

STUDENT ID: \_\_\_\_\_

HEMIS No. \_\_\_\_\_

For Internal Students of  
Royal Holloway

**DO NOT TURN OVER UNTIL TOLD TO BEGIN**

**EC5040 : ECONOMETRICS**

Mid-Term Examination No. 2

Time Allowed: 1 hour

**Answer All 3 questions**

**STATISTICAL TABLES ARE PROVIDED**

*Silent non-programmable calculators may be used*

**PRINT YOUR STUDENT NUMBER ON THE FRONT OF THIS TEST PAPER WHERE INDICATED**

**WRITE ALL YOUR ANSWERS (INCLUDING ROUGH WORKING) ON THIS TEST PAPER. THERE ARE EXTRA BLANK SHEETS TOWARD THE BACK OF THE PAPER**

1. Given the model

$$y = XB + u$$

a) What do you understand by the term endogeneity?

(5 marks)

*Is the term given to the situation when one or more of the regressors in the model are correlated with the error term such that*

$$E(X'u) \neq 0$$

*The 3 main causes of endogeneity are:*

- i) Measurement error in the right hand side variables*
- ii) Simultaneity (2 way causality) between dependent and right hand side variables*
- iii) Omitted variables*

b) Outline the consequences of endogeneity for the consistency of OLS estimates

(12 marks)

Given  $\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'u$

$$= \beta + \left(\frac{X'X}{N}\right)^{-1} \left(\frac{X'u}{N}\right)$$

*The behaviour of this estimator as the sample size gets larger can be determined by taking the probability limits*

$$p \lim(\hat{\beta}) = p \lim \left[ \beta + \left(\frac{X'X}{N}\right)^{-1} \left(\frac{X'u}{N}\right) \right] \quad (1)$$

Since  $\left(\frac{X'X}{N}\right) = \frac{1}{N} \sum_{i=1}^N x_i x_i'$

where  $x_i'$  is the  $i^{\text{th}}$  row of  $X$

*then this sample average will not go to zero as the sample size gets larger (it converges to a finite value, say  $Q_x$ )*

so  $p \lim \left(\frac{X'X}{N}\right) = p \lim \left(\frac{1}{N} \sum_{i=1}^N x_i x_i'\right) = Q_x$

However

$$\left( \frac{X'u}{N} \right) = \frac{1}{N} \sum_{i=1}^N x_i u_i$$

in the presence of endogeneity then the error and right hand side variable(s) are correlated so the average value of this term will **NOT** go to zero as the sample size gets larger (tends to infinity)

So

$$p \lim \left( \frac{X'u}{N} \right) = p \lim \left( \frac{1}{N} \sum_{i=1}^N x_i u_i \right) \neq 0$$

Hence in (1)

$$p \lim(\hat{\beta}) \neq \beta$$

b) Given the model (in mean-deviation form)

$$y = b_1 x_1 + b_2 x_2 + u \quad (1)$$

Suppose the variable,  $x_2$ , is measured with error

Given a partitioned matrix  $W = [y : x_1 : x_2 : x_3]$  where  $x_1$  and  $x_3$  are exogenous

$$W'W = \begin{bmatrix} 78 & 6 & 4 & 3 \\ 6 & 3 & 0 & 6 \\ 4 & 0 & 2 & 1 \\ 3 & 6 & 1 & 5 \end{bmatrix}$$

the sample size is 100

Find the instrumental variable (IV) estimates of the coefficients on  $x_1$  and  $x_2$  and the residual variance of the IV estimator

(16 marks)

