ECON0019 Past Paper 2018

Question A1

1. The sum of squared residual is defined as

$$SSR(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2 = 0$$

We wish to choose the parameter β that minimizes the sum of squared residual

$$\frac{\partial SSR}{\partial \beta} = -2 \sum_{i=1}^{n} x_i (y_i - \beta x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i = \beta \sum_{i=1}^{n} x_i^2$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

2. In order for the OLS estimator to be consistent, we need our model to be linear in parameters and our sample data to be independently and identically distributed (*i.i.d.*) while having a non-zero variation $\mathbb{E}[x^2] > 0$. Furthermore, we should also assume exogeneity, meaning that the error term is mean independent of regressor x.

$$\mathbb{E}[u|x] = 0$$

3. Rewrite the OLS estimator

$$\hat{\beta} = \frac{\sum_{i=1}^{n} \beta x_i^2 + x_i u_i}{\sum_{i=1}^{n} x_i^2} = \beta + \frac{\frac{1}{n}\sum_{i=1}^{n} x_i u_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

The exogeneity assumption implies that

$$\mathbb{E}[u] = 0 \text{ and } \mathbb{E}[ux] = 0 \rightarrow \text{Cov}(u, x) = 0$$

By LLN, any population moment can be consistently estimated by the corresponding sample moment, given that sample is *i.i.d.*. Also assuming non-degenerate and finite second moment:

$$\text{LLN} :: \frac{1}{n}\sum_{i=1}^{n} x_i u_i \xrightarrow{p} \mathbb{E}[x_i u_i] = \mathbb{E}\big[\mathbb{E}[x_i u_i|x]\big] = \mathbb{E}\big[x_i \mathbb{E}[u_i|x]\big] = 0$$

$$\text{LLN} :: \frac{1}{n}\sum_{i=1}^{n} x_i^2 \xrightarrow{p} \mathbb{E}[x^2] > 0$$

$$\text{CLT} :: \frac{1}{n}\sum_{i=1}^{n} x_i u_i \sim^a \mathcal{N}\left(0, \frac{\mathbb{E}[x^2u^2]}{n}\right)$$

Where the variance is derived as follow without assuming homoskedasticity

$$\text{Var}(x_i u_i) = \mathbb{E}[x^2u^2] - \underbrace{\mathbb{E}[x_i u_i]^2}_{=0} < \infty$$

Therefore

$$\frac{\frac{1}{n}\sum_{i=1}^{n} x_i u_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \sim^a \frac{\mathcal{N}\left(0, \frac{\mathbb{E}[x^2u^2]}{n}\right)}{\mathbb{E}[x^2]} = \mathcal{N}\left(0, \frac{\mathbb{E}[x^2u^2]}{n(\mathbb{E}[x^2])^2}\right)$$

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\mathbb{E}[x^2u^2]}{n(\mathbb{E}[x^2])^2}\right)$$

4. The variance of the OLS estimator is

$$\begin{aligned}\text{Var}(\hat{\beta}) &= \text{Var}\left(\frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}\right) \\ &= \frac{1}{(\sum_{i=1}^{n} x_i^2)^2}\text{Var}\left(\sum_{i=1}^{n} x_i u_i\right)\end{aligned}$$

Due to *i.i.d.* assumptions, all cross-covariance term vanishes, so we have the weighted sum of variance.

$$\text{Var}\left(\hat{\beta}\right) = \frac{1}{(\sum_{i=1}^{n} x_i^2)^2}\sum_{i=1}^{n}\text{Var}(x_i u_i) = \frac{\frac{1}{n}}{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)^2} \cdot \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x^2u^2] = \frac{1}{n}\frac{\mathbb{E}[x^2u^2]}{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)^2}$$

Notice that the denominator is consistently estimated by

$$\frac{1}{n}\sum_{i=1}^{n} x^2\hat{u}^2 \xrightarrow{p} \mathbb{E}[x^2u^2]$$

And the numerator consistently estimates

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \xrightarrow{p} \mathbb{E}[x^2]$$

