

## Section A

1. The following output is taken from a regression of the log of hourly wages, (*lnhw*), of a sample of individuals on a dummy variable for being female, (*female*), a dummy variable for part-time work, (*part\_time*), the age of the individual in years, (*age*) and the square of age, (*age2*).

Some of the information from the output has been concealed.

```
. reg lnhw female part_time age age2
```

Source	SS	df	MS	Number of obs =		
Model	18.2705596		4.56763989	F( 4, 300) =	16.27	
Residual	84.2341886	300		Prob > F =	0.0000	
				R-squared =	0.1782	
				Adj R-squared =	0.1673	
Total	102.504748		.337186672	Root MSE =	.52989	

  

lnhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-.209473	.066951	-3.13	0.002	-.3412262	-.0777199
part_time	-.2043635			0.011		
age	.0755914	.0161889	4.67	0.000	.0437332	.1074496
age2	-.000821	.0002015	-4.07	0.000	-.0012174	-.0004245
_cons	.701605	.3071563	2.28	0.023	.0971512	1.306059

The  $(X'X)^{-1}$  inverse matrix of sum of squares/sum of cross-products of the right hand side variables is given by

	female	part_time	age	age2	_cons
female	.016				
part_time	-.008	.023			
age	-.00005	.0001	.0009		
age2	1.853e-06	-2.151e-06	-.00001	1.446e-07	
_cons	-.007	-.002	-.017	.0002	.336

Find (to the nearest 3 decimal places where necessary)

- the sample size
- the estimated residual variance of the regression, ( $s^2$ )
- the standard error of the coefficient on the part-time dummy variable
- the effect of being female on the level of hourly pay

(10 marks)

i) Since we know *F* test of goodness of fit of model as a whole is given by value  $F(4, 300) = 16.27$  in the output above and this test  $\sim F(q, N-k)$  where *q* is the number of restrictions, in this case all the right hand side variables excluding the constant) *N* is the sample size and *K* the number of right hand side parameters then in this case  $q=4$  + the constant gives  $k=4+1=5$  so  $N-k=300$  and  $k=5$  implies that the sample size =305

ii) Since  $s^2 = RSS/N-k$  can calculate from info. in  $(X'X)^{-1}$  and standard error info. in regression output or simply calculate from information in stata output

“Residual | 84.2341886 300”

Hence  $s^2 = 84.234/300 = 0.28$

iii) Since  $\text{var}(\hat{\beta}) = s^2(X'X)^{-1}$  and variance of each coefficient is given by  $i^{\text{th}}$  element on main diagonal of  $(X'X)^{-1}$  multiplied by  $s^2$ .

Then  $\text{var}(\text{part\_time}) = 0.28 * 0.023 = 0.006$

and  $\text{se}(\text{part\_time}) = \text{sqrt}(0.006) = 0.077$

iv)

Since this is a semi-log model and age is entered as a quadratic then,

$$d\text{Ln}hw/d\text{age} = b_{\text{age}} + 2 b_{\text{age}^2}$$

= % change in level of pay when age rises by 1 year/100

$$= 0.076 + 2(-.0008)$$

$$= 0.074$$

ie 7.4% more

b) Test the hypothesis that the sum of the coefficients on female and part-time equals -0.5 (and hence the hypothesis that being female and in part-time work reduces log hourly pay by -0.5, other things equal ).

(5 marks)

Use fact that  $b_{\text{female}} + b_{\text{part\_time}} = -0.5 \Rightarrow b_{\text{nemploy}} + b_{\text{agef}} + 0.5 = 0$

and this test of a single linear restriction reduces becomes

$$F = \frac{(b_{\text{female}} + b_{\text{part\_time}} + 0.5)^2}{\text{Var}(b_{\text{female}} + b_{\text{part\_time}})} \sim F[q, N-k] \equiv t = \frac{(b_{\text{female}} + b_{\text{part\_time}} + 0.5)}{\sqrt{\text{Var}(b_{\text{female}} + b_{\text{part\_time}})}} \sim t_{N-k}$$

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

Follows that

$$b_{\text{nemploy}} + b_{\text{agef}} + 0.5 = -0.209 - 0.204 + 0.5 = 0.087$$

and

$$\text{Var}(b_{\text{female}}) = 0.28 * 0.016 = 0.004 \quad \text{and from above} \quad \text{var}(\text{part\_time}) = 0.006$$

$$\text{Similarly Cov}(b_{\text{female}}, b_{\text{part-time}}) = 0.28 * -.008 = -0.002$$

$$\text{Hence Var}(b_{\text{female}} + b_{\text{part-time}}) = \text{Var}(b_{\text{female}}) + \text{Var}(b_{\text{part-time}}) + 2 \text{Cov}(b_{\text{female}}, b_{\text{part-time}}) \\ = 0.004 + 0.006 + 2(-0.002) = 0.006$$

So  $F = (0.087)^2 / 0.006 = 1.262$

From F tables, 5% critical value for  $F(1, 300)$  is 3.88

Hence estimated  $F <$  critical value so accept null hypothesis that coefficients (returns) add to -0.5

(t test gives  $1.12 < 1.96$  critical value)

c) Outline – using matrix algebra if necessary – what the consequence of endogeneity among the right hand side variables will be for the OLS estimates.

