

## Mid-Term Test No. 1 2007/08 – Answers

1. Given the general linear model

$$y = X\beta + u$$

where  $y$  is an  $n \times 1$  vector of observations on the dependent variable,  $X$  is an  $n \times k$  matrix of observations on a set of explanatory variables,  $\beta$  is a  $k \times 1$  vector of parameters and  $u$  is an  $n \times 1$  vector of residuals

Assuming that the Gauss-Markov conditions hold then the ordinary least squares (OLS)

estimate of  $\beta$  is given by  $\hat{\beta} = (X'X)^{-1}X'y$  and the OLS residuals are given by  $\hat{u} = y - X\hat{\beta}$

a) Show that the mean of these OLS residuals is zero

(5 marks)

Re-write as  $X'X\hat{\beta} - X'y = 0$

$$\Rightarrow -X'(y - X\hat{\beta}) = 0$$

so  $X'\hat{u} = 0$

It follows that

OLS residuals add to zero  $\sum_{i=1}^N \hat{u}_i = 0$  and so the mean of OLS residuals is zero

b) Show that each regressor variable  $X$  is uncorrelated with the OLS residuals

(7 marks)

$$X'\hat{u} = 0 \Rightarrow \sum_{i=1}^N X_{ki} \hat{u}_i = 0$$

ie the sum of then  $N$  values on each variable  $X_k$  multiplied by the relevant OLS residual equals zero

$$\begin{aligned} \sum_{i=1}^N X_{ki} \hat{u}_i &= \sum_{i=1}^N (X_{ki} - \bar{X}_k + \bar{X}_k) \hat{u}_i = \sum_{i=1}^N (x_{ki} + \bar{X}_k) \hat{u}_i = \sum_{i=1}^N x_{ki} \hat{u}_i + \sum_{i=1}^N \bar{X}_k \hat{u}_i \\ &= \sum_{i=1}^N x_{ki} \hat{u}_i + \bar{X}_k \sum_{i=1}^N \hat{u}_i = \sum_{i=1}^N x_{ki} \hat{u}_i + 0 = Cov(X_k, u) \end{aligned}$$

Hence  $X'\hat{u} = 0 \Leftrightarrow Cov(X, u) = 0$

c) Derive an expression for the variance of the OLS estimator

(6 marks)

$$Var(\hat{\beta}) = E\left[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'\right]$$

$$\begin{aligned}
 \text{Var}(\hat{\beta}) &= E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] \\
 &= E\left[(X'X)^{-1}X'u u'X(X'X)^{-1}\right] \\
 &= (X'X)^{-1}X'E(uu')X(X'X)^{-1} \\
 &= (X'X)^{-1}X'\sigma^2I X(X'X)^{-1} \\
 &= \sigma^2(X'X)^{-1}
 \end{aligned}$$

d) Derive expression for the bias of the OLS estimate of  $\beta$

(5 marks)

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

So

$$E(\hat{\beta}) = \beta + (X'X)^{-1}X'E(u) = \beta \quad (\text{since } E(u)=0)$$

*OLS estimators are unbiased*

e) Suppose that one of the independent variables is subject to a linear transformation, (multiplied by a constant  $\lambda$ ) such that  $Z=X\Lambda$  where  $\Lambda$  is a diagonal matrix containing the transformation constant. Show the effect of this transformation on the OLS estimates of the parameters

(10 marks)

$$\text{Given } y=Z\gamma + v$$

$$\text{OLS implies } \hat{\gamma} = (Z'Z)^{-1}Z'y$$

$$\text{Sub. in } Z=X\Lambda$$

$$\hat{\gamma} = (\Lambda'X'X\Lambda)^{-1}\Lambda'X'y$$

*Using rules on inverse of a matrix product*

$$\hat{\gamma} = \Lambda^{-1}(X'X)^{-1}\Lambda^{-1}\Lambda'X'y$$

$$\hat{\gamma} = \Lambda^{-1}(X'X)^{-1}X'y$$

$$\hat{\gamma} = \Lambda^{-1}\hat{\beta}$$

*If the variable to be transformed is  $X_j$  then the transformation matrix looks like*

$$\Lambda = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & \lambda_j & \\ 0 & & & 1 \end{bmatrix}$$

ie a diagonal matrix with ones down the main diagonal except for the  $j$ th element which contains the constant of multiplication for the  $j^{\text{th}}$  variable

Since the inverse of a diagonal matrix is also diagonal with the reciprocal of each original element on the new main diagonal then

$$\Lambda^{-1} = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & 1/\lambda_j & \\ 0 & & & 1 \end{bmatrix}$$

So using the result in (3) it follows that then the corresponding regression coefficient is multiplied by  $1/a$  and all other coefficients are unchanged.