Finally, a consistent estimator of the variance of $\hat{\beta}$ is

$$\widehat{\text{Var}}(\hat{\beta}) = \frac{1}{n} \cdot \frac{\frac{1}{n}\sum_{i=1}^{n} x^2 \hat{u}^2}{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)^2}$$

5. If the true population model includes an intercept

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^{n} x_i^2} = \beta_1 + \beta_0 \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}$$

The conditional expectation of $\hat{\beta}$ is

$$\begin{aligned}
\mathbb{E}[\hat{\beta}|x] &= \beta_1 + \beta_0 \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} + \frac{1}{\sum_{i=1}^{n} x_i^2} \cdot \mathbb{E}[\textstyle\sum_{i=1}^{n} x_i u_i \,|x] \\
&= \beta_1 + \beta_0 \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} + \frac{1}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} \mathbb{E}[x_i u_i|x] \\
&= \beta_1 + \beta_0 \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} + \frac{1}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} x_i \underbrace{\mathbb{E}[u_i|x]}_{=0} \\
&= \beta_1 + \beta_0 \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

We have shown that the estimator is biased with bias equal to

$$\text{Bias} = \beta_0 \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

Question B2

1. The partial effect at the average (PEA) is the average of the partial effect

$$PEA = \frac{\partial \mathbb{E}[D_i|\text{Avail}, \text{Dist}]}{\partial \text{Dist}}\Big|_{\overline{\text{Dist}}} = \beta_{\text{Dist}} + \beta_{\text{Avail}\times\text{Dist}}\overline{\text{Avail}}_i$$

The partial average effect (PAE) is the partial effect evaluated at the average

$$\begin{aligned}
PAE = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbb{E}[D_i|\text{Avail}, \text{Dist}]}{\partial \text{Dist}} &= \frac{1}{n}\sum_{i=1}^{n} \beta_{\text{Dist}} + \beta_{\text{Avail}\times\text{Dist}}\text{Avail}_i \\
&= \beta_{\text{Dist}} + \beta_{\text{Avail}\times\text{Dist}}\overline{\text{Avail}}_i
\end{aligned}$$

Therefore, we see that PEA is identical to PAE for LPM model.

2. The probability of success is

$$\begin{aligned}\Pr(D_i = 1|\chi) &= \Pr(\beta_0 + \beta_1 \text{Avail}_i + \beta_2 \text{Dist}_i + \beta_3 \text{Avail}_i \times \text{Dist}_i - U_i \geq 0) \\ &= \Pr(U_i \leq \beta_0 + \beta_1 \text{Avail}_i + \beta_2 \text{Dist}_i + \beta_3 \text{Avail}_i \times \text{Dist}_i) \\ &= \Phi(\beta_0 + \beta_1 \text{Avail}_i + \beta_2 \text{Dist}_i + \beta_3 \text{Avail}_i \times \text{Dist}_i)\end{aligned}$$

The probability of failure is

$$\Pr(D_i = 0|\chi) = 1 - \Phi(\beta_0 + \beta_1 \text{Avail}_i + \beta_2 \text{Dist}_i + \beta_3 \text{Avail}_i \times \text{Dist}_i)$$

The likelihood function is

$$\prod_{i=1}^{n} [\Phi(\beta_0 + \beta_1 \text{Avail}_i + \beta_2 \text{Dist}_i + \beta_3 \text{Avail}_i \times \text{Dist}_i)]^{D_i} [1 - \Phi(\beta_0 + \beta_1 \text{Avail}_i + \beta_2 \text{Dist}_i + \beta_3 \text{Avail}_i \times \text{Dist}_i)]^{(1-D_i)}$$

The log-likelihood function is

$$\sum_{i=1}^{n} D_i \ln\big(\Phi(\beta_0 + \beta_1 \text{Avail}_i + \beta_2 \text{Dist}_i + \beta_3 \text{Avail}_i \times \text{Dist}_i)\big) + (1 - D_i) \ln[1 - \Phi(\beta_0 + \beta_1 \text{Avail}_i + \beta_2 \text{Dist}_i + \beta_3 \text{Avail}_i \times \text{Dist}_i)]$$

