

## Section A

1. The following output is taken from a regression of the log of hourly pay on a male/female dummy variable, the number of years in the current job, (tenure) and the number of years of education, (yrsted).

Some of the information from the output has been concealed.

Source	Sum of Squares	df	MS			
Model	120.000	3	40.000	Number of obs = 404		
Residual	200.000	400	0.500	F( , ) =		
				Prob > F =		
				R-squared = 0.375		
				Adj R-squared =		
Total	320.000	403	0.794	Root MSE = 0.707		

  

logw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-0.250					
tenure	.045					
yrsted	.040					
_cons	1.170644	.326782	3.58	0.002	.4961988	1.845089

The inverse of the matrix of sum of squares/sum of cross-products of the right hand side variables is given by

	Female	tenure	yrsted
Female	.080		
tenure	-.009	.002	
yrsted	-.040	.001	.004

Consider the variable female:

- Interpret the meaning of the estimated coefficient
- find the standard error of the coefficient
- find its estimated t value (under the null hypothesis that the true coefficient is zero)
- find the 95% confidence interval around this estimate

(10 marks)

i) Since this is a semi-log model, estimate suggests that women earn 25% less than men, on average, ( $d\log w/dfemale = b_{female} = (\% \text{ change in } y / 100)$ ) strictly since this is a dummy variable and a log-lin model should use the formula  $\exp(b_{female}) - 1 = e(-0.25) - 1 = -0.22$  ie 22% less

ii) since  $\text{var}(\hat{\beta}) = s^2(X'X)^{-1}$  and variance of each coefficient is given by  $i$ th element on main diagonal multiplied by  $s^2$ .  
 Since  $s^2 = \text{RSS}/N - k = 0.500$  (could calculate from info. in table or read directly from "MS" column if recognise in stata output)

Then  $\text{var}(\text{female}) = 0.5 * 0.08 = 0.04$   
 and  $\text{se}(\text{female}) = \text{sqrt}(0.04) = 0.2$

iii) Hence t value =  $(-0.25 - 0) / 0.2 = -1.25$

iv) Hence 95% confidence interval =

$$\Pr \left[ \hat{\beta}_1 - t_{n-k}^{\alpha/2} * s.e.(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{n-k}^{\alpha/2} * s.e.(\hat{\beta}_1) \right] = 0.95$$

Given  $n-k = 400$ , 5% critical value for 2 tailed test is 1.96

Hence  $-0.25 - (1.96 * 0.2) \leq \beta_1 \leq -0.25 + (1.96 * 0.2)$

ie 95% confident true female coefficient lies in range

$$-0.642 \leq \beta_1 \leq 0.142$$

b) Test the hypothesis that the return to a year of job tenure is equal to the return to a year of education

(4 marks)

Use fact that  $b_{tenure} = b_{yr sed} \Rightarrow b_{tenure} - b_{yr sed} = 0$

and this test of a single linear restriction reduces becomes

$$F = \frac{(\mathbf{b}_{\logwages} - \mathbf{b}_{\logwealth})^2}{\text{Var}(\mathbf{b}_{\logwages} - \mathbf{b}_{\logwealth})} \sim F[q, N - k] \equiv t = \frac{(\mathbf{b}_{\logwages} - \mathbf{b}_{\logwealth})}{\sqrt{\text{Var}(\mathbf{b}_{\logwages} - \mathbf{b}_{\logwealth})}} \sim t_{N-k}$$

(could do either and  $t$  is just square root of  $F$  for single restriction)

Follows that

$$b_{tenure} - b_{yr sed} = 0.045 - 0.04 = 0.005$$

$$\text{Var}(b_{tenure}) = 0.5 * 0.002 = 0.001$$

$$\text{Var}(b_{yr sed}) = 0.5 * 0.004 = 0.002$$

$$\text{Cov}(b_{tenure}, b_{yr sed}) = 0.001$$

$$\begin{aligned} \text{Hence } \text{Var}(b_{tenure} - b_{yr sed}) &= \text{Var}(b_{tenure}) + \text{Var}(b_{yr sed}) - 2 \text{Cov}(b_{tenure}, b_{yr sed}) \\ &= 0.001 + 0.002 - 2(0.001) = 0.001 \end{aligned}$$

$$\text{So } F = 0.005 / 0.001 = 5$$

From  $F$  tables, 5% critical value for  $F(1, 400)$  is 3.84

Hence estimated  $F >$  critical value so reject null hypothesis that coefficients (returns) are the same

