UNIVERSITY OF LONDON

DEPARTMENTAL EXAMINATION 2011

For Internal Students of
Royal Holloway

DO NOT TURN OVER UNTIL TOLD TO BEGIN

EC2203: ECONOMETRICS
EC4203: 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

The following output is taken from OLS regressions of three different models which try to establish the effect of output (measured in Kilograms) on total costs (measured in £). The first model regresses the level of costs, (*costs*), on the level of output, (*output)*. The second model regresses the natural log of costs, (*log_cost),* on the natural log of output, (*log_output)* and the third model regresses the level of costs on the level of output, the square of output, (*output_sq)* and the cube of output (*output_cub).*

Some of the regression output has been hidden.

Model 1

```
reg costs q
      Source |       SS       df       MS              Number of obs =
-------------+------------------------------           F(  1,    58) =  662.73
       Model |  733.336303     1  733.336303           Prob > F      =  0.0000
    Residual |  97.3749935    58  1.10653402           R-squared     =  0.8828
-------------+------------------------------           Adj R-squared =  0.8814
       Total |  830.711297    59  9.33383479           Root MSE      =  1.0519
------------------------------------------------------------------------------
       costs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           q |   .5000000   .0250000            0.000
       _cons |   .6501553   .1677777     3.88   0.000     .3167323    .9835782
------------------------------------------------------------------------------
```

Model 2

```
reg log_cost log_output

      Source |       SS       df       MS              Number of obs =
-------------+------------------------------           F(  1,    58) =  185.50
       Model |                 1                       Prob > F      =  0.0000
    Residual |  10.0000000    58  .113636360           R-squared     =
-------------+------------------------------           Adj R-squared =  0.9155
       Total |  100.000000    59  1.69491530           Root MSE      =  1.3019
------------------------------------------------------------------------------
    log_cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  log_output |   .6000000   .0272426    22.39   0.000     .5556884    .6639662
       _cons |  -2.447097   .1569509   -15.59   0.000    -2.759004    -2.13519
------------------------------------------------------------------------------
```

Model 3

```
reg costs output output_sq output_cub

      Source |       SS       df       MS              Number of obs =
-------------+------------------------------           F(   ,      ) =
       Model |  855.000000     3   285.00000           Prob > F      =  0.0000
    Residual |  95.0000000    56  1.69642860           R-squared     =  0.9000
-------------+------------------------------           Adj R-squared =  0.8806
       Total |  950.000000    59  16.1016950           Root MSE      =  4.0126
------------------------------------------------------------------------------
       costs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           q |   .8000000   .2000000     4.00   0.000
          q2 |  -0.003000   0.003000    -1.00   0.290    -9.07e-06    2.74e-06
          q3 |   0.000001   0.000001     0.95   0.347    -1.36e-09    3.83e-09
       _cons |   .4343922   .2503542     1.74   0.086    -.0632955    .9320799
------------------------------------------------------------------------------
```

continued over

a) Find the sample size in model 1

(5 marks)

- Since we know in a regression output that the regression degrees of freedom are calculated as N-k and the F test of goodness of fit is also distributed Ff[q, N-k]

where N is the sample size and k is the number of right hand side parameters and q Is the number of restrictions, then can see from the above output on the F test that

58 = N- k and since in this model k=2 then 58= N-2 so N=60

b) Calculate the R^2 value in model 2

(5 marks)

Need R^2 = 1-(RSS/TSS) = 1-(10/100) = 1-0.1 = 0.9

Equivalently R^2 =ESS/TSS where ESS = 100-10=90 and 90/100 = 0.9

c) Interpret the effect of the estimated effect of output on costs in **each** model

(10 marks)

Model 1 is a linear model in levels so estimated effect is $\frac{dCost}{dOutput} = \beta_{output}$

and the change in food expenditure is given by $dCost = \beta_{output} * dOutput$

*so a 1KG rise in output, increases costs by 0.5*1 = 50pence (and a 10KG increase raises costs by £5)*

- *3 marks*

Model 2 equation is log-linear so estimated effect is $\frac{dLog(Cost)}{d\log(Output)} = \beta_{\log(output)}$

which is the definition of an elasticity and the effect can be interpreted as the percentage change in costs wrt an x% change in output $d\log(Cost) = \beta_{\log(output)} * d\log(Output)$

*so a 1% rise in output, increases costs by 0.6*1 = 0.67%*

(Note in order to use this model both y and X variables should always be positive – can't take (natural) log of a negative number)

Note also that the implied slope of this predicted line of Y against X changes at very value of X, but the elasticity is constant at every value of X

- 3 marks

Model 3 is a polynomial (cubic) in output.

$$Cost = b_0 + b_q Q + b_{q2} Q^2 + b_{q3} Q^3 + u$$

This means that the effect of output on costs is non-linear & not constant (unlike a straight line where the slope is constant)

$\frac{dCost}{dOutput} = \beta_{output} + 2\beta_{output\_sq} Q + 3\beta_{output\_sq} Q^2$ *so effect varies with level of output*

