta_1 \text{Deaths}_i})^2}$$

The APE measures the overall average impact of a small change in log number of deaths on the probability of having good quality water in parish *i*.

[McFadden/ Efron R-Squared] Since the estimates acquired from iterative MLE process do not minimize variance so the conventional R-Squared does not apply in Logit model. Instead, we may use McFadden's Pseudo R-Squared to measure goodness-of-fit.

$$R^2_{\text{McF}} = 1 - \frac{\ln \mathcal{L}_{ur}}{\ln \mathcal{L}_0}$$

It measures the improvement from null model to the fitted model.