Mid-Term Test No. 1 2005/06 – Answers

1. Given the general linear model y = XB + u where y is n_x1, B is n_x1 , X is n_xk and u ~ N(0, σ^2),

show that the OLS estimate of the coefficient vector B, $\hat{\beta}$ = $(X'X)^{-1}X'y$, is an unbiased estimator of B

(5 marks)

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

(5 marks)

c) Show that this variance is the smallest variance of any linear unbiased estimators

(15 marks)

The answers to all these questions have been covered in the lectures. See lecture notes for proofs.

2. The following output is taken from a regression of the log of hourly wages on years of education (edage), a quadratic in age (age , age^2) and dummy variables for gender (1 = female 0 = male) and part-time working (1=part-time 0 =full-time) Some of the output has been hidden.

```
. reg lnhw edage female part_time age age2

      Source |       SS       df       MS              Number of obs =    7471
-------------+------------------------------           F(  5,  7465) =  365.50
       Model |  577.441557     5  115.488311           Prob > F      =  0.0000
    Residual |  2358.73895                             R-squared     =
-------------+------------------------------           Adj R-squared =  0.1961
       Total |  2936.18051        .393062987           Root MSE      =  .56211

------------------------------------------------------------------------------
        lnhw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       edage |   .0622461   .0023077    26.97   0.000
      female |  -.1593637   .0143709            0.000    -.1875348   -.1311926
   part_time |  -.2524919   .0169652   -14.88   0.000    -.2857485   -.2192352
         age |   .0696136   .0041935    16.60   0.000     .0613931    .0778341
        age2 |  -.0007595   .0000534   -14.24   0.000    -.0008641    -.000655
       _cons |  -.2992242    .086984    -3.44   0.001    -.4697374    -.128711
------------------------------------------------------------------------------
```

What is the estimated effect of part-time working on hourly wages?

since this is a log-lin model the coefficients are semi-elasticities,

*$\delta LnW / \delta x_i = \hat{\beta}_i = (\delta W/W) / \delta x_i$ so % change in wage=($\hat{\beta}_i * \delta x_i$)*100*

however this is a dummy variable so need to transform by (exp($\hat{\beta}_i$)-1

$\hat{\beta}_i$ ***= -0.252 so the effect of part-time is exp(-0.252)-1 = -0.223*100 = 22.3%***

What is the effect of an extra year of age on hourly pay?

The effect of age is the combined effect of the 2 variables age and age^2, since

$$dLn(w)/dAge = b_{age}+2b_{age2}Age$$
$$= 0.069-2(0.00076)Age$$
$$= 0.069-0.0015Age$$

so the marginal effect of age on pay is not constant but declines with age (and will become negative after Age=0.069/0.0015 = 45)

What is the R^2 value ?

R^2= ESS/TSS = 577.4/2936.1 = 0.197

What is the t value on the female dummy variable and does the variable have zero explanatory power in the model?

$$\hat{t} = \frac{\hat{\beta}_i - \beta^{null}}{s.e.(\hat{\beta}_i)} \sim t_{(N-k)} = -.159/.014 \quad \sim t_{7471-6}$$

= -11.36

which since 95% critical value t_{7465} (2 tailed test) = 1.96 then absolute value of estimated t lies outside acceptance region. So reject null that female dummy has no explanatory power

What is the 95% confidence interval around the estimate of the edage variable?
95% confidence interval for each individual forecast observation given by

$$\Pr[\hat{\beta}_1 - t^{\alpha/2}_{n-k}.SE(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 - t^{\alpha/2}_{n-k}SE(\hat{\beta}_1)] = 1-.05$$

critical value for $t^{\alpha/2}_{n-k}$ = $t_{7471-6}{}^{.05/2}$ = 1.96

$$= \hat{\beta}_1 \pm t_{.05/2} SE(\hat{\beta}_1) \qquad = 0.062 \pm 1.96 * 0.002 \qquad = 0.062 \pm .004 = [.058,.066]$$

The OLS estimate of the residual variance s^2 ?

s^2 = RSS/N-k = 2358.7/(7471-6) = 0.316

Assuming that the female and part-time dummy variables are orthogonal, test the hypothesis that the true effect of these variables is the same in the model

H0: $\hat{\beta}_{female} = \hat{\beta}_{part-time} \equiv \hat{\beta}_{female} - \hat{\beta}_{part-time} = 0$

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

If variables orthogonal $X_1'X_2$ =0 and hence

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

So $Cov(\hat{\beta}_{female},\hat{\beta}_{part-time})=0$

Hence F = $(-.159--.252)^2$ / $(.014^2+.017^2)$

= *.0086/(0.000485) = 17.7 ~ F[1,7471-6]*

From F tables $F_{critical}^{5\%}$ *[1,7465] = 3.84*

$\hat{F}$ > $F_{critical}^{5\%}$ *so reject null that coefficients sum are equal.*