2. The following regression output is taken from a regression of the % of the household budget spent on food, (values from 0 to 100), on the log of household expenditure, age, age left education and dummy variables for being female, living in London and being employed. Some of the regression output has been obscured.

```
reg foodsh2 logexpeq age edage female london employed
```

Source	SS	df	MS	Number of obs = 2800		
Model	124066.01		20677.6684	F( , ) =	Prob > F = 0.0000	
Residual	229083.645	2793	82.0206392	R-squared =	Adj R-squared = 0.3499	
Total	353149.655		126.169938	Root MSE =	9.0565	

  

foodsh2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
logexpeq	-9.888498	.2890556	-34.21	0.000	-10.45528	-9.321714
age	.0566277	.0136578			.0298473	.0834082
edage	.0544532	.0675362	0.81	0.420	-.0779727	.1868792
female	.2937349	.3666487	0.80	0.423	-.4251949	1.012665
london	1.829847	.5969375	3.07	0.002	.659364	3.000331
employed	.4999172	.4426271	1.13	0.259	-.3679922	1.367827
_cons	70.57788	1.894482	37.25	0.000	66.86315	74.2926

a) Interpret the meaning of the coefficient on the London dummy variable

(3 marks)

*this is a "levels" equation so the impact of living in London relative to not (net of differences in mean values of control variables between London sub-sample and others – see problem set 4) is 1.83*

*Other things equal households in London spend 1.8 **percentage points** (not percent) more of their budget on food than other households*

b) Interpret the meaning of the coefficient on the log of expenditure

(4 marks)

*The coefficient on log expenditure is a semi-elasticity and gives the percentage point change in the budget share of food following a 1% change in total household expenditure, multiplied by 100, (since  $dw/d\text{Log}(x) = b_i = dw/(dx/x) = \text{unit change in } w \text{ with respect to a 1 percentage change in } x * 100$ )*

*So a 1% increase in expenditure is associated with a 0.099 percentage point fall (-9.89/100) in the share of the household budget spent on food. The negative sign confirms that food is a necessity (expenditure share falls as income rises)*

c) Find the estimate of  $R^2$

(4 marks)

$$R^2 \text{ (the coefficient of determination)} = ESS/TSS = 1 - (RSS/TSS)$$

*From information in the regression output (highlighted in yellow)*

$$R^2 = ESS/TSS = 124066.01/353149.655 = 0.351$$

d) Test the hypothesis that the true coefficient on age is significantly different from zero

(5 marks)

Using

$$t = \frac{\hat{\beta}_i - \beta_i^{null}}{s.e.(\hat{\beta}_i)} \sim t_{(N-k)}$$

Again relevant information is given (highlighted in green) in regression output

$$= 0.0566/0.01365 = 4.15$$

which since 95% critical value  $t_{2793}$  (2 tailed test) = 1.96 then absolute value of estimated  $t$  lies outside acceptance region. So **reject** null that age has no explanatory power in the model

e) Test the hypothesis that the true coefficient on the female dummy is significantly different from 0.3

(5 marks)

Again using

$$t = \frac{\hat{\beta}_i - \beta^{null}}{s.e.(\hat{\beta}_i)} \sim t_{(N-k)} = (0.294 - 0.3)/0.367 = -0.016$$

which since 95% critical value  $t_{2793}$  (2 tailed test) = 1.96 then absolute value of estimated  $t$  lies inside acceptance region. So **accept** null that female effect is not significantly different from 0.3

f) Test the hypothesis the hypothesis that the model as a whole is a good fit (that the effect of all the variables in the model excluding the constant is zero)