The Average Partial Effect (PAE) is estimated by

$$\mathbb{E}\left[\frac{\Pr(D_i = 1|\chi)}{\partial \text{Dist}}\right] = \frac{1}{n} \sum_{i=1}^{n} \phi\big(\hat{\beta}_0 + \hat{\beta}_1 \text{Avail}_i + \hat{\beta}_2 \text{Dist}_i + \hat{\beta}_3 \text{Avail}_i \times \text{Dist}_i\big) \times \big(\hat{\beta}_2 + \hat{\beta}_3 \text{Avail}_i\big)$$

3. We may adopt two-stage least square (2SLS) estimation protocol to consistently estimate the model where D_i is potentially endogenous. Firstly, we regress the endogenous regressor on the instrument plus all exogenous variables and record the fitted value

$$D_i = \pi_0 + \pi_1 \text{Avail}_i \times \text{Dist}_i + \pi_2 \text{Avail}_i + \pi_3 \text{Dist}_i + u_i \Rightarrow \widehat{D}_i$$

Secondly, we replace D_i in the model by its first-stage estimates and run OLS.

$$Y_i = \delta_0 + \delta_D \widehat{D}_i + \delta_{\text{avail}} \text{Avail}_i + \delta_{\text{Dist}} \text{Dist}_i + V_i$$

We observe a 0.44 reduction in course grade if the student attended the course online. As the TSLS produces a more negative estimate than OLS, it's possible that those who opt for online learning tend to have characteristics that contributes to higher score and bias the OLS estimate toward zero.

We may implement the Hausman test to test whether D_i is indeed endogenous, the intuition is that under exogeneity, both OLS and 2SLS estimate are asymptotically consistent and statistically insignificant difference between them. First, we regress the 2SLS first stage regression, yet we instead estimate the residual, which reflects any endogenous part of D_i that is not explained by the instrument or exogenous component.

$$D_i = \pi_0 + \pi_1 \text{Avail}_i \times \text{Dist}_i + \pi_2 \text{Avail}_i + \pi_3 \text{Dist}_i + u_i \Rightarrow \hat{u}_i$$

Second, we add $\hat{u}_i$ to the original model and run OLS

$$Y_i = \delta_0 + \delta_D D_i + \delta_{\text{avail}} \text{Avail}_i + \delta_{\text{Dist}} \text{Dist}_i + \theta \hat{u}_i + V_i$$

We use *t*-test to check whether $\hat{\theta}$ is statistically significant, where we reject $\mathcal{H}_0$ if significant and conclude OLS estimate is consistent and D_i is indeed endogenous.

4. Notice that equation (5) is exactly the 2SLS first stage regression, a *F*-statistics of 100 is significantly greater than the rejection rule of $F > 10$ proposed by Stock & Goyo (2005), hence we reject the null hypothesis that the instrument is weak (i.e. does not correlate with D_i) and relevance condition is satisfied.
   Intuitively, the interaction term (i.e. instrument) must be significant (i.e. relevant) because the student must travel extra distance if the course is not offered at the student's home campus and therefore contributes to probability of choosing the course online.

   Two homogeneity conditions the authored flagged underpinned the exclusion restriction of instrument, meaning that the interaction term only affects Y_i through D_i, rather than some other unknown pathways that moves Y_i.

5. This $AR(1)$ process violates strict exogeneity assumption, so it is uncertainly not unbiased. However, it could be consistent if contemporaneous exogeneity is not violated.

6. Since strict exogeneity is violated, we may use the Durbin (1970) alternative test as it does not assume strict exogeneity. First, we regress OLS for $\bar{y}_t$ on $\bar{y}_{t-1}$ and obtain estimated residuals $\hat{u}_t$ for all $t = 1,2, \dots, n$. Then, we run an auxiliary regression with coefficient

$$\hat{u}_t = \alpha + \rho\, \hat{u}_{t-1}$$

   Then, we use *t*-test to test whether $\rho = 0$. We reject no autocorrelation is ρ is significant.