R and Stata Workshop: Using Stata

Economic and Finance Department
Brunel University London, UK

Michael Willox

2024-06-09

Setup

Use *mus08psidextract.dta* to set up panel data in Stata, and define panel variable *id*, and time variable *t*. Consider the following model:

$$lwage_{it} = \alpha + \beta_1 exp_{it} + \beta_2 exp2_{it} + \beta_3 wks_{it} + \beta_4 ed_{it} + \mu_i + \epsilon_{it}$$

Summarize and describe the dataset.

```
use "../Data/mus08psidextract.dta", clear
summarize
describe
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         exp |      4,165    19.85378    10.96637          1         51
         wks |      4,165    46.81152    5.129098          5         52
         occ |      4,165    .5111645    .4999354          0          1
         ind |      4,165    .3954382    .4890033          0          1
       south |      4,165    .2902761    .4539442          0          1
-------------+---------------------------------------------------------
        smsa |      4,165    .6537815     .475821          0          1
          ms |      4,165    .8144058    .3888256          0          1
         fem |      4,165     .112605    .3161473          0          1
       union |      4,165    .3639856    .4812023          0          1
          ed |      4,165    12.84538    2.787995          4         17
-------------+---------------------------------------------------------
         blk |      4,165    .0722689    .2589637          0          1
       lwage |      4,165    6.676346    .4615122    4.60517      8.537
          id |      4,165         298    171.7821          1        595
           t |      4,165           4     2.00024          1          7
       tdum1 |      4,165    .1428571    .3499691          0          1
-------------+---------------------------------------------------------
       tdum2 |      4,165    .1428571    .3499691          0          1
       tdum3 |      4,165    .1428571    .3499691          0          1
       tdum4 |      4,165    .1428571    .3499691          0          1
```

```
       tdum5 |      4,165    .1428571    .3499691          0          1
       tdum6 |      4,165    .1428571    .3499691          0          1
-------------+---------------------------------------------------------
       tdum7 |      4,165    .1428571    .3499691          0          1
        exp2 |      4,165     514.405    496.9962          1       2601
```

```
Contains data from ../Data/mus08psidextract.dta
 Observations:         4,165                  PSID wage data 1976-82 from
                                                Baltagi and Khanti-Akom (1990)
    Variables:            22                  26 Nov 2008 17:15
                                              (_dta has notes)
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
exp             float   %9.0g                 years of full-time work
                                                experience
wks             float   %9.0g                 weeks worked
occ             float   %9.0g                 occupation; occ==1 if in a
                                                blue-collar occupation
ind             float   %9.0g                 industry; ind==1 if working in a
                                                manufacturing industry
south           float   %9.0g                 residence; south==1 if in the
                                                South area
smsa            float   %9.0g                 smsa==1 if in the Standard
                                                metropolitan statistical area
ms              float   %9.0g                 marital status
fem             float   %9.0g                 female or male
union           float   %9.0g                 if wage set be a union contract
ed              float   %9.0g                 years of education
blk             float   %9.0g                 black
lwage           float   %9.0g                 log wage
id              float   %9.0g
t               float   %9.0g
tdum1           byte    %8.0g                 t== 1.0000
tdum2           byte    %8.0g                 t== 2.0000
tdum3           byte    %8.0g                 t== 3.0000
tdum4           byte    %8.0g                 t== 4.0000
tdum5           byte    %8.0g                 t== 5.0000
tdum6           byte    %8.0g                 t== 6.0000
tdum7           byte    %8.0g                 t== 7.0000
exp2            float   %9.0g
-------------------------------------------------------------------------------
Sorted by: id  t
```

Questions 1. Determine if this panel data is the short or long panel.

Given that the intercept term in equation 1 of the assignment is common to all units, it suggests that the model being considered is characterized by random effects, which assumes the term u_i is not correlated with the regressors, X_{it}.

Use the distinct command to count distinct values for the time and panel variables. The dataset is short because $n = 595 > T = 7$. The total number of observations, N, is 4165.

```
use "../Data/mus08psidextract.dta", clear
distinct id t
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))
```

```
--------------------------
    |     total   distinct
----+---------------------
 id |      4165        595
  t |      4165          7
--------------------------
```

Questions 2. Run pooled OLS, fixed effects and random effects regressions?