( $t$  test gives  $2.24 > 1.96$  critical value)

c) Suppose that the right hand side variables in the above model (which can be written in matrix form as  $X_{N \times 4}$ ) are subject to a scalar transformation of the form  $Z = XA$ , where  $A$  is a  $4 \times 4$  matrix of constants. Show the consequences for the OLS estimates of a) the coefficients b) the  $R^2$  value in the new model

(6 marks)

$$\begin{aligned} \text{Given } Z = XA \text{ then OLS of } y \text{ on } Z \text{ gives } & (Z'Z)^{-1} Z'y \\ &= (A'X'XA)^{-1} A'X'y \\ &= A^{-1} (X'X)^{-1} A^{-1} A'X'y \\ &= A^{-1} (X'X)^{-1} X'y \\ &= A^{-1} \hat{\beta} \end{aligned}$$

so original OLS coefficients are rescaled by the inverse of A

Also since OLS residuals  $\hat{u} = y - X\hat{\beta} = y - X(X'X)^{-1}X'y = [I - X(X'X)^{-1}X']y$

Then in transformed regression  $\hat{v} = y - Z\hat{\gamma} = y - Z(Z'Z)^{-1}Z'y = [I - Z(Z'Z)^{-1}Z']y$

$\hat{v} = [I - XA(A'X'XA)^{-1}A'X']y = [I - X(X'X)^{-1}X']y = \hat{u}$

so vector of ols residuals are the same in each regression and since  $R^2 = 1 - \text{RSS}/\text{TSS}$

and y values are unaffected so TSS is unaffected then  $R^2$  will be the same in both regressions

d) Outline why and how you would test for the presence of heteroskedasticity in your estimates

(5 marks)

Cross-section data often susceptible to heteroskedasticity – variance in pay not likely to be constant with years of education or tenure.

Outline either Goldfeld-Quandt test or Breusch-Pagan

1. Given the following model, (in mean deviation form), you suspect that the variable,  $x_1$ , is endogenous

$$y = b_1x_1 + b_2x_2 + u \quad (1)$$

a) What are the consequences for OLS estimation of the coefficient  $b_1$  in (1) ?  
(5 marks)

Should be able to show that endogeneity means  $\text{Plim}(X'u/N) \neq 0$  and hence prove that OLS estimates are inconsistent using

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

Given a set of exogenous variables  $X = [x_1 : x_2 : x_3]$  and the partitioned matrix  $W = [y : x_1 : x_2 : x_3]$

$$W'W = \begin{bmatrix} 60 & 5 & 4 & 3 \\ 5 & 6 & 1 & 1 \\ 4 & 1 & 2 & 1 \\ 3 & 1 & 1 & 5 \end{bmatrix}$$

the sample size is 300

b) Find the IV estimates of the coefficients on  $x_1$  and  $x_2$

(6 marks)

IV estimates given by  $(z'x)^{-1}z'y$  where  $x = [x_1 : x_2]$  and  $z = [x_3 : x_2]$

$$z'x = \begin{bmatrix} x_3'x_1 & x_3'x_2 \\ x_2'x_1 & x_2'x_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \quad \text{and} \quad z'y = \begin{bmatrix} x_3'y \\ x_2'y \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

$$\text{so} \quad \hat{\beta}_{IV} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 2/1 & -1/1 \\ -1/1 & 1/1 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

c) Find the variance of the IV residuals

(5 marks)

$$\text{Need } s^2_{IV} = u_{IV}' u_{IV} / n = (y - xb_{IV})' (y - xb_{IV}) / n = (y'y - b_{IV}'x'x b_{IV}) / n$$

$$\text{From answer to above and using } x'x = \begin{bmatrix} x_1'x_1 & x_1'x_2 \\ x_2'x_1 & x_2'x_2 \end{bmatrix} = \begin{bmatrix} 6 & 1 \\ 1 & 2 \end{bmatrix}$$

$$s^2_{IV} = \frac{\left( 60 - \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 6 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} \right)}{300} = 30/300 = 0.1$$

d) Hence find the standard error of the IV estimates and comment on your findings

(4 marks)

$$\text{So } \text{Var}(b_{IV}) = s^2_{IV} (Z'X)^{-1}(X'X)(Z'X)^{-1}$$

$$\text{Var}(\hat{\beta}_{IV}) = 0.1 \left( \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 6 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \right) = 0.1 \begin{bmatrix} -22 & -11 \\ -11 & 6 \end{bmatrix} = \begin{bmatrix} -2.2 & -1.1 \\ -1.1 & 0.6 \end{bmatrix}$$

$$\text{So } \text{s.e.}(\hat{\beta}_{IV}) = \begin{bmatrix} 1.48 \\ 0.77 \end{bmatrix}$$

