

UNIVERSITY OF LONDON
DEPARTMENTAL EXAMINATION 2009

For Internal Students of
Royal Holloway

DO NOT TURN OVER UNTIL TOLD TO BEGIN

EC5040: ECONOMETRICS

Mid-Term Examination No. 1

Time Allowed: 1 hour

Answer **every** question

Please answer each question on a separate page

College Calculators are provided
Statistical Tables are attached

© Royal Holloway University of London 2009

Mid-Term Test No. 1 2009/10 – 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, β 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 β

(10 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) Derive an expression for the variance of the OLS estimator

(7 marks)

Since $\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'u$ *and OLS estimate of* β *is unbiased*

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

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

(12 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

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

Given OLS normal equations $X'X\hat{\beta} = X'y$

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

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 both of the association between x_1 and the other variables, $(x_1'X_2)$, and the effect of the other variables on the dependent variable, B_2 .

Note that $(x_1'x_1)^{-1} X_1'X_2$ is a k_1 by k_2 matrix of coefficients equal to the ols coefficients from a regression of each of the X_2 variables on all the X_1 variables

d) Show that $s^2 = \frac{\hat{u}'\hat{u}}{N-k}$ is an unbiased estimate of the (unobserved) true residual variance σ^2

(21 marks)

$$\begin{aligned} \text{Since } \hat{u} &= y - X\hat{\beta} &= y - X(X'X)^{-1}X'y \\ & &= [I - X(X'X)^{-1}X']y \end{aligned}$$

$$=My$$

where M is a "residual maker" symmetric, idempotent matrix with the properties that

$$MX = 0 \quad My = \hat{u} \quad M\hat{u} = \hat{u}$$

$$\begin{aligned} \text{It follows that } Mu &= [I - X(X'X)^{-1}X'] [y - X\beta] = y - X(X'X)^{-1}X'y - X\beta + X(X'X)^{-1}X'X\beta \\ &= y - X\beta \\ &= u \end{aligned}$$

$$\therefore \hat{u}'\hat{u} = u'M'Mu = u'Mu \quad (\text{since } M \text{ idempotent})$$

$$\text{and so } E(\hat{u}'\hat{u}) = E(u'Mu)$$

Since $u'Mu$ is a scalar 1×1

use the result that the trace of a scalar = sum of elements on the main diagonal which in this case is just the value of the scalar

$$\begin{aligned} \text{so } E(\hat{u}'\hat{u}) &= E(u'Mu) = E[\text{tr}(u'Mu)] = E[\text{tr}(Mu u')] \\ &\quad - \text{using the result that } \text{tr}(ABC) = \text{tr}(BAC) = \text{tr}(BCA) = \text{tr}(CBA) \end{aligned}$$

Since M is assumed to be non-stochastic can take it outside the expectation

$$\begin{aligned} \text{so } E(\hat{u}'\hat{u}) &= \text{tr}[ME(uu')] \\ &= \sigma^2 \text{tr}(M) \\ &= \sigma^2 [\text{tr}(I_N - [X(X'X)^{-1}X'])] \\ &= \sigma^2 [\text{tr}(I_N) - \text{tr}[X(X'X)^{-1}X']] \\ &= \sigma^2 [\text{tr}(I_N) - \text{tr}[(X'X)^{-1}X'X]] \\ &= \sigma^2 [\text{tr}(I_N) - \text{tr}[I_k]] \end{aligned}$$

since the trace is the sum of the elements on the main diagonal

$$E(\hat{u}'\hat{u}) = \sigma^2 [N - k]$$

Hence it follows that $s^2 = ((\hat{u}'\hat{u}) / (N - k))$ will be an unbiased estimator of σ^2

$$\begin{aligned} E(\hat{u}'\hat{u} / (N - k)) &= (\sigma^2 [N - k]) / (N - k) \\ &= \sigma^2 \end{aligned}$$

2. The following regression output is taken from a regression of the log of food expenditure (*lfood*) on income (*income*), years of education (*yrsed*), age and the square of age, (*age*, *age2*) a dummy variable for being female, (*female*).