Since there are as many instruments as there are endogenous variables ( $l=k$ ) the formula for the IV estimator

$$\hat{\beta}_{GLS} = \hat{\beta}_{IV} = (X'Z(Z'Z)^{-1}Z'X)^{-1} X'Z(Z'Z)^{-1}Z'y$$

simplifies to

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'y$$

In this example

$$W'W = \begin{bmatrix} y'y & y'x_1 & y'x_2 & y'x_3 \\ x_1'y & x_1'x_1 & x_1'x_2 & x_1'x_3 \\ x_2'y & x_2'x_1 & x_2'x_2 & x_2'x_3 \\ x_3'y & x_3'x_1 & x_3'x_2 & x_3'x_3 \end{bmatrix} = \begin{bmatrix} 78 & 6 & 4 & 3 \\ 6 & 3 & 0 & 6 \\ 4 & 0 & 2 & 1 \\ 3 & 6 & 1 & 5 \end{bmatrix}$$

so

$$z'x = \begin{bmatrix} x_1 & x_3 \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} x_1' \\ x_3' \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} x_1'x_1 & x_1'x_2 \\ x_3'x_1 & x_3'x_2 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 6 & 1 \end{bmatrix} \quad \text{and} \quad z'y = \begin{bmatrix} x_1'y \\ x_3'y \end{bmatrix} = \begin{bmatrix} 6 \\ 3 \end{bmatrix}$$

$$\text{so } \hat{\beta}_{IV} = \begin{bmatrix} \hat{\beta}_{IV}^1 \\ \hat{\beta}_{IV}^2 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 6 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 6 \\ 3 \end{bmatrix} = \begin{bmatrix} 1/3 & -0/3 \\ -6/3 & 3/3 \end{bmatrix} \begin{bmatrix} 6 \\ 3 \end{bmatrix} = \begin{bmatrix} 1/3 & 0 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 6 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ -9 \end{bmatrix}$$

To obtain the IV residual variance

need to use the consistent estimator

$$s_{IV}^2 = u_{IV}' u_{IV} / n = (y - Xb_{IV})' (y - Xb_{IV}) / n = (y'y - 2y'Xb_{IV} + b_{IV}'X'Xb_{IV}) / n$$

$$= \frac{78 - 2 \begin{bmatrix} 6 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ -9 \end{bmatrix} + \begin{bmatrix} 2 & -9 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 6 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -9 \end{bmatrix}}{100} = \frac{78 + 48 + 174}{100} = \frac{300}{100} = 3$$

2. Given

$$Y_{it} = X_{it}\beta + u_{it} \quad \begin{matrix} i = 1, 2, \dots, N \\ t = 1, 2, \dots, T \end{matrix}$$

where  $u_{it} = a_i + e_{it}$

$a_i$  is a group specific unobserved component

and  $e_{it} \sim \text{iid}(0, \sigma_u^2)$

a) Show how panel data can deal with unobserved heterogeneity using within-group estimation

(10 marks)

Given

$$Y_{it} = X_{it}\beta + u_{it}$$

Unobserved heterogeneity implies that  $Y_{it} = X_{it}\beta + a_i + e_{it}$

Where  $a_i$  is a set of unobserved characteristics

Assume that the unobserved component  $a_i$  can be modelled as an individual constant term

Follows that summing and averaging within each group of  $T$  observations for individual  $i$   
 Can obtain the set of within-group means and hence

$$\bar{y}_i = \bar{X}_i \beta + a_i + \bar{e}_i$$

so

$$y_i - \bar{y}_i = (X_i - \bar{X}_i) \beta + (e_i - \bar{e}_i)$$

Assuming no autocorrelation or heteroskedasticity in the error term then

$$\tilde{\varepsilon} = (e_i - \bar{e}_i) \sim N(0, \sigma_{\tilde{\varepsilon}}^2)$$

So can remove fixed effect and obtain a consistent estimate of  $B$  by estimating the model in mean deviation form

b) Show that the method of 1<sup>st</sup> differencing will introduce negative autocorrelation into the (differenced) error term with correlation coefficient  $\rho = -0.5$

(15 marks)

Given

$$Y_{it} = X_{it} \beta + u_{it}$$

1<sup>st</sup> differencing implies that

$$\Delta Y_{it} = \Delta X_{it} \beta + \Delta u_{it}$$

where  $\Delta Y_{it} = Y_{it} - Y_{it-1}$                        $\Delta u_{it} = u_{it} - u_{it-1}$