(6 marks)

$$\text{Test of goodness of fit of the model is given by } F = \frac{ESS / k - 1}{TSS / N - k} = \frac{R^2 / k - 1}{(1 - R^2) / N - k} \sim F[k - 1, N - k]$$

$$\text{So } F = \frac{124066/7-1}{353149/2800-7} = \frac{0.351/7-1}{(1-0.351)/2800-7} \sim F[7-1, 2800-7]$$

$$= 252.1 \sim F[6, 2793]$$

So estimated  $F$  is greater than 5% critical value ( $F(6, \infty) = 2.09$ ) so reject null that model as a whole has no explanatory power

The variance/covariance matrix of the OLS parameter estimates (excluding the constant) is given by

	logexpeq	age	edage	female	london	employed
logexpeq	.08355316					
age	-.00042658	.00018654				
edage	-.00418811	.00024251	.00456114			
female	.00922894	.00058968	.00013298	.13443127		
london	.00757201	-7.814e-06	-.00670241	-.00574375	.3563344	
employed	-.04441312	.00284793	-.00050682	.02653622	.00499103	.19591879

2. Consider a simple model of 100 observations split equally into two sub-samples such that

$$y_i = a_1 + b_1 X_i + u_i \quad i=1..N_1 \quad \text{in sub-sample 1}$$

and

$$y_i = a_2 + b_2 X_i + u_i \quad i=N_1+1..N \quad \text{in sub-sample 2}$$

a) Show how all four parameters could be obtained from a single OLS regression. (9 marks)

b) Suppose that  $RSS_1 = 10$  and  $RSS_2 = 5$  and that the RSS from the pooled regression is 17. Test the hypothesis of no structural change across the two sub-samples at the 5% level. (7 marks)

The unrestricted form of the model (intercepts and the slopes vary in two periods) in (partitioned) matrix form is given by

$$y = \begin{bmatrix} y_1 \\ \dots \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & : & 0 \\ \dots & \dots & \dots \\ 0 & : & X_2 \end{bmatrix} \begin{bmatrix} a_1 \\ b_1 \\ a_2 \\ b_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ \dots \\ u_2 \end{bmatrix} = X\beta + u \quad (1)$$

where  $X_1$  is an  $N_1$  by 2 matrix of observations from the 1<sup>st</sup> sub-sample and  $X_2$  is an  $N_2$  by 2 matrix of observations from the 2<sup>nd</sup> sub-sample with  $N = N_1 + N_2$

ie stacking the data from the second period below that of the observations from the 1st period in a way that allows the coefficients to differ between the periods

OLS on (1) gives 
$$\hat{\beta} = (X'X)^{-1} X'y$$

which using rules on inverse of partitioned matrices (the inverses of the elements on a diagonal partitioned matrix are just the inverses of the elements themselves) can be written as

$$\hat{\beta} = \begin{bmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{bmatrix} \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix} = \begin{bmatrix} (X_1'X_1)^{-1} X_1'y \\ (X_2'X_2)^{-1} X_2'y \end{bmatrix}$$

which is identical to those obtained by running OLS separately on the two sub-samples

Compare this with estimates from the restricted (pooled) model based on

$$y = \begin{bmatrix} y_1 \\ \dots \\ y_2 \end{bmatrix} = \begin{bmatrix} i & X_1 \\ i & X_2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} u_1 \\ \dots \\ u_2 \end{bmatrix} = X\beta + u$$

To test formally use

$$F = \frac{(RSS_{restricted} - RSS_{unrestricted})/q}{RSS_{unrestricted}/N - k} \sim F[q, N - k]$$

which in this case becomes the Chow test

$$F = \frac{(RSS_{restricted} - RSS_1 + RSS_2)/q}{RSS_1 + RSS_2 / N - 2k} \sim F[q, N - 2k]$$

(remember that there are 4 parameters in the unrestricted model so  $k=4$  and  $q=2$  restrictions)

$$\text{hence } \hat{F} = \frac{(17 - (10 + 5))/2}{(10 + 5)/100 - 2 * 2} = 6.4$$

From Tables the 5% critical value given the degrees of freedom  $F^{05}[2, 96] = 3.1$

$\hat{F} > F_{critical}$  so reject null (of no structural change)

c) You suspect that this model should contain additional explanatory variables. Outline the consequences for OLS estimation of such a mis-specified model.

