Section A

1. The following output is taken from a regression of the log of output, logo, of a sample of firms on a dummy variable for manufacturing sector, (manuf), the number of employees at the firm, (nemploy) and the age of the firm in years, (agef).
   Some of the information from the output has been concealed.

```
      Source |Sum of Squares  df       MS                Number of obs =     104
-------------+----------------------------               F(   ,     ) =
       Model |   33.000        3   11.000                Prob > F      =
    Residual |  200.000      100    2.000                R-squared     =   0.248
-------------+----------------------------               Adj R-squared =
       Total |  233.000      103    2.262                Root MSE      =   1.414
------------------------------------------------------------------------------
        logo |      Coef.   Std. Err.      t     P>|t|      [95% Conf. Interval]
-------------+----------------------------------------------------------------
       manuf|  0.300
     nemploy|  0.030
        agef|  0.020
       _cons |   1.17        .326        3.58   0.002     .496        1.845
------------------------------------------------------------------------------
```

The inverse of the matrix of sum of squares/sum of cross-products of the right hand side variables is given by

```
            Manuf        nemploy            agef
 Manuf        .005
 nemploy    -.015        .020
 agef       -.033        .001               .040
```

Consider the variable manuf:

i) Interpret the meaning of the estimated coefficient

ii) find the standard error of the coefficient

iii) find its estimated t value (under the null hypothesis that the true coefficient is zero)

iv) find the 95% confidence interval around this estimate

(10 marks)

i) Since this is a semi-log model, estimate suggests that manufacturing firms produce, on average, (dLogo/dmanuf = b_{manuf} *= % change in 0 /100 when move from non-manufacturing to manufacturing firms*

since this is a dummy variable and a log-lin model should use the formula $\exp(b_{manuf})-1$ *= e(0.30)-1= 0.35 ie 35% more*

ii) since $\text{var}(\hat{\beta}) = s^2(X'X)^{-1}$ *and variance of each coefficient is given by* i^{th} *element on main diagonal multiplied by* s^2 *.*

Since s^2 *= RSS/N-k =2.00 (could calculate from info. in table or read directly from"MS" column if recognise in stata output)*

*Then var(manuf) = 2.00*0.005 =0.01*

and se(manuf) = sqrt(.01) = 0.1

Hence t value =(0.30)/0.1 = 3.00

iv) Hence 95% confidence interval =

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

Given n-k = 100, 5% critical value for 2 tailed test is 1.99 (NOT 1.96)

Hence $0.30 - (1.99*0.1) <= \beta_1 <= 0.30+(1.99*0.1)$
Ie 95% confident true manufacturing coefficient lies in range
$0.101 <= \beta_1 <= 0.499$

b) Test the hypothesis that the return to a year of job tenure is equal to the return to a year of education

(4 marks)

Use fact that $b_{nemploy} = b_{agef} \Rightarrow b_{nemploy} - b_{agef} = 0$
and this test of a single linear restriction reduces becomes

$$F = \frac{(b_{\text{nemploy}} - b_{\text{agef}})^2}{\text{Var}(b_{\text{nemploy}} - b_{\text{agef}})} \sim F[q, N-k] \equiv t = \frac{(b_{\text{nemploy}} - b_{\text{agef}})}{\sqrt{\text{Var}(b_{\text{nemploy}} - b_{\text{agef}})}} \sim t_{N-k}$$

(could do either and t is just square root of F for single restriction)

Follows that
$b_{nemploy} - b_{agef} = 0.030 - 0.02 = 0.010$
$Var(b_{nemploy}) = 0.1*0.020 = 0.002$ $Var(bagef) = 0.1*0.040 = 0.004$
$Cov(b_{tenure}, b_{yrsed}) = 0.001$

Hence $Var(b_{tenure} - b_{yrsed}) = Var(b_{tenure}) + Var(b_{yrsed}) - 2\, Cov(b_{tenure}, b_{yrsed})$
$= 0.002 + 0.004 - 2(0.001) = 0.004$

So $F = (0.010)^2 / 0.004 = .025$

From F tables, 5% critical value for F(1, 100) is 3.90
Hence estimated F < critical value so accept null hypothesis that coefficients (returns) are the same
(t test gives 0.158 < 1.99 critical value)

c) Outline – using matrix algebra if necessary – what the consequence of omitting relevant variables will be for the bias and variance of OLS estimates. (Assume that the variance of the unknown residuals, σ^2_u, is known).