Pooled OLS with a single intercept.

```
use "../Data/mus08psidextract.dta", clear
reg lwage exp exp2 wks ed
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))

      Source |       SS           df       MS      Number of obs   =     4,165
-------------+----------------------------------   F(4, 4160)      =    411.62
       Model |  251.491445         4  62.8728613   Prob > F        =    0.0000
    Residual |  635.413457     4,160  .152743619   R-squared       =    0.2836
-------------+----------------------------------   Adj R-squared   =    0.2829
       Total |  886.904902     4,164  .212993492   Root MSE        =    .39082

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         exp |    .044675   .0023929    18.67   0.000     .0399838    .0493663
        exp2 |  -.0007156   .0000528   -13.56   0.000    -.0008191   -.0006121
         wks |    .005827   .0011827     4.93   0.000     .0035084    .0081456
          ed |   .0760407   .0022266    34.15   0.000     .0716754     .080406
       _cons |   4.907961   .0673297    72.89   0.000     4.775959    5.039963
------------------------------------------------------------------------------
```

Fixed effects with unit specific intercepts and unit-specific, time-invariant error term that is uncorrelated with the explanatory variables. u_i is assumed to be correlated with the regressors, X_{it}.

```
use "../Data/mus08psidextract.dta", clear
xtreg lwage exp exp2 wks ed, fe
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))

note: ed omitted because of collinearity.

Fixed-effects (within) regression               Number of obs     =      4,165
Group variable: id                              Number of groups  =        595
```

```
R-squared:                                      Obs per group:
     Within  = 0.6566                                         min =          7
     Between = 0.0276                                         avg =        7.0
     Overall = 0.0476                                         max =          7

                                                F(3,3567)         =    2273.74
corr(u_i, Xb) = -0.9107                         Prob > F          =     0.0000

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         exp |   .1137879   .0024689    46.09   0.000     .1089473    .1186284
        exp2 |  -.0004244   .0000546    -7.77   0.000    -.0005315   -.0003173
         wks |   .0008359   .0005997     1.39   0.163    -.0003399    .0020116
          ed |          0  (omitted)
       _cons |   4.596396   .0389061   118.14   0.000     4.520116    4.672677
-------------+----------------------------------------------------------------
     sigma_u |  1.0362039
     sigma_e |  .15220316
         rho |  .97888036   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(594, 3567) = 53.12                  Prob > F = 0.0000
```

Random effects with single intercept and a unit specific error term (random effect), which is uncorrelated with the explanatory variables, varies across units, is constant over time for each unit and is separate from the idiosyncratic error term.

```
use "../Data/mus08psidextract.dta", clear
xtreg lwage exp exp2 wks ed, re
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))


Random-effects GLS regression                   Number of obs     =      4,165
Group variable: id                              Number of groups  =        595

R-squared:                                      Obs per group:
     Within  = 0.6340                                         min =          7
     Between = 0.1716                                         avg =        7.0
     Overall = 0.1830                                         max =          7

                                                Wald chi2(4)      =    3012.45
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         exp |   .0888609   .0028178    31.54   0.000     .0833382    .0943837
        exp2 |  -.0007726   .0000623   -12.41   0.000    -.0008946   -.0006505
         wks |   .0009658   .0007433     1.30   0.194     -.000491    .0024226
          ed |   .1117099   .0060572    18.44   0.000     .0998381    .1235818
       _cons |   3.829366   .0936336    40.90   0.000     3.645848    4.012885
```

```
-------------+----------------------------------------------------------------
     sigma_u |  .31951859
     sigma_e |  .15220316
         rho |  .81505521   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Questions 3. Does this model have multicollinearity or heteroscedasticity?

Pooled OLS exhibits multicollinearity.

```
use "../Data/mus08psidextract.dta", clear

reg lwage exp exp2 wks ed
estat vif
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))

      Source |       SS           df       MS      Number of obs   =     4,165
-------------+----------------------------------   F(4, 4160)      =    411.62
       Model |  251.491445         4  62.8728613   Prob > F        =    0.0000
    Residual |  635.413457     4,160  .152743619   R-squared       =    0.2836
-------------+----------------------------------   Adj R-squared   =    0.2829
       Total |  886.904902     4,164  .212993492   Root MSE        =    .39082

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         exp |    .044675   .0023929    18.67   0.000     .0399838    .0493663
        exp2 |  -.0007156   .0000528   -13.56   0.000    -.0008191   -.0006121
         wks |    .005827   .0011827     4.93   0.000     .0035084    .0081456
          ed |   .0760407   .0022266    34.15   0.000     .0716754     .080406
       _cons |   4.907961   .0673297    72.89   0.000     4.775959    5.039963
------------------------------------------------------------------------------