Some of the regression output has been hidden

```
reg lfood income yrsed age age2 female
```

Source	SS	df	MS	Number of obs =		
Model	6.71746248	5	1.3434925	F(5, 120)	=	7.44
Residual	21.6718034	120	.180598362	Prob > F	=	0.0000
-----				R-squared	=	0.2366
Total	28.3892659	125	.227114127	Adj R-squared	=	0.2048
-----				Root MSE	=	.42497
lfood	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.0600000					
yrsed	.0200000	.0100000	2.00	0.049	.0002000	.0398000
age	.0396647	.0183701	2.16	0.033	.0032932	.0760362
age2	-.0003862	.0001706	-2.26	0.025	-.000724	-.0000485
female	.0347106	.0794144	0.44	0.663	-.1225244	.1919455
_cons	2.327685	.517741	4.50	0.000	1.302594	3.352776

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

	income	yrsed	age	age2	female
income	0.0009000				
yrsed	-0.0000500	.00010000			
age	-1.838e-08	6.937e-06	.00033746		
age2	6.770e-10	6.832e-08	-3.097e-06	2.909e-08	
female	1.179e-06	.00005742	.00013342	-1.220e-06	.00630664

a) What is the effect of being female on food expenditure?

(8 marks)

this is a "semi-log" equation so the impact of being female relative to being male (net of differences in mean values of control variables) is equal to the % difference /100 in hourly pay of being female relative to being male, (since in the continuous variable case $d\ln w/d(x) = b_i = dw/dw(x) = \% \text{ change in } w /100 \text{ with respect to a unit change in } x$)

However since the coefficient is a dummy variable this is only an approximation to the proportionate

difference and the true effect is closer to $\exp(\hat{\beta}_{female}) - 1$

So other things equal the point estimate indicates that women earn $\exp(-.034) - 1 = .034 = 3.4\%$ less than men

However the t value indicates that this variables effect is statistically indistinguishable from zero so can only conclude that being female has no significant effect on food expenditure

b) Find the standard error on the estimate of income and hence test the hypothesis (at the 95% level) that income has no effect on food expenditure

(10 marks)

Since elements on the main diagonal of the variance/covariance matrix are the variances of the coefficient estimates then the standard error of income is simply

$$\sqrt{0.0009} = 0.03$$

$$\text{Hence using } \hat{t} = \frac{\hat{\beta} - \beta^0}{s.e.(\hat{\beta})} \text{ where } \beta^0 \text{ (the null hypothesis)} = 0$$

$$\text{Then } \hat{t} = \frac{0.06 - 0}{0.03} = 2$$

To find critical value of t distribution (at the 95% level) need to find degrees of freedom N-k

From regression output can see that N-k=120 (look at F test of goodness of fit of the model)

From t tables nearest critical value at the 95% level given N-k = 120 = 1.98 (not 1.96)

Hence **absolute** value of estimated t > $t_{critical}$ (2 > 1.98) so **reject null hypothesis that variable has no explanatory power**. Income does appear to have a (positive) effect on food expenditure

c) Find the 95% confidence interval in which the true effect of income lies.

(5 marks)

Since $\alpha=0.05$ (5%) and this is a 2-tailed test then the confidence interval is given by

$$\Pr \left[\hat{\beta}_1 - t_{\frac{0.05}{2}, N-k} * s.e.(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{0.05}{2}, N-k} * s.e.(\hat{\beta}_1) \right] = 0.95$$

so can be 95% confident that true value lies in range

$$0.06 - (1.98 * 0.03) \leq \beta \leq 0.06 + (1.98 * 0.03) \\ 0.0006 \leq \beta \leq 0.119$$

d) Test the hypothesis that the return to education (yrseid) and the return to income sum to 0.1 in the model above

(8 marks)