(8 marks)

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

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

taking expectations

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

Unless X_1 *and* X_2 *are orthogonal,* $X_1'X_2=0$*, estimates in omitted variable equation are biased*

Using rules on partitioned matrices

$$\sigma_u^2(X_1'X_1)^{-1} = \sigma_u^2[X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1]^{-1} \tag{1}$$

compare with variance estimated in case of omitted variables

$$\sigma_u^2(X_1'X_1)^{-1} \tag{2}$$

so (1) < (2)

and omitted variable estimates have smaller variance
The more highly correlated X_1 *and* X_2 *the greater the difference*

d) Outline why and how you would test for the presence of outliers in your estimates (5 marks)

In small data sets possible that one observation can have large influence on regression estimates. If regression line passes close to that observation and it lies a long way from main body of data then observation said to have high leverage

$h_i = x_i(X'X)^{-}1x_i$ *where* x_i *is ith row of X*

(h measures distance form k dimensional means of X and implies predicted value is a weighted average of each element of the y vector where weights reflect contribution of y_j *to predicted value of* y_i

Detection: 1. Inspection *2. DFITS test =* $r_i\sqrt{h_i}/(1-h_i)$

Where r_i *= studentised residuals =* $\dfrac{\hat{u}_i}{s\sqrt{1-h_i}}$ *and s = RSS/N-k*

If DFIT > $2\sqrt{k/N}$ *observation worth investigating*

1. Given the following model, (in mean deviation form), you suspect that the variable, x_2, is measured with error

$$y = b_1x_1 + b_2x_2 + u \qquad (1)$$

a) What are the consequences for OLS estimation of (1) ?

(5 marks)

True: $y = X^t\beta+u$

Observe: $X=X^t + w$ *where u and w are iid*$\sim(0,\sigma^2_i)$ $i = w, u$

So observe $y =X\beta+ u - \beta w = X\beta+v$ *where* $v= u - \beta w$

So OLS gives $$\hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + v)$$

$$\hat{\beta} = \beta + (X'X)^{-1}v = \beta + (X'X)^{-1}(u - \beta w)$$

Consistency OF OLS depends on

$Plim(X'v/N) = 0$

Since $plim(X'u/N) = 0$ *by assumption and* $plim(X'X/N) = plim(X^{t\prime}X^t/N) + plim(w'w/N)$

$= \Sigma_x^t + \Omega$

ie both terms converge to non-zero values

Now $Plim(X'v/N) =$ $$p\lim[\frac{(X^t + w)'(u - w\beta)}{N}]$$

$$= p\lim[\frac{X^{t\prime}u}{N}] - \beta p\lim\left[\frac{(X^t w)}{N}\right] + p\lim\left[\frac{w'u}{N}\right] - \beta p\lim\left[\frac{w'w}{N}\right]$$

since x^t *uncorrelated with u and w by assumption and measurement error w uncorrelated with u then*

$Plim(X'v/N) =$ $$-\beta p\lim\left[\frac{w'w}{N}\right] = -\Omega\beta \neq \beta$$

So $$p\lim\hat{\beta} = \beta - [\Sigma_{xt} + \Omega]^{-1}\Omega\beta$$

(attenuation bias, OLS estimates closer to zero than true values)

Given a partitioned matrix W = [y: x_1 : x_2 : x_3] where x_1 and x_3 are exogenous

$$W'W = \begin{bmatrix} 217 & 3 & 4 & 6 \\ 3 & 3 & 1 & 2 \\ 4 & 1 & 5 & 2 \\ 6 & 2 & 2 & 2 \end{bmatrix}$$

the sample size is 100

b) Find the IV estimates of the coefficients on x_1 and x_2

(6 marks)

