1. The following is the output from a regression of the log of hourly wages on years of education,(yearsed), years of work experience (xper) and its square, (xpersq) on a sample of 6225 individuals.

```
  Source |       SS       df       MS              Number of obs =    6225
---------+------------------------------           F(  3,  6221) =  512.58
   Model |  457.732594     3  152.577531           Prob > F      =  0.0000
Residual |  1851.79026  6221  .297667619           R-squared     =  0.1982
---------+------------------------------           Adj R-squared =  0.1978
   Total |  2309.52285  6224  .371067296           Root MSE      =  .54559
---------------------------------------------------------------------------
  lnhpay |  Coef.      Std. Err.    t     P>|t|  [95% Conf. Interval]
---------+-----------------------------------------------------------------
 yearsed | .0740664   .0023269    31.830  0.000  .0695048    .0786279
    xper | .0160000   .002000      8.000  0.000  .0200000    .0120000
  xpersq |-.0005000   .000100      5.000  0.000  -.000700    -.000300
   _cons | .5000000   .0372179    16.552  0.000  .5430626    .6889826
---------------------------------------------------------------------------
```

Interpret the regression output

(20 marks)

Answer:

Some general discussion along the following lines.

4 variable linear regression. Log-lin model so coefficients are semi-elasticities. $dLn(hourpay)/dx_i$ = d Hourpay/Hourpay/ dx_I = % change in hourly pay given unit change in right hand side variable. So 1 extra year of education raises mean hourly pay by 7.4%, (since regression line passes through mean of dependent variable). Variable is statistically significantly different from zero, (t=31.8). Confidence interval in which can be 95% certain that true parameter lies goes from .069 to .079

Experience is entered as a quadratic so effect is non-linear

$$dLn(hourpay)/dx_i = b_{xper} + 2b_{xper2}*xper$$

Returns to experience are highest (1st order condition for maximum) when $dLn(hourpay)/dx_i = 0 = 0.016 - 2*.0005xper$

So that xper = 16 (16 years of work experience).

Both level and squared term are statistically significant from zero. R^2 suggests that 20% of total variation in hourly pay is explained by this variable (quite good in cross-section data)

F test is test of joint significance of all explanatory variables = 512.6 . P value (on next line below) indicates that this is statistically significantly different from critical value. Hence reject null that model has no explanatory power.

2. Say whether the following statements are ALWAYS TRUE, SOMETIMES TRUE or ALWAYS FALSE. Give a (short) justification for each answer.

a) The OLS estimator $\hat{b} = (X'X)^{-1}X'y$ is an unbiased estimator of the true parameters.

(10 marks)

b) Adding another variable to a regression will never reduce the adjusted R^2 value

(10 marks)

Answer:

a) SOMETIMES TRUE.

Unbiased if just two (not all) of the assumptions of General linear model are satisfied, E(u) = 0 E(X'u) = 0, since only these required in proof of unbiasedness of OLS estimator (see lecture notes).

b) FALSE.

Since $\bar{R}^2 = 1 - \frac{(N-1)}{N-k}(1-R^2)$, then it is possible that adjusted R^2 can fall if new variables are added, if addition to unadjusted R^2 is less than the loss in the degrees of freedom (N-k). For proof see Exercise 1 question 5.

3. Outline a test you might use to for the presence of outliers and leverage in your data.

(10 marks)

Answer:

Use DFITS test as in computer exercise No. 2. This is the product of the outlier effect and leverage embedded in any single observation.

DFITS = $r_i\sqrt{\frac{h_i}{1-h_i}}$ where r_i is the studentised residual (normalised by regression standard error times square root of one minus leverage), for the ith observation and h_i is the leverage for the ith observation (measure of dispersion from main mass of X values). Inspect variables where estimated DFIT >2*√(k/N).

(see lecture notes).

4. Given the model $y_t = b_0 + b_1x_{1t} + b_2x_{2t} + e_t$

is estimated in mean deviation form, $y_t = b_1x_{1t} + b_2x_{2t} + e_t$
(lower case letters in this example denote variables measured in mean deviation form)

with $(x'x) = \begin{bmatrix} 10 & 10 \\ 10 & 30 \end{bmatrix}$ $\quad x'y = \begin{bmatrix} 5 \\ -5 \end{bmatrix}$ $\quad y'y = 37.5$

using a sample of 63 observations, find the OLS estimates of b1 and b2.

Test the hypotheses that

i) $b_1 = 0$

ii) $b_1 + b_2 = 0$

(30 marks)

Answer: attached

5. Consider the multiple regression model y = XB + u

Suppose that just one of the X variables, X_1, was subject to a linear transformation $z=X_1\lambda$ where λ is a constant. Show the consequences for OLS estimation of the coefficients in the model.

(20 marks)

Answer: See exercise sheet 1, question 6.