Hence neither IV estimate of  $b$  is statistically significant at 5% level. Large standard errors could be a sign of weak instruments.

e) Outline a possible test for the endogeneity of  $x_1$  in (1)

(6 marks)

*Wu-Hausman test – compare IV and OLS estimates accounting for sampling variation, Under null IV & OLS consistent but OLS efficient. Under alternative only IV consistent.*

*Reject null if estimated Wu-Hausman exceeds critical value*

*An asymptotic equivalent version of the test is to take the residuals from the 1<sup>st</sup> stage of the regression (the regression of the endogenous variable on its instruments), and include them as additional regressor(s) in the original OLS equation (include an estimated residual for each endogenous rhs variable). Reject null (of no endogeneity) if  $t$  value(s) on added residuals is significant.*

2. Given the model

$$y = X_1\beta_1 + u$$

you suspect that the model has omitted a set of relevant variables  $X_2$

Outline and comment on the consequences for

- i) the bias of the OLS estimates of the coefficient vector  $\beta_1$  (6 marks)
- ii) the standard errors of the OLS coefficient estimates (without using formal proofs)

(7 marks)

*Should be able to show that*

$$\hat{\beta}_1^{ols} = (X_1'X_1)^{-1}X_1'y = (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + u)$$

$$\hat{\beta}_1^{ols} = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'u$$

$$E(\hat{\beta}_1^{ols}) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 \quad \text{since } E(X_1'u) = 0$$

*ie estimates in omitted variable model are biased (unless  $X_1$  and  $X_2$  are orthogonal)*

*Sign of bias depends on i) sign of  $(X_1'X_2)$  which approximates to sign of covariance between  $X_1$  &  $X_2$  (or equivalently on  $(X_1'X_1)^{-1}X_1'X_2$  which is a  $k_1$  by  $k_2$  matrix of coefficients from a regression of each of the  $X_2$  variables on all of the  $X_1$  variables)*

*ii) sign of effect of missing variables on  $y$  ie  $\beta_2$*

*b) parameter variance is also biased in 2 ways*

*i) because (using rules on partitioned matrices)*

$$\text{Var}(\hat{\beta}_1) = \sigma^2(X_1'X_1)^{-1} \text{ but } \text{Var}(\hat{\beta}_{1,2}) = \sigma^2[(X_1'X_1) - X_1'X_2(X_2'X_2)^{-1}X_2'X_1]^{-1}$$

$$\text{and so } \text{Var}(\hat{\beta}_1) < \text{Var}(\hat{\beta}_{1,2})$$

*ii) because OLS estimate of residual variance  $s^2$  is also biased in presence of omitted variables*

$$E(u_1'u_1) > \sigma^2(N - k_1)$$

*(biased up by an amount equal to increase in RSS when  $X_2$  dropped from the regression)*

*This goes in opposite direction to 1<sup>st</sup> bias*

b) What are the consequences for OLS estimation if you include irrelevant variables?

(4 marks)

If include relevant variables then equivalent to imposing (correct) restriction that  $\beta_2 = 0$

But since OLS delivers unbiased estimates of true parameters then expected value of estimates of  $\beta_2$  should = 0 (and expected values of estimates of  $\beta_1 = \beta_1$ )

Only problem is that standard errors are larger than should be since we know that

$$\text{Var}(\hat{\beta}_1) < \text{Var}(\hat{\beta}_{1,2}) \text{ from above}$$

c) Outline the form of a test to check the functional form of your regression. What would you do if the  $X_1$  and  $X_2$  variables were not nested?

(8 marks)

Ramsey RESET test – save predicted values, add higher order polynomials (predicted value is linear functions of original rhs variables) to original equation. If significant reject null of no specification bias

If variables are non-nested use J test

Given

$$H_0: y = X_1 B_1 + u_1 \quad (1)$$

$$H_1: y = X_2 B_2 + u_2 \quad (2)$$

Consider augmented regression

$$y = (1-\alpha) X_1 B_1 + \alpha X_2 B_2 + v$$

and test whether  $\alpha=0$

Asymptotically can show that this is equivalent to taking predicted values,  $X_2 b_2$ , from (2) and using F test of their joint significance