

Mid-Term Test No. 1 2006/07 – Answers

1. Given the general linear model

$$y = XB + u$$

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, B is a $k \times 1$ vector of parameters and u is an $n \times 1$ vector of residuals

a) Derive, from first principles, an expression for the ordinary least squares (OLS) estimate of B

(8 marks)

Minimising the sum of squared residuals implies

$$\begin{aligned} \text{Min}_{\beta} \hat{u}'\hat{u} &= \begin{bmatrix} \hat{u}_1 & \hat{u}_2 & \dots & \hat{u}_n \end{bmatrix} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix} = \hat{u}_1^2 + \hat{u}_2^2 + \dots + \hat{u}_n^2 \\ &= (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

Since all terms are scalars (1x1) can add

$$= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$$

F.O.C. minimum

$$\frac{\partial \hat{u}'\hat{u}}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

which gives k normal equations $X'X\hat{\beta} = X'y$

and the k variable OLS solution $\hat{\beta} = (X'X)^{-1}X'y$

b) Suppose the independent variables are subject to a linear transformation $Z = X\Lambda$ where Λ is a diagonal matrix of constants. Show the effect of this transformation on the OLS estimates of the parameters

(8 marks)

Given $y = Z\gamma + v$

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

Sub. in $Z = X\Lambda$

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

Using rules on inverse of a matrix product

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

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

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

Since Λ is diagonal matrix, so original i th OLS estimate is multiplied by reciprocal of value on i th element on main diagonal of Λ

c) Show that the OLS estimate of any single coefficient, b_i , in a multiple regression is the same as that obtained in a simple regression together with a correction factor that takes account of the association between x_i and the other variables

(10 marks)

Given

$$y = X\hat{\beta} + u$$

Consider partitioning the X matrix into $X=[x_1 : X_2]$

ie the $N \times 1$ vector of observations on a single variable (x_1)

and

the $N \times (k-1)$ matrix of observations on the other $k-1$ right-hand side variables (including the constant)

so

$$y = \begin{bmatrix} x_1 & X_2 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{\beta}_1 \end{bmatrix} + u$$

Given OLS normal equations

$$X'X\hat{\beta} = X'y$$

$$\begin{bmatrix} x_1' \\ X_2' \end{bmatrix} \begin{bmatrix} x_1 & X_2 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} x_1' \\ X_2' \end{bmatrix} y$$

so

$$\begin{bmatrix} x_1'x_1 & x_1'X_2 \\ X_2'x_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} x_1'y \\ X_2'y \end{bmatrix}$$

1st row can be written

$$(x_1'x_1)^{-1}\hat{b}_1 + (x_1'X_2)\hat{\beta}_1 = x_1'y$$

so

$$\hat{b}_1 = (x_1'x_1)^{-1}x_1'y - (x_1'x_1)^{-1}(x_1'X_2)\hat{\beta}_1$$

Hence OLS estimate of coefficient b_1 in multiple regression is the same as that obtained in a simple regression together with a correction factor that takes account of the association between x_1 and the other variables

d) Outline the form of a technique that could be used to test for specification error bias
(7 marks)

Either

Ramsey RESET Test

- if model is good fit then addition of extra variables should not be statistically significant

rather than add higher order terms of original variables a more parsimonious alternative is to use fact that

$$\hat{y} = X \hat{\beta}$$

so predicted values are linear function of all the X variables (weighted by their estimated coefficients)

and hence $(\hat{y})^j = (X \hat{\beta})^j$

are linear functions of higher powers of all the X variables

$$y = X\beta + \delta_2 \hat{y}^2 + \delta_3 \hat{y}^3 + \dots + \delta_j \hat{y}^j + u$$

and test null $H_0: \delta_2 = \delta_3 = \dots \delta_j = 0$

If estimated F value greater than critical value reject null that functional form is acceptable.

OR

LM Test of Omitted Variables

1. Run restricted regression (no higher order terms)

2. save residuals

3. Regress residuals on unrestricted model (containing higher order values of X (or the \hat{y}_j) - the auxiliary regression

Can show

$$NR^2_{aux} \stackrel{a}{\sim} \chi^2_{(No.ofrestrictions)}$$

If estimated Chi-squared value greater than critical value reject null that functional form is acceptable.

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(,)	=	
Residual	229083.645	2793	82.0206392	Prob > F	=	0.0000
				R-squared	=	
Total	353149.655		126.169938	Adj R-squared	=	0.3499
				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)

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

g) test the hypothesis that the coefficient on the employed dummy equals the coefficient on the female dummy

(6 marks)

In this case

$$H_0: \hat{\beta}_{female} = \hat{\beta}_{employed} \equiv \hat{\beta}_{female} - \hat{\beta}_{employed} = 0$$

$$\text{So use } F = \frac{\left(\hat{\beta}_{\text{female}} - \hat{\beta}_{\text{employed}} - 0\right)^2}{\text{Var}(\hat{\beta}_{\text{female}} - \hat{\beta}_{\text{employed}})} = \frac{\left(\hat{\beta}_{\text{female}} - \hat{\beta}_{\text{employed}}\right)^2}{\text{Var}(\hat{\beta}_{\text{female}}) + \text{var}(\hat{\beta}_{\text{employed}}) - 2\text{Cov}(\hat{\beta}_{\text{female}}, \hat{\beta}_{\text{employed}})}$$

The relevant variances and covariances are highlighted in red above

$$\begin{aligned} \text{Hence } F &= (.294 - .499)^2 / (.134 + .196 - 2 * .027) \\ &= .042 / (0.276) = 0.152 \sim F[1, 2800 - 7] \end{aligned}$$

From F tables $F_{\text{critical}}^{5\%} [1, 2793] = 3.84$

$F < F_{\text{critical}}^{5\%}$ so accept null that coefficients are equal.