(10 marks)

True:  $y = X_1\beta_1 + X_2\beta_2 + e$

Estimate:  $y = X_1\beta_1 + u$

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y$$

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

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

so

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

estimates of the coefficients on the set of  $X_1$  variables are biased in the presence of omitted variables

and sign of bias depends on a) the effect of the omitted variables on  $y$ ,  $\beta_2$ ,  
b) the coefficient from a regression of  $X_1$  on  $X_2$

d) Given the following regression output

```
reg lhw exper grad
```

Source	SS	df	MS			
Model	854.145433	2	427.072717	Number of obs =	17321	
Residual	5312.01729	17318	.306733878	F( 2, 17318) =	1392.32	
				Prob > F =	0.0000	
				R-squared =	0.1385	
				Adj R-squared =	0.1384	
Total	6166.16272	17320	.356014014	Root MSE =	.55384	
lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	.0080000	.0004000	20.00	0.000	.0072936	.0086807
grad	.6000000	.0120000	50.00	0.000	.59087	.6379807
_cons	1.691547	.0091817	184.23	0.000	1.67355	1.709544

```
reg grad exper
```

Source	SS	df	MS			
Model	80.2003462	1	80.2003462	Number of obs =	17321	
Residual	2123.91997	17319	.122635254	F( 1, 17319) =	653.97	
				Prob > F =	0.0000	
				R-squared =	0.0364	
				Adj R-squared =	0.0363	
Total	2204.12032	17320	.127258679	Root MSE =	.35019	

  

grad	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	-.0060000	.0002000	-30.00	0.000	-.0060467	-.0051857
_cons	.2500000	.0050000	50.00	0.000	.2600229	.2813061

Find the OLS estimate of exper in a simple regression of lhw on exper

(7 marks)

- using the omitted variable bias formula above to give

$$\hat{\beta}_{\text{exper}}^{2 \text{ variable}} = \hat{\beta}_{\text{exper}}^{3 \text{ variable}} + \left( \hat{\beta}_{\text{grad}}^{\text{auxillary}_\text{grad.exper}} \right) \hat{\beta}_{\text{grad}}^{3 \text{ variable}}$$

$$= 0.008 + (-.006 * .6)$$

$$= .0044$$

3. The following regression output is based on the OLS estimates from a sample of 800 individuals who were asked about the number of cigarettes smoked a day (*numcig*) and the log of the average price in the local area, (*lncigp*), together with a set of demographic characteristics: age and its square, log weekly income (*lninc*), years of education (*years*ed).

```
. reg numcigs age age2 yearsed lncigp lninc, robust
```

Regression with robust standard errors

```
Number of obs =      807
F(   5,   801) =    11.36
Prob > F       =    0.0000
R-squared      =    0.0451
Root MSE     =    13.45
```

numcigs	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
age	.7806352	.1401858	5.569	0.000	.5054602	1.05581
age2	-.0091056	.0014817	-6.146	0.000	-.012014	-.0061972
years	-.5141422	.1627735	-3.159	0.002	-.8336552	-.1946291
lncigp	-2.853151	5.989163	-0.476	0.634	-14.60946	8.903157
lninc	.7582931	.5979058	1.268	0.205	-.4153542	1.93194
_cons	5.368736	25.37082	0.212	0.832	-44.43241	55.16988

a) Find the effect of

i) age

ii) price

on the number of cigarettes consumed

(6 marks)

- Since age is entered as a quadratic, an extra year of age influences the number of cigarettes smoked each day by around

$$d(\text{numcigs})/d\text{Age} = 0.78 - 2 \cdot 0.009\text{Age}$$

so that cigarette consumption reaches a (theoretical) maximum at  $0.78/0.018 = 43$  and then begins to fall back

Price appears in logarithmic form. Hence the coefficient is a semi-elasticity. However the *t* value is so low that can't reject the null hypothesis that the effect of price on cigarette consumption is zero

b) The regression uses the "robust" correction for heteroskedasticity. Derive the formula for the variance of the OLS estimator under heteroskedasticity and show how the robust estimate of this variance is estimated in practice

(8 marks)

$$\begin{aligned} \text{Var}(\hat{\beta}) &= E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] \\ &= E[(X'X)^{-1}X'uu'X(X'X)^{-1}] \\ &= (X'X)^{-1}X'E(uu')X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2\Omega X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1} \end{aligned} \quad (1)$$

White (1980) showed that a consistent estimate of the true OLS variance is given by

$$\text{Var}(\hat{\beta}_{OLS}) = (X'X)^{-1} S (X'X)^{-1}$$

$$\text{where } S = \frac{\sum_{i=1}^N u_i^2 x_i' x_i}{N} \rightarrow E(u^2 X'X) = \sigma^2 X' \Omega X$$

This variance/covariance matrix is said to be **robust** to the presence of unknown heteroskedasticity

c) Outline the form of the Breusch-Pagan test for heteroskedasticity (8 marks)

In general it is unlikely that you will know the variable causing heteroskedasticity or there may be more than one variable responsible.