(5 marks)

Endogeneity means inconsistency in OLS estimates

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

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

where  $x_i'$  is the  $i$ th row of the X matrix

The  $x_i x_i'$  term will converge to a finite value as the sample size increases and the average value of  $x_i u_i$  will NOT approach zero as N increases if there is endogeneity ( $x_i$  and  $u_i$  are correlated)

Hence  $p \lim(\hat{\beta}_{OLS}) \neq \beta$

d) Outline how you would test for the presence of endogeneity in one of your right hand side variables

(5 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.*

## SECTION B

1. The following is a regression output from an OLS regression of the log of hourly earnings on years of education, (*educ*) and age for a sample of 2296 married women.

```
. reg lhw educ age
```

Source	SS	df	MS			
Model	90.9340802	2	45.4670401	Number of obs =	2296	
Residual	870.040215	2293	.379433151	F( 2, 2293) =	119.83	
				Prob > F =	0.0000	
				R-squared =	0.0946	
				Adj R-squared =	0.0938	
Total	960.974295	2295	.418725183	Root MSE =	.61598	

  

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0791092	.005117	15.46	0.000	.0690748	.0891435
age	.0037069	.0013281	2.79	0.005	.0011024	.0063113
_cons	5.468114	.0960707	56.92	0.000	5.279719	5.656508

a) What do you understand by the term selectivity?

(3 marks)

*If selectivity exists then these coefficients may not be applicable to all married women (working and non-working).*

b) Show what the consequences will be for OLS estimation if selectivity is not accounted for

(6 marks)

*Let presence in treatment group depend on a set of variables Z*

$$T_i^* = Z_i\gamma + \varepsilon_i$$

*And let dummy variable  $T_i=1$  if the underlying continuous variable  $T_i^*>0$  and the individual is observed in the treatment group, = 0 otherwise*

$$E(Y_i/T_i=1) = E(Y_i/T_i^*>0) = E(Y_i/\varepsilon_i > -Z_i\gamma) = E(X_i\beta + u_i/\varepsilon_i > -Z_i\gamma) = X_i\beta + E(u_i/\varepsilon_i > -Z_i\gamma)$$

*So the expected value of the error term is non-zero and so OLS will be biased in the presence of selectivity*

c) Outline how you would control for selectivity the above example.

(12 marks)

*To determine whether selection is a problem, first estimate the probability of being in work, (the probability of being treated) as a function of the original control variables **and** an additional identifying variable. This additional identifying variable is assumed to affect the probability of participation in work, but is assumed not to influence wages on offer once in work, (in practice this assumption should be tested).*

*Using formula for mean of a truncated bivariate normal*

$$E(Y/X>a) = \mu_y + \rho_{xy} \sigma_y \lambda(\alpha)$$

Where  $\alpha = (a - \mu_x) / \sigma_x$  and  $\lambda(\alpha) = \varphi(\alpha) / (1 - \Phi(\alpha))$

Then  $E(Y_i / T_i = 1) = X_i \beta + E(u_i / \varepsilon_i > -Z_i \gamma)$   
becomes

$$E(Y_i / T_i = 1) = X_i \beta + \rho_{u\varepsilon} \sigma_u \lambda(\alpha)$$

Where now  $\alpha = (-Z_i \gamma - 0) / \sigma_\varepsilon$  and  $\lambda(\alpha) = \varphi(Z_i \gamma) / \Phi(Z_i \gamma)$

So by including  $\lambda(\alpha)$  as an extra term in the original model can account for the non-random nature of the error term and hence OLS will be consistent

The next step is to construct the inverse mills ratio  $\lambda = \varphi(Z_i \gamma) / \Phi(Z_i \gamma)$

Where  $\varphi(Z_i \gamma)$  is the standard normal pdf and  $\Phi(Z_i \gamma)$  is the standard normal distribution function evaluated at  $Z_i \gamma$  (the sum of each variable evaluated at its mean value multiplied by its probit estimate =  $\bar{z}_1 \hat{\gamma}_1 + \bar{z}_2 \hat{\gamma}_2 + \dots + \bar{z}_k \hat{\gamma}_k$ )

The final step is to include this lambda term as an additional regressor in the original model and use OLS on this augmented regression

c) The following output contains a selection correction term, (*lambda*)

```
. reg lhw educ age lambda if sex==2 & married==1
```

Source	SS	df	MS			
Model	92.733141	3	30.911047	Number of obs =	2296	
Residual	868.241154	2292	.378813767	F( 3, 2292) =	81.60	
Total	960.974295	2295	.418725183	Prob > F =	0.0000	
				R-squared =	0.0965	
				Adj R-squared =	0.0953	
				Root MSE =	.61548	

  

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0772027	.0051871	14.88	0.000	.0670308	.0873746
age	.0103897	.0033414	3.11	0.002	.0038373	.0169421
lambda	-.3269318	.1500193	-2.18	0.029	-.6211196	-.032744
_cons	5.418822	.098621	54.95	0.000	5.225426	5.612218

Interpret what the sign of the coefficient on the lambda variable implies

(4 marks)

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 participation more likely tend to be associated with lower wages.

