UNIVERSITY OF LONDON

DEPARTMENTAL EXAMINATION 2011

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 **both** questions

Please answer each question on a separate page

College Calculators are provided
Statistical Tables are attached

1. Given the general linear model

$$y = X\beta + u$$

where y is an n_x1 vector of observations on the dependent variable, X is an n_xk matrix of observations on a set of explanatory variables, β is a k_x1 vector of parameters and u is an n_x1 vector of residuals

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

(8 marks)

Minimising the sum of squared residuals with respect to the unknown b parameters implies

Min$_\beta$ $\hat{u}'\hat{u} = \begin{bmatrix} \hat{u}_1 & \hat{u}_2 & .. & \hat{u}_n \end{bmatrix} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ : \\ \hat{u}_N \end{bmatrix} = \hat{u}_1^2 + \hat{u}_2^2 + ...\hat{u}_N^2$

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

Since all terms are scalars (1_x1) 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) Show that the mean of these OLS residuals is zero

(5 marks)

Re-write the OLS coefficient estimator as $X'X\hat{\beta} - X'y = 0$

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

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

It follows from the 1st row of this k by 1 vector (the row corresponding to the constant so that $x_1 = 1$) that

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

c) 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

$$\sum_{i=1}^{N} X_{ki} \hat{u}_i = \sum_{i=1}^{N} (X_{ki} - \bar{X}_k + \bar{X}_k) \hat{u}_i = \sum_{i=1}^{N} (x_{ki} + \bar{X}_k) \hat{u}_i = \sum_{i=1}^{N} x_{ki} \hat{u}_i + \sum_{i=1}^{N} \bar{X}_k \hat{u}_i$$

where x_{ki} is the variable x_k in mean deviation form

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

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

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

(6 marks)

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

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

$= E[(X'X)^{-1}X'uu'X(X'X)^{-1}]$

$= (X'X)^{-1}X'E(uu') X(X'X)^{-1}$

$= (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1}$

$\sigma^2(X'X)^{-1}$

e) 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

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

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

ie the N_x1 vector of observations on a single variable (x_1)
and
the N_x(k-1) matrix of observations on the other k-1 right-hand side variables (including the constant)

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

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

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

1^{st} row can be written

$$(x_1'x_1)^{-1}\hat{b}_1 + (x_1'X_2)\hat{\beta}_2 = x_1'y$$

so

$$\hat{b}_1 = (x_1'x_1)^{-1}x_1'y - (x_1'x_1)^{-1}(x_1'X_2)\hat{\beta}_2$$

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

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

(14 marks)

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

$= [I-X(X'X)^{-1}X']y$
$=My$

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

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

It follows that $Mu = [I-X(X'X)^{-1}X'][y-X\beta] \quad = y - X(X'X)^{-1}X'y - X\beta + X(X'X)^{-1}X'X\beta$
$=y - X\beta$
$= u$

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

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

Since $u'Mu$ is a scalar 1_x1

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

so $E(\hat{u}'\hat{u})=E(u'Mu)=E[tr(u'Mu)]=E[tr(Muu')]$
- using the result that $tr(ABC)=tr(BAC)=tr(BCA)=tr(CBA)$

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

so $E(\hat{u}'\hat{u})=tr[ME(uu')]$
$= \sigma^2 tr(M)$
$= \sigma^2[tr(I_N-[X(X'X)^{-1}X']]$
$= \sigma^2[tr(I_N)-tr[X(X'X)^{-1}X']]$
$= \sigma^2[tr(I_N)-tr[(X'X)^{-1}X'X]]$
$= \sigma^2[tr(I_N)-tr[I_K]]$

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

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

2. The following output is taken from 2 regressions of the share of the log of hourly pay (*lhw*) on
i) years of work experience (*exper)*, the square of work experience, (*exper2*) and dummy variables for being a graduate, (*grad*), having intermediate level qualifications (*inter*) and being a member of a trade union (*union*) (Model A)
ii) as Model A but without the union dummy variables (Model B)