IV estimates given by $(z'x)^{-1}z'y$ where $x= [x_1 : x_2]$ and $z = [x_1 : x_3]$

$$z'x = [x_1 \quad x_3]'[x_1 \quad x_2] = \begin{bmatrix} x_1' \\ x_3' \end{bmatrix}[x_1 \quad x_2] = \begin{bmatrix} x_1'x_1 & x_1'x_2 \\ x_3'x_1 & x_3'x_2 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}$$ *and*

$$z'y = \begin{bmatrix} x_1'y \\ x_3'y \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$$

so $$\hat{\beta}_{IV} = \begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}^{-1}\begin{bmatrix} 3 \\ 6 \end{bmatrix} = \begin{bmatrix} 2/4 & -2/4 \\ -1/4 & 3/4 \end{bmatrix}\begin{bmatrix} 3 \\ 6 \end{bmatrix} = \begin{bmatrix} 1/2 & -1/2 \\ -1/4 & 3/4 \end{bmatrix}\begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 3/4 \end{bmatrix}$$

c) Find the variance of the IV residuals

(4 marks)

Need $s^2_{IV} = u_{IV}'\, u_{IV} / n = (y - xb_{IV})'\,(y - xb_{IV}) / n = (y'y - 2\, b_{IV}'X'y + b_{IV}'x'x\, b_{IV}) / n$

From answer to above and using $x'x = \begin{bmatrix} x1'x_1 & x_1'x_2 \\ x_2'x_1 & x_2'x_2 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}$ *and* $x'y = \begin{bmatrix} x_1'y \\ x_2'y \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$

$$s^2_{IV} = \left(217 - 2[1/23/4]\begin{bmatrix} 3 \\ 4 \end{bmatrix} + [1/2 \quad 3/4]\begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}\begin{bmatrix} 1 \\ 6 \end{bmatrix}\right) \Big/ 100 = (217 - 9 + 27)10/100 = 2.36$$

d) Hence find the standard error of the IV estimates and comment on your findings

(5 marks)

So $Var(b_{IV}) = s^2_{IV}\ (Z'X)^{-1}(Z'Z)(Z'X)^{-1}$

$$Var(\hat{\beta}_{IV}) = 2.36\left(\begin{bmatrix} 1 & -1 \\ -1 & 3/2 \end{bmatrix}\begin{bmatrix} 3 & 2 \\ 2 & 5 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ -1 & 3/2 \end{bmatrix}\right) = 2.36\begin{bmatrix} 4 & -11/2 \\ -11/2 & 33/4 \end{bmatrix} = \begin{bmatrix} 9.44 & -12.98 \\ -12.98 & 19.47 \end{bmatrix}$$

So standard errors are square roots of elements on main diagonal $s.e.(\hat{\beta}_{IV}) = \begin{bmatrix} 3.07 \\ 4.41 \end{bmatrix}$

Hence only IV estimate of mismeasured variable is statistically significant at 5% level.

e) Outline a possible test for the validity of the instrument
(4 marks)

If an instrument Z is valid (exogenous) it is uncorrelated with u

To test this simply regress u on ***all*** *the possible instruments.*

$$u = d_0 + d_1Z_1 + d_2Z_2 + \ldots\ d_lZ_l + v$$

If the instruments are exogenous they should be uncorrelated with u and so the coefficients d_1 .. d_l should all be zero (ie the Z variables have no explanatory power)

Since u is never observed have to use a proxy for this which turns out to be the residual from the 2SLS estimation estimated using all the possible instruments

$$\hat{u}^{2sls} = y - \hat{b}_0^{2sls} - \hat{b}_1^{2sls} X$$

(since this is a consistent estimate of the true unknown residuals)

So to Test Overidentifying Restrictions

1. *Estimate model by 2SLS and save the residuals*
2. *Regress these residuals on all the exogenous variables (including those X variables in the original equation that are not suspect)*

$$\hat{u}^{2sls} = d_0 + b_1X_1 + d_1Z_1 + d_2Z_2 + \ldots\ d_lZ_l + v$$

and save the R^2

3. *Compute N^*R^2*

4. *Under the null that all the instruments are uncorrelated then*
   $N^*R^2 \sim \chi^2$ with L-k degrees of freedom

   (L is the number of instruments and k is the number of endogenous right hand side variables in the original equation)

Note that can only do this test if there are more instruments than endogenous right hand side variables (in just identified case the residuals and right hand side variables are uncorrelated by construction)