    Variable |       VIF       1/VIF
-------------+----------------------
         exp |     18.77    0.053271
        exp2 |     18.77    0.053282
          ed |      1.05    0.951887
         wks |      1.00    0.996914
-------------+----------------------
    Mean VIF |      9.90
```

The *VIF* measures how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. A *VIF* value greater than 10 is often considered indicative of high multicollinearity, which can affect the stability and interpretation of the regression coefficients.

Although the mean *VIF* of 9.90 is just below 10, the individual *VIF* values for exp and exp^2 are of more concern. Centering variables can help reduce multicollinearity. This involves subtracting the mean of a variable from each of its values and then using this centered variable in the regression. Note that $estatvif$ does not work with $xtreg$ combined with the fe or re options.

Here, the bar over the variable represents the centred or demeaned variable.

$$\overline{lwage_{it}} = \alpha + \beta_1 \overline{exp_{it}} + \beta_2 \overline{exp_{it}^2} + \beta_3 \overline{wks_{it}} + \beta_4 \overline{ed_{it}} + \mu_i + \epsilon_{it}$$

```
use "../Data/mus08psidextract.dta", clear

* Generate de-meaned variables
bys id: egen mean_lwage = mean(lwage)
bys id: egen mean_exp = mean(exp)
bys id: egen mean_exp2 = mean(exp2)
bys id: egen mean_wks = mean(wks)
bys id: egen mean_ed = mean(ed)

gen lwage_dm = lwage - mean_lwage
gen exp_dm = exp - mean_exp
gen exp2_dm = exp2 - mean_exp2
gen wks_dm = wks - mean_wks
gen ed_dm = ed - mean_ed

reg lwage_dm exp_dm exp2_dm wks_dm ed_dm

* Calculate VIFs
estat vif
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))
```

```
note: ed_dm omitted because of collinearity.

      Source |       SS           df       MS      Number of obs   =     4,165
-------------+----------------------------------   F(3, 4161)      =   2652.37
       Model |  158.018789         3  52.6729298   Prob > F        =    0.0000
    Residual |  82.6324168     4,161  .019858788   R-squared       =    0.6566
-------------+----------------------------------   Adj R-squared   =    0.6564
       Total |  240.651206     4,164  .057793277   Root MSE        =    .14092

------------------------------------------------------------------------------
    lwage_dm | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      exp_dm |   .1137879   .0022859    49.78   0.000     .1093063    .1182694
     exp2_dm |  -.0004244   .0000506    -8.39   0.000    -.0005235   -.0003252
      wks_dm |   .0008359   .0005552     1.51   0.132    -.0002527    .0019244
       ed_dm |          0  (omitted)
       _cons |  -7.66e-09   .0021836    -0.00   1.000     -.004281     .004281
------------------------------------------------------------------------------
```

```
    Variable |       VIF       1/VIF  
-------------+----------------------
     exp2_dm |      4.39    0.227862
      exp_dm |      4.38    0.228124
      wks_dm |      1.00    0.995637
-------------+----------------------
    Mean VIF |      3.26
```

```
use "../Data/mus08psidextract.dta", clear

reg lwage exp exp2 wks ed
estat hettest, iid
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))

      Source |       SS           df       MS      Number of obs   =     4,165
-------------+----------------------------------   F(4, 4160)      =    411.62
       Model |  251.491445         4  62.8728613   Prob > F        =    0.0000
    Residual |  635.413457     4,160  .152743619   R-squared       =    0.2836
-------------+----------------------------------   Adj R-squared   =    0.2829
       Total |  886.904902     4,164  .212993492   Root MSE        =    .39082

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         exp |    .044675   .0023929    18.67   0.000     .0399838    .0493663
        exp2 |  -.0007156   .0000528   -13.56   0.000    -.0008191   -.0006121
         wks |    .005827   .0011827     4.93   0.000     .0035084    .0081456
          ed |   .0760407   .0022266    34.15   0.000     .0716754     .080406
       _cons |   4.907961   .0673297    72.89   0.000     4.775959    5.039963
------------------------------------------------------------------------------


Breusch-Pagan/Cook-Weisberg test for heteroskedasticity 
Assumption: i.i.d. error terms
Variable: Fitted values of lwage

H0: Constant variance

    chi2(1) =   0.00
Prob > chi2 = 0.9763
```

The result, *Prob > chi2 = 0.9763* indicates that the null hypothesis that the residuals are homoscedastic and cannot be rejected at standard levels of statistical significance.

Questions 4. Which method is suitable for this model, pooled OLS regression or a random effects model?

```
use "../Data/mus08psidextract.dta", clear

reg lwage exp exp2 wks ed
```

```
estimates store OLS

reg lwage exp exp2 wks ed, vce(cluster id)
estimates store OLSR

xtreg lwage exp exp2 wks ed, fe
estimates store FE

xtreg lwage exp exp2 wks ed, re
estimates store RE

estimate table OLS OLSR FE RE, star(.05 .01 .001) b(%7.2f)
estimate table OLS OLSR FE RE, b(%7.2f) se(%7.2f) p(%7.2f)  stats(N r2_a)
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))