Some of the regression output has been hidden

```
(A) reg lhw exper exper2 grad inter union

      Source |       SS       df       MS              Number of obs =
-------------+------------------------------           F(  5,   994) =   59.28
       Model |  84.2284197     5  16.8456839           Prob > F      =  0.0000
    Residual |  282.470998   994  .284176055           R-squared     =  0.2297
-------------+------------------------------           Adj R-squared =  0.2258
       Total |  366.699418        .367066485           Root MSE      =  .53308
------------------------------------------------------------------------------
         lhw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0407455   .0053528     7.61   0.000     .0302414    .0512497
      exper2 |  -.0006682   .0001158    -5.77   0.000    -.0008954   -.0004411
        grad |   .8000000   .0609305    13.26   0.000     .6880887    .9272229
       inter |   .3000000   .0400000     7.50   0.000     .2194622    .3983451
       union |   .2000000               5.00   0.000
       _cons |   1.066693   .0678323    15.73   0.000     .9335819    1.199804
```

```
(B) reg lhw exper exper2 grad inter

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  4,   995) =   64.93
       Model |  75.9008761     4   18.975219           Prob > F      =  0.0000
    Residual |  290.798542   995  .292259841           R-squared     =  0.2070
-------------+------------------------------           Adj R-squared =  0.2038
       Total |  366.699418   999  .367066485           Root MSE      =  .54061
------------------------------------------------------------------------------
         lhw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0464292    .005323     8.72   0.000     .0359836    .0568748
      exper2 |  -.0007707   .0001158    -6.65   0.000    -.0009979   -.0005434
        grad |   .8500000   .0612537    13.89   0.000     .7308568    .9712594
       inter |   .3239396   .0461365     7.02   0.000     .2334036    .4144756
       _cons |   1.050576    .068724    15.29   0.000      .915715    1.185436
```

The matrix of variances and covariances of the coefficient estimates from the model A, ($Var\left(\hat{\beta}\right)$ is given by

```
          exper      exper2        grad       inter       union
exper   .00002865
exper2  -5.947e-07  1.340e-08
grad    .00002869   9.143e-08   .00371252
inter  -3.581e-06   4.649e-07   .00174792   .00160000
union  -.00003984   7.179e-07               .00035000   .001600
```

a) What is the sample size in Model A?

(5 marks)

Since the degrees of freedom for the regression are N-k and K=6 (including the constant) then from the regression output it can be seen that the test of goodness of fit of the model has degrees of freedom N- k = 994

So if k=6 then N=1000

b) What is the estimated effect of being a graduate on the hourly pay in Model A?

(5 marks)

since the coefficient is a dummy variable the semi-log calculation is only an approximation to the proportionate difference and the true effect is closer to $\exp(\hat{\beta}_{grad}) - 1$

So other things equal women earn exp(.800)-1 = 1.23 = 123% more than the default of individuals with ***less than intermediate qualifications*** (there is an intermediate dummy in the model so all education estimates are relative to the default group)

c) What is the estimated effect of work experience on hourly pay?

(7 marks)

The coefficient is a semi-elasticity and gives the percentage change in hourly pay following a 1 year (unit) change in work experience (multiplied by 100)

(since work experience enters as a quadratic $d\log w_i/d(exper) = b_{exper} + 2 * b_{exper^2} * exper$

$= 0.041 + 2*-0.0007*exper$

So the effect is not constant but varies with the level of work experience

With 1 year of experience an extra year of work experience will raise hourly pay by
.041 – 0.0014 = .040

Ie 4%

With 10 years of experience the estimated effect is to raise hourly pay by
.041 – 0.007 = .034

Ie 3.4%

So a 1 % increase in income is associated with a 0.1 ***percentage point fall*** *(-10.0/100) 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)*

d) Find the standard error on the estimate of the union dummy variable.