Given a model of the form

$$Y_{it} = X_{it}\beta + a_i + u_{it} \qquad i = 1, 2\ldots.N \quad t = 1, 2\ldots T$$

where a_i is unobserved

Outline and explain the 2 most popular estimation methods for dealing with models of this type. Give details of the assumptions that underlie these 2 approaches

(10 marks)

discussion of fixed and random effects methods

Fixed effects treats a_i as individual specific constants so effectively become unknown parameters that are $y = X_{it}\beta + \delta a_i + u_{it}$
Rather than estimate N parameters can either 1st difference or use within-group estimation to

Random effects treats a+i+ as individual specific error components so that estimate random effects by GLS, $\hat{\beta}_{RE} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$, where

$$\Omega = \begin{bmatrix} \Sigma & & & & \\ & \Sigma & 0 & & \\ & & \Sigma & & \\ & 0 & & .. & \\ & & & & \Sigma \end{bmatrix}$$

ie a block diagonal matrix, where the Σ are the residual covariance matrices for each of the I individuals in the data set

We need to find $\Sigma^{-1/2}$, since transformation required for GLS estimation is to multiply each y and X observation by square root of Σ^{-1}

Given = $\Sigma = \sigma^2_e I_T + \sigma^2_a i\, i'$ (see lecture notes)

Follows that $\Sigma^{-1/2} = \frac{1}{\sigma_e}\left[I - \frac{\theta}{T} ii'\right]$ where $\theta = 1 - \frac{\sigma_e}{\sqrt{T\sigma_a^2 - \sigma_e^2}}$

ie $\Sigma^{-1/2} y_i = \frac{1}{\sigma_e}\begin{bmatrix} y_{i1} - \theta \bar{y}_i \\ y_{i2} - \theta \bar{y}_i \\ \vdots \\ y_{iT} - \theta \bar{y}_i \end{bmatrix}$

Now this is similar (but <u>not</u> identical) to within groups estimation when θ = 1

This (ie θ = 1) would mean that $\sigma^2_e = 0$, so any variation across individuals is caused by the different a_i ‘s once the X variables are accounted for. In this case fixed and random effects models would be indistinguishable.

b) Hausman test

Under null Cov(X_i a_i) = 0 then random effects consistent and efficient. If not inconsistent

$$H=(\hat{\beta}_{re}-\hat{\beta}_{fe})'[Var(\hat{\beta}_{re})-Var(\hat{\beta}_{fe})]^{-1}(\hat{\beta}_{re}-\hat{\beta}_{fe})\sim\chi^2(k)$$

Or equivalently
Include within-group terms to random effects regression and F test for joint significance. Under null should have no effect

c) Measurement error in levels implies

$$\hat{\beta}=\beta\left[1-\frac{\sigma^2_w}{\sigma^2_{xt}+\sigma^2_w}\right]=\beta\left[\frac{\sigma^2_{xt}}{\sigma^2_{xt}+\sigma^2_w}\right]$$

In 1st diffs

$$\hat{\beta}=\beta\left[1-\frac{\sigma^2_{\Delta w}}{\sigma^2_{\Delta xt}+\sigma^2_{\Delta w}}\right]=\beta\left[\frac{\sigma^2_{\Delta xt}}{\sigma^2_{\Delta xt}+\sigma^2_{\Delta w}}\right]$$

where

$$\sigma^2_{\Delta w}=\mathrm{var}(w_t-w_{t-1})=2\sigma^2_w(1-\rho_w)$$

and $\sigma^2_{\Delta xt}=\mathrm{var}(xt_t-xt_{t-1})=2\sigma^2_{xt}(1-\rho_{xt})$

so

$$\hat{\beta}==\beta\left[\frac{2\sigma^2_{xt}(1-\rho_{xt})}{2\sigma^2_{xt}(1-\rho_{xt})+2\sigma^2_w(1-\rho_w)}\right]=\beta\left[\frac{\sigma^2_{xt}}{\sigma^2_{xt}+\sigma^2_w(1-\rho_w)\big/(1-\rho_{xt})}\right]$$

If $\rho_{xt} > \rho_w$ then reliability ratio is smaller in 1st differences than in levels. Attenuation bias is larger in 1st differences