(effectively the standard errors around the estimates are so large that the 95% confidence intervals for both variables include the values of the other coefficient, as can be seen in the regression output above)

3. Given the general linear model

$$y = XB + u$$

you suspect the presence of heteroskedasticity of the form $E(uu') = \sigma^2 \Omega$

a) Write down expressions for the form of the GLS estimator and its variance when heteroskedasticity is of this form

(6 marks)

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

$$\text{Var}(\hat{\beta}_{GLS}) = \sigma^2 (X' \Omega^{-1} X)^{-1}$$

b) Suppose the data underlying the model above is grouped such that

$$\bar{y}_g = \bar{X}_g \beta + \bar{u}_g \quad g = 1, G \text{ groups}$$

Write down the form of the residual variance/covariance matrix in this case. Give reasons for your answer.

(5 marks)

$$\sigma^2 \Omega = \sigma^2 \begin{bmatrix} 1/N_1 & & & \\ & 1/N_2 & & \\ & & \vdots & \\ & & & 1/N_G \end{bmatrix}$$

$$\begin{aligned} \text{Var}(\bar{u}_g) &= \text{Var}\left[\frac{1}{N_g} \sum_i u_i\right] = \frac{1}{N_g^2} \text{Var}\left[\sum_i u_i\right] \\ &= \sigma^2/N_g \end{aligned}$$

so the residual variance **falls** as the size of the group on which the average is based **rises**

Suppose that the X matrix consists of a constant and observations on a single explanatory variable. The weighted matrices of products and cross-products are given by

$$\sum_{j=1}^G N_g x_j x_j' = \begin{bmatrix} 10 & 5 \\ 5 & 3 \end{bmatrix} \quad \sum_{j=1}^G N_g x_j y_j = \begin{bmatrix} 5 \\ 1 \end{bmatrix} \quad \sum_{j=1}^G N_g y_j y_j = 15$$

Assume the sample size is 8

c) Find the GLS (weighted least squares) estimates of the constant and the slope (10 marks)

So

$$\hat{\beta}_{GLS} = \begin{bmatrix} 10 & 5 \\ 5 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 5 \\ 1 \end{bmatrix} = \begin{bmatrix} 3/5 & -5/5 \\ -5/5 & 10/5 \end{bmatrix} \begin{bmatrix} 5 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

d) Find the GLS (weighted least squares) standard errors for the constant and slope estimates (12 marks)

$$\text{Var}(\hat{\beta}_{GLS}) = \sigma^2 (X' \Omega^{-1} X)^{-1}$$

but σ^2 unobserved, so replace with unbiased estimator which in this case = s_{GLS}^2

$$\text{where } s_{GLS}^2 = \frac{(y^* - X^* \hat{\beta}_{GLS})'(y^* - X^* \hat{\beta}_{GLS})}{N - k} = \frac{\hat{u}^* \hat{u}^*}{N - k}$$

$$\text{using } \hat{u}^* \hat{u}^* = (y^* - X^* \hat{\beta}_{GLS})'(y^* - X^* \hat{\beta}_{GLS}) = y^{*'} y^* - \hat{\beta}'_{GLS} X^{*'} y^* = y' \Omega^{-1} y - \hat{\beta}'_{GLS} X' \Omega^{-1} y$$

$$\text{and } \hat{u}^* \hat{u}^* = 25 - \begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} 5 \\ 1 \end{bmatrix} = 25 - 13 = 12$$

$$\hat{u}^* \hat{u}^* = 25 - \begin{bmatrix} 2 & -3 \end{bmatrix} \begin{bmatrix} 5 \\ 1 \end{bmatrix} = 25 - 10 + 3 = 18$$

$$\text{so } s_{GLS}^2 = \frac{\hat{u}^* \hat{u}^*}{N - k} = \frac{12}{8 - 2} = 2$$

$$\frac{\hat{u}'\hat{u}}{N-k} = \frac{18}{8-2} = 3$$

and hence the GLS variance is given by

$$\text{Var}(\hat{\beta}_{GLS}) = s^2(X'\Omega^{-1}X)^{-1} = 3 \begin{bmatrix} 3/5 & -1 \\ -1 & 2 \end{bmatrix}$$

Since the variance of the i^{th} parameter estimates occurs at the i^{th} position on the main diagonal, it follows that

$$\text{Var}(\hat{\beta}_0) = 9/5 \quad \text{Var}(\hat{\beta}_1) = 6$$

Hence

$$\text{s.e.}(\hat{\beta}_0) = 1.34 \quad \text{s.e.}(\hat{\beta}_1) = 2.45$$

(on this basis neither parameter estimate is significantly different from zero)

Bonus Mark

The LM test for functional form is (asymptotically) follows which distribution?

Chi-squared