Follows that

$$\text{Var}(\Delta u_{it}) = \text{Var}(u_{it} - u_{it-1}) = \text{Var}(u_{it}) + \text{Var}(u_{it-1}) - 2\text{Cov}(u_{it}, u_{it-1})$$

Assuming the random error term is uncorrelated across individuals and over time then  
 $\text{Cov}(u_{it}, u_{it-1}) = 0$

and so  $\text{Var}(\Delta u_{it}) = \sigma_u^2 + \sigma_u^2 = 2\sigma_u^2$

Similarly

$$\text{Cov}(u_{it}, u_{it-1}) = E(\Delta u_{it}, \Delta u_{it-1})$$

$$(since E(\Delta u_{it}) = E(\Delta u_{it-1}) = 0)$$

so 
$$\begin{aligned} \text{Cov}(u_{it}, u_{it-1}) &= E[(u_{it} - u_{it-1}), (u_{it-1} - u_{it-2})] \\ &= E(u_{it}, u_{it-1}) - E(u_{it-1}^2) + E(u_{it-1}, u_{it-2}) - E(u_{it-2}^2) \end{aligned}$$

given no autocorrelation in the original residuals

$$\text{Cov}(u_{it}, u_{it-1}) = -E(u_{it-1}^2) = -\sigma_u^2$$

It follows that

$$\rho = \text{Corr}(\Delta u_{it}, \Delta u_{it-1}) = \frac{\text{Cov}(\Delta u_t, \Delta u_{t-1})}{\sqrt{\text{Var}(\Delta u_t) * \text{Var}(\Delta u_{t-1})}} = \frac{-\sigma_u^2}{\sqrt{2\sigma_u^2 * 2\sigma_u^2}} = -\frac{\sigma_u^2}{2\sigma_u^2} = -\frac{1}{2}$$

c) Outline a test that can help determine whether to use fixed or random effects estimation in panels

(8 marks)

*Fixed or Random Effects?*

*Crucial assumption is that  $\text{Cov}(X_i a_i) = 0$*

*If this is wrong then random effects will be inconsistent. If correct then random effects is both consistent and more efficient (it is the GLS estimator)*

*Can use a variant of the Hausman test based on this logic*

$$H = \left( \hat{\beta}_{RE} - \hat{\beta}_{FE} \right)' \left[ \text{Var}(\hat{\beta}_{RE}) - \text{Var}(\hat{\beta}_{FE}) \right]^{-1} \left( \hat{\beta}_{RE} - \hat{\beta}_{FE} \right) \sim \chi^2_{(k)}$$

*where k is the number of right hand side parameters excluding the constant*

*If the estimated test value exceeds the critical value then reject the null that the estimates from the two procedures are the same, allowing for sampling variation. In this case random effects would appear to be inconsistent.*

3.

The following is the output from a Heckman two-step estimation of the determinants of the proportion of household budget spent on tobacco

a) Explain why OLS estimation is biased in the presence of selectivity.

(10 marks)

*If only the subset of individuals subject to the treatment (in this case smoking) is observed this is said to be a selected sample*

*It follows that*

$$T_i^* = Z_i\gamma + e_i \quad (1)$$

*$T_i^*$  said to be a latent variable since its value is typically unobserved and in this case can be thought of as representing the propensity to be treated. In practice we only observe whether someone is treated or not  $T_i = 1$  or  $0$ .*

*So presence in the treatment depends on a set of observed factors Z and a residual*

$$\begin{aligned} E(Y_i / T_i = 1) &= E((Y_i / T_i^* > 0)) \\ &= E((Y_i / e_i > -Z_i\gamma)) \\ &= E(X_i\beta_1 + u_{i1} / e_i > -Z_i\gamma) \end{aligned}$$

If  $X$  is non-stochastic then

$$E(Y_i / T_i = 1) = X_i B_1 + E(u_{i1} / e_i > -Z_i \gamma)$$

So the mean value of the residual from a regression of the selected sub-sample is non-zero and hence any OLS estimation on this sub-sample will produce biased and inconsistent estimates of  $B$

b) The probit coefficients in the sub-section of the output headed "select" have been replaced with their marginal effects. Explain the meaning of the term marginal effect and hence evaluate the marginal effects of age and the female dummy variable