However since both the quadratic and cubic terms are insignificantly different from zero then these terms do not contribute so the effect of output on costs is really only due to the level term =0.8

- 4 marks

d) Test the hypothesis that the variable *output* has some explanatory power in model 1
(use the 95% significance level for your test and the nearest critical value in the Table for the relevant degrees of freedom)

(5 marks)

Use $\hat{t} = \frac{\hat{\beta} - \beta^0}{s.e.(\hat{\beta})}$ *=(0.500-0)/0.250= 2*

From tables nearest critical value at the 95% level given N-k = 60-2 = 58 degrees of freedom (2-tailed test) ***is 2.01*** *(**not** 1.96)*

*Hence **absolute** value of estimated t <* $t_{critical}$ *(2<2.01) so **CANT reject** null hypothesis **that variable has no explanatory power.** Output appears to have (positive) effect on costs*

e) Find the 95% confidence interval for the true value of the coefficient on *output* in model 3

(5 marks)

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

$$\Pr\left[\hat{\beta}_1 - t^{.05/2}_{N-k} * s.e.(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t^{.05/2}_{N-k} * s.e.(\hat{\beta}_1)\right] = 0.95$$

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

*0.8– (2*0.2) <=β <= 0.8 + (2*0.2)*
0.402 <= 1.198

(note N-k = 56 in model 3 but nearest critical value in Tables is for N-k=60)

f) Calculate the F test of goodness of fit of the model as a whole in model 3

(5 marks)

F test for goodness of fit of model given by

$\frac{R^2 / k-1}{1-R^2 / N-k}$ *~ F[k-1,N-k]*

So $\hat{F} = \frac{0.9/4-1}{1-0.9/60-4}$ *~ F[3, 56]*

$\hat{F}$ *= 168 ~ F[3,56]*

From F tables $F_{critical}^{5\%\,[3,56]}$ *= 2.76*

$\hat{F} > F_{critical}^{5\%}$ *so reject null that model as a whole has zero explanatory power.*

g) Explain, briefly, how the adjusted R^2 helps with model selection. Use this to help you choose whether you prefer models 1, 2 or 3

(10 marks)

One problem with using the R^2 in a multiple regression is (can show) that the R^2 (and the ESS) will never fall when add regressors. (this is because OLS minimises the RSS so whenever a variable is dropped the RSS will always increase because the size of the residual increases)

- If so may be tempted to add many, many variables in order to increase the fit of the model.

- Problem (see notes on multicolinearity) that this will increase the chance of introducing correlation between rhs variables which will inflate the estimated standard errors so running the risk of type II error (failing to reject a false null) and make it difficult to assess the contribution of individual variables (if standard errors are all large)

*Useful therefore to also report the **adjusted R^2***

$$\bar{R}^2 = 1 - \frac{RSS/N-k}{TSS/N-1} = 1-(1-R^2)\frac{N-1}{N-k}$$

which contains an adjustment factor so that while RSS never ↑ (and usually falls) when new variables added there is a penalty to adding new variables because N-k ↓ (so moving in the opposite direction to the effect of adding more variables on RSS)

*Can show that **adjusted R^2** will only increase if the t value on the new variable > 1 (in absolute value)*

On this basis would prefer model 1 to model 3 since the adjusted R^2 is lower in model 3

Can **not** use this to compare models 1 & 2 since the dependent variables are different in the two models and so total sum of squares are different and not directly comparable. To do this you d have to do a box-cox test.

h) The regression output in Model 3 suggests the presence of what issue that arises in many multiple regression models? Give reasons for your answer.

(8 marks)

In the 3 variable model can show that

$$Var(\hat{\beta}_1) = \frac{s^2}{N * Var(X)} * \frac{1}{1-r^2_{X_1X_2}}$$

r^2_{x1x2} is the square of the correlation coefficient between X_1 & X_2

(compared with $Var(\hat{\beta}_1) = \frac{s^2}{N * Var(X)}$ in the 2 variable model)

So an increased correlation between X_1 & X_2 will make the OLS estimates of the effects of the X variables less precise (can't distinguish between the contribution of the individual variables if correlation is high)

*The high correlation is called **multicolinearity***

Detection:

1) *Low t values and high R^2*

i) How might you address this issue?

(9 marks)