2, Given the general linear model  

$$y = X\beta + u$$

and  $E(uu') = \sigma^2\Omega \neq \sigma^2I$

where  $\Omega$  is a diagonal matrix

a) What are the consequences for OLS estimation of the  $\beta$  vector and estimation of its associated variance/covariance matrix?

(7 marks)

$\Omega$  is diagonal implies Heteroskedasticity exists, so

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

Taking expectations

$$E\left[\hat{\beta}_{OLS}\right] = \beta \quad \text{so OLS estimates remain unbiased in presence of heteroskedasticity}$$

but

$$\text{Var}\left(\hat{\beta}_{OLS}\right) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = E[(X'X)^{-1}X'uu'X(X'X)^{-1}]$$

$$\text{Var}\left(\hat{\beta}_{OLS}\right) = \sigma^2[(X'X)^{-1}X'\Omega X(X'X)^{-1}] \neq \sigma^2(X'X)^{-1}$$

so standard errors in OLS based on latter are biased (in an unknown direction which depends on  $\Omega$ ) and since  $t$  and  $F$  values also calculated using latter they are also biased if use OLS.

b) Derive the GLS estimator as a solution to this problem

(7 marks)

Idea is to transform the model  $y = XB + u$  such that residual covariance matrix is homoskedastic rather than heteroskedastic

Consider the transformation matrix  $T$  such that  $Ty = TXB + Tu$  and  $E(Tuu'T) = \sigma^2T'\Omega T = \sigma^2I$

If  $\Omega$  is a positive definite matrix there will always be a transformation such that  $\Omega^{-1} = T'T$

Where  $T = (C\Lambda^{-1/2})$

and  $C$  is a matrix of characteristic vectors of  $\Omega$

and  $\Lambda$  is a diagonal matrix whose non-zero elements are the characteristic roots of  $\Omega$

Hence  $\Omega = (T'T)^{-1} = T^{-1}(T')^{-1}$

and  $T'\Omega T = T[T^{-1}(T')^{-1}]T' = I$

so OLS on  $Ty = TXB + Tu$

gives

$$\hat{\beta}_{GLS} = (X'T'TX)^{-1}(X'T'Ty) = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

which is unbiased and the most efficient estimator in the presence of heteroskedasticity

c) Outline and describe the White (robust) adjustment procedure to OLS estimation

(7 marks)

If functional form of  $\Omega$  is unknown – or wrongly assumed - then FGLS will be inconsistent and asymptotically inefficient and hence OLS may be preferred but with standard errors adjusted to take account of any heterogeneity ie using

$$\text{Var}\left(\hat{\beta}_{OLS}\right) = \sigma^2[(X'X)^{-1}X'\Omega X(X'X)^{-1}] \neq \sigma^2(X'X)^{-1}$$

Consistent estimates of (unknown)  $\Omega$  can be obtained by using

$$\left[ \sum_{i=1}^N \sigma_i^2 x_i x_i' \right] = X'\Omega X = \begin{bmatrix} \cdot & \cdot & \cdot \\ x_1 & \dots & x_N \\ \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix} \begin{bmatrix} \cdot & x_1' & \cdot \\ \cdot & \vdots & \cdot \\ \cdot & x_n' & \cdot \end{bmatrix}$$

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

and replacing the unknown residual variances with the square of the OLS residuals for

each observation  $(u_i = y - x_i' \hat{\beta})^2$

which can be shown to be consistent estimates for  $\sigma_i^2$

Since these are consistent estimates, the correction is only valid asymptotically. The corrected standard errors are consistent but not efficient. However if the exact form of heteroskedasticity is unknown this may be the best that can be done.

d) Outline the White test for the presence of heteroskedasticity

(4 marks)

regress the squared residuals from the model (as proxy for residual variances) on the original right hand side variables together with their squares and cross products. Then  $N \cdot R^2$  from this auxiliary regression is  $\sim \chi^2_{(p)}$  where  $p$  is the number of rhs variables in the auxiliary regression excluding the constant.

If the estimated chi-squared value exceeds the critical value then reject null of no homoskedasticity

(though since this is a joint test, it is hard to say which variable is causing the problem)