N.B. could use fact that t^2*= F and hence equivalent test is*

$$\frac{\left(\hat{\beta}_{female}-\hat{\beta}_{part-time}-0\right)}{S.E.(\hat{\beta}_{female}-\hat{\beta}_{part-time})}=\frac{\left(\hat{\beta}_{female}-\hat{\beta}_{part-time}\right)}{s.e.(\hat{\beta}_{female})+s.e.(\hat{\beta}_{part-time})}$$ = *2916* ~ t_{7465}

Show what the consequences of orthogonality are for the bias of OLS in the presence of omitted variables.

Partition X such that x = $[X_1 \mid X_2]$

True: $y = X_1B_1 + X_2B_2 + u$ *(1)*
Estimate: $y = X_1B_1 + u$ *(2)*

OLS on (2) gives

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

taking expectations

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

Only if X_1 and X_2 are orthogonal, $X_1'X_2=0$, estimates in omitted variable equation are unbiased

Tests: RESET test and LM Test
Former is test of higher order polynomials of existing variables. Latter is more general in that can test exclusion of different variables

3. Given a simple 3 variable model $y = b_0 + b_1X_1 + b_2X_2 + u$ estimated in mean deviation form

the summary statistics from an OLS regression are of the above model are

N=50 TSS = 100 ESS = 85.1 R^2 =0.851

The model is then estimated for 2 sub-groups, one a sample of 20 observations, the other a sample of 30 observations

Sub_sample 1

$X'X = \begin{bmatrix} 15 & 5 \\ 5 & 5 \end{bmatrix}$ $X'y = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ **y'y =5.8** **N =20**

Sub_sample 2

$X'X = \begin{bmatrix} 10 & 5 \\ 5 & 5 \end{bmatrix}$ $X'y = \begin{bmatrix} 10 \\ 5 \end{bmatrix}$ **y'y =15** **N=30**

Test the hypothesis that the coefficients are the same in both sub-samples

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

So $\hat{u}'_1\hat{u}_1 = 5.8 - [1 \quad 2]\begin{bmatrix} 15 & 5 \\ 5 & 5 \end{bmatrix}^{-1}\begin{bmatrix} 1 \\ 2 \end{bmatrix} = 5/8 - [1 \quad 2]\begin{bmatrix} 5/50 & -5/50 \\ -5/50 & 15/50 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix} = 5.8 - 0.9 = 4.9$

and $\hat{u}'_2\hat{u}_2 = 15 - [10 \quad 5]\begin{bmatrix} 10 & 5 \\ 5 & 5 \end{bmatrix}^{-1}\begin{bmatrix} 10 \\ 5 \end{bmatrix} = 15 - [10 \quad 5]\begin{bmatrix} 5/25 & -5/25 \\ -5/25 & 10/25 \end{bmatrix}\begin{bmatrix} 10 \\ 5 \end{bmatrix} = 15 - 10 = 5$

This gives unrestricted sum of squares (since equivalent to running OLS on the 2 sub-samples and allowing the intercept and the slope values to vary across the 2 sub-samples)

ie $RSS_{unrest} = \hat{u}_1'\hat{u}_1 + \hat{u}_2'\hat{u}_2 = 4.9 + 5 = 9.9$

Now restricted model implies imposing a common slopes on the two-sub-samples

ie $y = XB+u$

where $y=\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ $X=\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ $B=\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ $u=\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$

Could use fact that in restricted model $X'X = X_1'X_1 + X_2'X_2$ *and* $y'y=y_1'y_1 + y_2'y_2$
But
given $RSS_{restrict} = TSS - ESS = 100-85.1 = 14.9$

Then quicker to use $F = \frac{RSSrest - RSS_{unrest}/q}{RSS_{unrest}/n - k_{unrest}} = \frac{RSSrest - (RSS_1 + RSS_2)/q}{(RSS_1 + RSS_2)/n - k_{unrest}}$

$$F = \frac{14.9 - 9.9/2}{9.9/50-4} = \frac{5/2}{9/46} = 11.7 \sim F[2,46]$$

95% critical F value at F[2,46] = 3.23

so estimated $F > F_{critical}$

Hence reject null hypothesis that the intercept and slopes are the same in both sub-samples

Outline and comment on 2 tests that you could do to assess the specification of any estimated model

(16 marks)

RESET test
LM test

See lecture notes for details
Remember RESET test is a test that higher order powers of original variables in model are significant. LM test allows variables not included in original model to be tested.

Bonus Question (1 mark)

Which US state was the outlier in Computer Exercise 2?

Washington DC