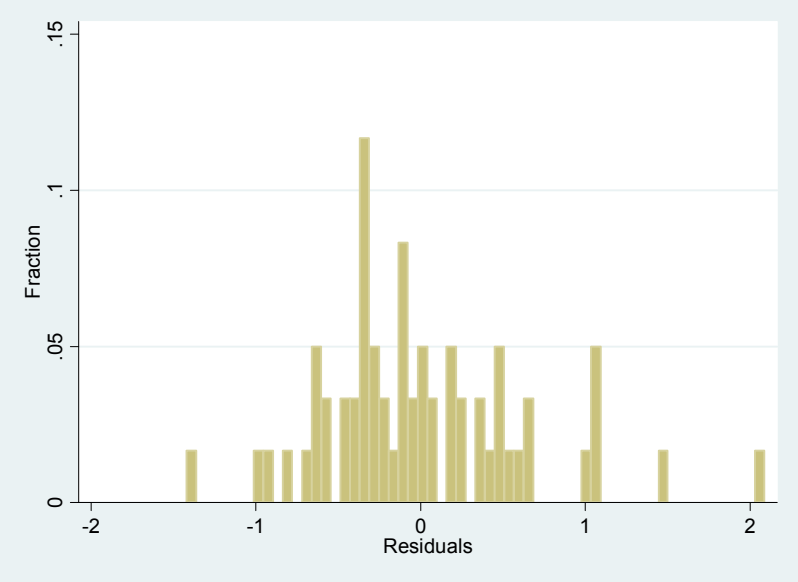
a) *increase sample size – since more information/observations increases precision other things equal*
b) *get more right hand side variables since this will reduce residual variance and again increase precision*
c) *drop one or more right hand side variables. Not enough information in the data set to distinguish individual effects of each variable*

j) The following output describes the distribution of OLS regression residuals estimated in model 1

```
su rhat, detail
```

```
                          Residuals
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -1.421913      -1.421913
 5%    -.8853922      -.9557221
10%    -.6503446      -.9424092       Obs
25%    -.3559197      -.8283751       Sum of Wgt.

50%    -.0958435                      Mean           -1.96e-09
                        Largest       Std. Dev.       .6224607
75%     .3609289       1.082467
90%     .8353969       1.086639       Variance        .3874573
95%     1.084553       1.474241       Skewness        1.000000
99%     2.087922       2.087922       Kurtosis        4.000000
```



j) Look at the output and outline with reasons – before doing a formal test - whether the residuals are likely to follow a normal distribution

(6 marks)

Histogram of residuals show signs of right skewness (residuals bunched to left, the median is less than the mean of zero, not centred on zero – so not symmetric). In output, Skewness value is above zero – which again indicates right skewness

and kurtosis appears to be leptokurtic – since peak of distribution higher than expected for a normal distribution where kurtosis is equal to 3

So conclude that residuals unlikely to be normally distributed

k) Why might we worry if the OLS residuals are not normally distributed?

(5 marks)

The assumption of normality is needed to derive the t test, confidence intervals and F tests need for hypothesis testing. If residuals not normal then t values may not follow a t distribution and so values and hence hypothesis testing and inference may be biased

l) Now do a formal test of normality in these residuals.

(7 marks)

To test more formally construct Jarque-Bera test

$$JB = N * \left[\frac{Skewness^2}{6} + \frac{(Kurtosis - 3)^2}{24} \right]$$

jb = (60((1^2)/6+(((1-3)^2)/24)) = 12.5*

*The statistic has a Chi^2 distribution with **2** degrees of freedom, (one for skewness one for kurtosis).*

From tables critical value at 5% level for 2 degrees of freedom is 5.99

*So JB>$\chi^2_{critical}$, so **reject** null that residuals are normally distributed.*

Suggests should try another functional form to try and make residuals normal, otherwise t stats may be invalid.

Outline the intuition that underlies the Chow forecast test of parameter stability

(7 marks)

If a model forecasts well out of sample then we would expect all the out-of-sample residuals to be close to zero. Intuitively if the model fits well the RSS from the combined regression should be close to that from the in-sample regression. A "large" difference suggest the RSS are different and so model does not forecast well)

Given a null hypothesis that the model is stable out of sample (predicts well) then if

$$\hat{F} > F_{critical}{}^{\alpha}[N_o, N-k]$$

***reject** null of model stability out-of-sample*

The following is the regression output produced when model 2 is estimated over an **additional 10** observations

```
      Source |       SS       df       MS              Number of obs =
-------------+------------------------------           F(  1,    68) =  279.18
       Model |  58.6012142     1  58.6012142           Prob > F      =  0.0000
    Residual |  18.6200000    68  .321034480           R-squared     =  0.8041
-------------+------------------------------           Adj R-squared =  0.8013
       Total |  72.8749761    69  1.05615907           Root MSE      =  .56656

------------------------------------------------------------------------------
    log_cost |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
  log_output |   .5689644   .0340523    16.71   0.000     .5010141    .6369147
       _cons |  -2.290508   .1805286   -12.69   0.000    -2.650747   -1.930269
-------------------------------------------------------------------------------
```

Do the formal test of forecast stability given the information in the regression output above
(10 marks)

It can be shown that the joint test of all the out-of-sample-residuals being close to zero is given by:

$$F = \frac{RSS_{in+out} - RSS_{in} / N_o}{RSS_{in} / N - k} \sim F[N_o, N - k]$$

where N_o *is the number of out-of-sample observations*
N is the number of in-sample observations
k is the number of RHS coefficients

is given by

$$F = \frac{(18.62000 - 10) / 10}{10 / 60 - 2} = 5$$

The 95%critical value at the relevant degrees of freedom is the same as before

. F(10,58)=1.99

so $\hat{F} > F_{critical}^{\alpha}[N_O, N - k]$ *so* ***reject null*** *that model predicts well out of sample*