The test of value of the sum of income and education effects is given by a form of the F test

$$(\hat{\beta}_{income} + \hat{\beta}_{yrseid} - 0.1)^2 / \text{Var}(\hat{\beta}_{income} + \hat{\beta}_{yrseid}) \sim F(1, N-k)$$

the value of the denominator can be obtained from the variance/covariance matrix of the OLS estimates since the i^{th} element on the main diagonal is the variance on the parameter estimate of the i^{th} variable in the regression and the off diagonal terms are the covariances of the parameter estimates. The relevant covariance is highlighted (in green)

$$\text{Since } \text{Var}(\hat{\beta}_{income} + \hat{\beta}_{yrseid}) = \text{Var}(\hat{\beta}_{income}) + \text{Var}(\hat{\beta}_{yrseid}) + 2\text{Cov}(\hat{\beta}_{income}, \hat{\beta}_{yrseid})$$

$$\text{Then } \text{Var}(\hat{\beta}_{income} + \hat{\beta}_{yrseid}) = .0009 + .0001 + 2(-.00005) = .0009$$

$$\text{then } F = \frac{(.08 - 0.1)^2}{.0009} = .44 \quad \sim F[1, 120]$$

Since 95% critical value = 3.92 then $\hat{F} < F_{critical}$

So can **accept** null that coefficients sum to 0.1 (the relatively large standard errors on income and education ensure that the sampling variance is wide enough to incorporate the value 0.1)

e) Consider a simple model of 54 observations split equally into two sub-samples such that

$$\begin{aligned} y_i &= a_1 + b_1 X_i + u_i & i=1..N_1 & \quad \text{in sub-sample 1} \\ \text{and} & & & \\ y_i &= a_2 + b_2 X_i + u_i & i=N_1+1..N & \quad \text{in sub-sample 2} \end{aligned}$$

Suppose that the residual variance (s^2) from sub-sample 1 is 0.64, the residual variance (s^2) from sub-sample 2 is 0.16 and the residual variance (s^2) from a pooled regression of all 54 observations is 0.5

Test the hypothesis of no structural change across the two sub-samples at the 5% level. (9 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 1st sub-sample and X_2 is an N_2 by 2 matrix of observations from the 2nd 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

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)

Need to find RSS in each sample

Since $s^2 = \hat{u}' \hat{u} / (N-k)$

Then in sub-sample 1

$$\hat{u}' \hat{u} = s^2 * (N-k) = .64 * (27-2) = 16$$

in sub-sample 2

$$\hat{u}' \hat{u} = s^2 * (N-k) = .16 * (27-2) = 4$$

and in the (restricted) pooled sample

$$\hat{u}' \hat{u} = s^2 * (N-k) = .5 * (54-2) = 26$$

(note there are only 2 parameters in the restricted pooled sample and 4 in the unrestricted sample)

$$\text{hence } \hat{F} = \frac{(26 - (16 + 4)) / 2}{(16 + 4) / 54 - 2 * 2} = 7.5$$

From Tables the 5% critical value given the degrees of freedom $F^{0.05}[2, 50] = 3.2$

\hat{F}

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

e) Outline the form of a test that could be used to detect the presence of outliers in a data set

(10 marks)

DFITS test

- summarises contribution of both leverage h and residual

$$= r_i \sqrt{\frac{h_i}{1 - h_i}}$$

$$\text{where standardised residuals } r_i = \frac{\hat{u}_i}{s \sqrt{1 - h_i}}$$

s = standard error of regression equation = $\text{RSS} / (N-k)$

leverage $h_i = x_i (X'X)^{-1} x_i'$

where x_i is the i^{th} row of X

be the i^{th} element on the main diagonal of H . Said to be the "leverage" of the i^{th} observation

In 2 variable model can show

$$h_i \approx \frac{1}{N} + \frac{(x_i - \bar{X})^2}{\sum_{j=1}^N (x_j - \bar{X})^2}$$

$$0 \leq h_i \leq 1$$

(1 is high leverage)

(normalised by its standard error becomes scale invariant)

Can show that $DFITS > 2\sqrt{(k/N)}$ is worth investigating