Instead we would like to test whether the residual variance depends on a set of variables, Z

$$u_i^2 = d_0 + Z_i \delta + e_i \quad (1)$$

Since  $u_i^2$  is unobserved replace with OLS residuals (which are consistent estimates of  $u_i$  if the model is correctly specified)

$$\hat{u}_i^2 = d_0 + Z_i \delta + e_i \quad (2)$$

and estimate (2) by OLS

Can show that

$$N^* R^2 \stackrel{asy}{\sim} \chi_r^2$$

where  $r$  = number of right hand side variables in Z (ie excluding constant)

[this is the **Breusch-Pagan** test for heteroskedasticity and belongs to the set of test known as LM tests – see Johnston & DiNardo ch. 5 for derivation]

If  $N^* R^2 > \chi_{critical}^2$  then reject null of homoskedasticity

d) You have data of N individuals, with observations on each case for two consecutive periods. The model is given by

$$y_{it} = X_{it} B + u_{it} \quad i=1, 2 \dots N; \quad t=1, 2$$

The residual contains a case specific effect,  $m_i$

$$u_{it} = m_i + e_{it}$$

where

$$m_i \sim \text{iid } N(0, \sigma_m^2) \quad \text{and} \quad e_{it} \sim \text{iid } N(0, \sigma_e^2)$$

Write down the form for the variance/covariance matrix of u,  $E(uu')$

(11 marks)

- Given the above information then the residual consists of a group-specific component that is invariant over time ( $m_i$ ) and a component  $e_{it}$  that varies over both individuals and time.

Ordering the data by i) individual ii) time gives a vector of residuals

$$u' = [u_{11} \quad u_{12} \quad u_{21} \quad u_{22} \quad \dots \quad u_{N2}]$$

and a residual covariance matrix

$$uu' = \begin{bmatrix} u_{11}^2 & u_{11}u_{12} & & \dots & u_{11}u_{N2} \\ u_{12}u_{11} & u_{12}^2 & & & \\ & & u_{21}^2 & & \\ & & & \vdots & \\ u_{N2}u_{11} & & \dots & u_{N2}u_{N1} & u_{N2}^2 \end{bmatrix}$$

$$\text{Now } \text{Cov}(u_{it}u_{is}) = E \left[ (u_{it} - \bar{u})(u_{is} - \bar{u}) \right] = E((u_{it}u_{is}))$$

Since  $E(\bar{u}) = 0$

$$\text{Hence } \text{Cov}(u_{it}u_{is}) = E[(m_i + \varepsilon_{it})(m_i + \varepsilon_{is})] = E[m_i m_i + m_i \varepsilon_{is} + \varepsilon_{it} m_i + \varepsilon_{it} \varepsilon_{is}]$$

and so  $\text{Cov}(u_{it}u_{is})$

$$\begin{aligned} &= \sigma_m^2 + \sigma_e^2 && \text{if } i=j \text{ and } t=s \\ &= \sigma_m^2 && \text{if } i=j \text{ and } t \neq s \\ &= 0 && \text{if } i \neq j \end{aligned}$$

Hence with  $t=1$  or  $2$  then

$$E(uu') = \begin{bmatrix} \sigma_m^2 + \sigma_e^2 & \sigma_m^2 & 0 & \dots & 0 \\ \sigma_m^2 & \sigma_m^2 + \sigma_e^2 & & & \\ 0 & & \sigma_m^2 + \sigma_e^2 & \sigma_m^2 & \\ \vdots & & \sigma_m^2 & \vdots & \sigma_m^2 \\ 0 & \dots & 0 & \sigma_m^2 & \sigma_m^2 + \sigma_e^2 \end{bmatrix}$$

ie a block diagonal matrix which can in principle be estimated using GLS given consistent estimates of the individual elements