(9 marks)

Estimates of the determinants of the probability that an individual belongs to the treatment group do not have the same interpretation as with OLS coefficients (they are simply values that maximise the likelihood function). To obtain coefficients which can be interpreted in a similar way to OLS, need marginal effects

In the case of probit model used in the Heckman two-step method this is given by

$$\frac{\delta \text{Prob}(T_i = 1)}{\delta X_i} = \beta_i \phi(Z_i \gamma)$$

The interpretation of these marginal effects is the impact that a unit change in the variable  $X_i$  has on the probability of belonging to the treatment group. Note that unlike OLS these marginal effects are not constant but vary with the levels of all the  $Z$  variables. (For a discrete explanatory variable it may be better to simply compare the difference between the estimated  $\text{Prob}(T_i=1)$  when  $X_i=1$  and  $X_i=0$  with all other variables in  $Z$  set to their sample mean values.)

So the marginal effect of age is that 1 extra year of age reduces the probability of smoking by 0.7 percentage points

And the effect of being female reduces the probability of being a smoker by 3.2 percentage points.

iii) What do you understand is the role of the variable named "lambda"? Interpret the meaning of the estimated value in the output below

(14 marks)

The term  $\lambda_a$  is called the inverse Mills ratio

In a selected sample

$$\begin{aligned} Y_i / T_i = 1 &= E(Y_i / T_i = 1) + v_i = X_i B_1 + E(u_{i1} / e_i > -Z_i \gamma) + v_i \\ &= X_i B_1 + \rho_{ue} \sigma_u \lambda_a + v_i \end{aligned}$$

equation (5) suggests can deal with the issue like an omitted variable problem. Given an estimate of  $\lambda_a$  could include this as an additional variable and this would correct for the non-randomness of the selected sample. To estimate  $\lambda_a = \frac{\phi(Z_i\gamma / \sigma_e)}{\Phi(Z_i\gamma / \sigma_e)}$  requires an

estimate of  $\gamma$

the coefficient on  $\lambda_a$  would then be an estimate of  $\rho_{ue}\sigma_u$ , the sign of which will therefore indicate the direction of correlation between the unobserveables  $e$  that determine participation in the treatment and the unobserveables  $u$  that determine the level of the outcome variable

Can see that the lambda term is significant and negatively signed – which suggests that the error terms in the selection and primary equations are negatively correlated (since The coefficient on lambda =  $\rho_{eu}\sigma_u$ ). So (unobserved) factors that make smoking more likely tend to be associated (perhaps surprisingly) with a lower budget share of tobacco.

```
. heckman tobacsh logexp age female work , select(age logexp female work edage
) two
```

```
Heckman selection model -- two-step estimates      Number of obs      =      6757
(regression model with sample selection)          Censored obs       =      4952
                                                    Uncensored obs     =      1805

                                                    Wald chi2(8)       =      930.61
                                                    Prob > chi2        =      0.0000
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----						
tobacsh						
logexp	-.043631	.0023872	-18.28	0.000	-.0483099	-.0389522
age	.0007194	.0002001	3.60	0.000	.0003273	.0011116
female	.00508	.0035391	1.44	0.151	-.0018565	.0120165
work	-.0077767	.0042862	-1.81	0.070	-.0161774	.0006241
_cons	.3150016	.0148337	21.24	0.000	.285928	.3440751
-----						
select						
age	-.0077393	.0004111				
female	-.0317458	.0113592				
work	-.0645719	.0141603				
logexp	.0090177	.0077847				
edage	-.088913	.0073195				
-----						
mills						
lambda	-.0224732	.0110691	-2.03	0.042	-.0441683	-.0007781
-----						
rho	-0.31869					
sigma	.07051765					
lambda	-.02247316	.0110691				
-----						

**BONUS MARK**

In a model with measurement error, what is the phrase used to describe the fact that OLS parameter estimates will be closer to zero than the true parameter values?  
 - *attenuation bias*