(5 marks)

Using $\hat{t} = \dfrac{\hat{\beta} - \beta^0}{s.e.(\hat{\beta})}$ *where* β^0 *(the null hypothesis) = 0*

Then $5 = \dfrac{0.2 - 0}{s.e.\left(\hat{\beta}_{london}\right)}$ *so* $s.e.\left(\hat{\beta}_{london}\right) = \dfrac{0.2}{5.0} = 0.04$

(or could obtain form variance/covariance matrix of parameter estimates =$(0.0016)^{1/2}$

e) Hence calculate the 95% confidence interval around this estimate for union status

(5 marks)

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

$$\Pr\left[\hat{\beta}_{union} - t_{N-k}^{.05/2} * s.e.(\hat{\beta}_{lunion}) \le \beta_{union} \le \hat{\beta}_{union} + t_{N-k}^{.05/2} * s.e.(\hat{\beta}_{union})\right] = 0.95$$

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

From output (or from answer to section a) can see that N-k= 994

From t tables nearest critical value at the 95% level = 1.96)

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

$$0.2 - (1.96*(0.04)) <= \beta <= 0.2 + (1.96*(0.04))$$

$$0.12 <= \beta <= 0.28$$

f) Test the hypothesis that the coefficient on the union dummy variable is the same as that on the intermediate qualifications dummy variable in Model A

(7 marks)

The test of value of the equality of the union and intermediate effects is given by a form of the F test where RB=r is now $\beta_{inter} - \beta_{union} = 0$*. (since equality is the same as the difference in the coefficients being zero) Can show that in this case the test of linear restrictions becomes*

$$(\beta_{inter} - \beta_{union})^2 / Var(\hat{\beta}_{inter} - \hat{\beta}_{union}) \sim F(1, N-k)$$

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)*

Since $Var(\hat{\beta}_{inter} - \hat{\beta}_{union}) = Var(\hat{\beta}_{inter}) + Var(\hat{\beta}_{union}) - 2Cov(\hat{\beta}_{inter}\ \hat{\beta}_{union})$

Then Var($\hat{\beta}_{union}$ - $\hat{\beta}_{union}$) = *0.0016 + 0.0016 - 2(.00035) =.0025*

then $F = \frac{(.3-.2)^2}{.0025}$ *= 4* *~ F[1,994]*

Since 95% critical value = 3.84 then $\hat{F} > F_{critical}$

So can ***reject*** *null that coefficients are equal (the relatively small standard errors on union and inter ensure that the sampling variance is too small to incorporate the value of the other coefficient)*

g) Model B is a restricted version of Model A. Work out the implied OLS estimated effect from a regression of graduate status on the union dummy variable, net of all the other variables in the model.

(8 marks)

The answer relies on knowledge of omitted variable bias. Given the formula for omitted variable bias tells us that

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

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, β_2,
b) the covariance of X_1 and X_2

In this case X_2= Union

so

$$E(\hat{\beta}_{graduate}^{restricted\ model}) = \hat{\beta}_{graduate}^{unrestricted} + (X_1'X_1)^{-1}X_1'X_{union}\hat{\beta}_{union} \qquad (1)$$

The OLS estimate of graduate in the restricted model equals the (true) OLS coefficient on graduate in the unrestricted model plus a correction factor which is equal to the coefficient from a regression of union on graduate and all the other variables in the restricted model (X_1), $(X_1'X_1)^{-1}X_1'X_2$*, multiplied by the OLS coefficient on union in the full model*

Note that the restricted beta vector has 5 rows (1 for each coefficient) and the omitted variable bias formula applies to each row (ie each estimated coefficient) since $(X_1'X_1)^{-1}X_1'X_2$ *Is – in this case – 5 by 1*

Hence for the graduate dummy variable it must be that

$$0.850 = 0.800 + (X_1^{'}X_1)^{-1} X_1^{'}X_{union} * 0.200$$

so $\frac{0.85 - 0.8}{.20} = (X_1^{'}X_1)^{-1} X_1^{'}X_{union}$ = 0.25

Since the missing term can be interpreted as the OLS regression coefficient from a regression of age on log income net of all the other variables originally in the model, (the other variables in X1) this shows that graduate status and union status are positively correlated.

Interpreting the effect as a % difference, then graduates are 25 **percentage points** more likely to be union members than those with less than intermediate qualifications net of the other variables in the model

h) Show that OLS estimates remain unbiased in the presence of irrelevant variables in the model

(8 marks)