      Source |       SS           df       MS      Number of obs   =     4,165
-------------+----------------------------------   F(4, 4160)      =    411.62
       Model |  251.491445         4  62.8728613   Prob > F        =    0.0000
    Residual |  635.413457     4,160  .152743619   R-squared       =    0.2836
-------------+----------------------------------   Adj R-squared   =    0.2829
       Total |  886.904902     4,164  .212993492   Root MSE        =    .39082

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         exp |    .044675   .0023929    18.67   0.000     .0399838    .0493663
        exp2 |  -.0007156   .0000528   -13.56   0.000    -.0008191   -.0006121
         wks |    .005827   .0011827     4.93   0.000     .0035084    .0081456
          ed |   .0760407   .0022266    34.15   0.000     .0716754     .080406
       _cons |   4.907961   .0673297    72.89   0.000     4.775959    5.039963
------------------------------------------------------------------------------


Linear regression                               Number of obs     =      4,165
                                                F(4, 594)         =      72.58
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2836
                                                Root MSE          =     .39082

                                  (Std. err. adjusted for 595 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         exp |    .044675   .0054385     8.21   0.000     .0339941     .055356
        exp2 |  -.0007156   .0001285    -5.57   0.000    -.0009679   -.0004633
         wks |    .005827   .0019284     3.02   0.003     .0020396    .0096144
          ed |   .0760407   .0052122    14.59   0.000     .0658042    .0862772
       _cons |   4.907961   .1399887    35.06   0.000     4.633028    5.182894
------------------------------------------------------------------------------
```

```
note: ed omitted because of collinearity.

Fixed-effects (within) regression               Number of obs     =      4,165
Group variable: id                              Number of groups  =        595

R-squared:                                      Obs per group:
     Within  = 0.6566                                         min =          7
     Between = 0.0276                                         avg =        7.0
     Overall = 0.0476                                         max =          7

                                                F(3,3567)         =    2273.74
corr(u_i, Xb) = -0.9107                         Prob > F          =     0.0000

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         exp |   .1137879   .0024689    46.09   0.000     .1089473    .1186284
        exp2 |  -.0004244   .0000546    -7.77   0.000    -.0005315   -.0003173
         wks |   .0008359   .0005997     1.39   0.163    -.0003399    .0020116
          ed |          0  (omitted)
       _cons |   4.596396   .0389061   118.14   0.000     4.520116    4.672677
-------------+----------------------------------------------------------------
     sigma_u |  1.0362039
     sigma_e |  .15220316
         rho |  .97888036   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(594, 3567) = 53.12                  Prob > F = 0.0000


Random-effects GLS regression                   Number of obs     =      4,165
Group variable: id                              Number of groups  =        595

R-squared:                                      Obs per group:
     Within  = 0.6340                                         min =          7
     Between = 0.1716                                         avg =        7.0
     Overall = 0.1830                                         max =          7

                                                Wald chi2(4)      =    3012.45
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         exp |   .0888609   .0028178    31.54   0.000     .0833382    .0943837
        exp2 |  -.0007726   .0000623   -12.41   0.000    -.0008946   -.0006505
         wks |   .0009658   .0007433     1.30   0.194     -.000491    .0024226
          ed |   .1117099   .0060572    18.44   0.000     .0998381    .1235818
       _cons |   3.829366   .0936336    40.90   0.000     3.645848    4.012885
-------------+----------------------------------------------------------------
     sigma_u |  .31951859
     sigma_e |  .15220316
         rho |  .81505521   (fraction of variance due to u_i)
```

Variable	OLS	OLSR	FE	RE
exp	0.04***	0.04***	0.11***	0.09***
exp2	-0.00***	-0.00***	-0.00***	-0.00***
wks	0.01***	0.01**	0.00	0.00
ed	0.08***	0.08***	(omitted)	0.11***
_cons	4.91***	4.91***	4.60***	3.83***

Legend: * p<.05; ** p<.01; *** p<.001

Variable	OLS	OLSR	FE	RE
exp	0.04	0.04	0.11	0.09
	0.00	0.01	0.00	0.00
	0.00	0.00	0.00	0.00
exp2	-0.00	-0.00	-0.00	-0.00
	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00
wks	0.01	0.01	0.00	0.00
	0.00	0.00	0.00	0.00
	0.00	0.00	0.16	0.19
ed	0.08	0.08	(omitted)	0.11
	0.00	0.01		0.01
	0.00	0.00		0.00
_cons	4.91	4.91	4.60	3.83
	0.07	0.14	0.04	0.09
	0.00	0.00	0.00	0.00
N	4165	4165	4165	4165
r2_a	0.28	0.28	0.60	

Legend: b/se/p

Based on the results, a random effects model appears to more suitable since the Wald chi-squared statistic (3012.45) with a p-value of 0.0000 indicates that the model is also highly significant. For the fixed effects model, the F-statistic (F(3, 3567) = 2273.74) with a p-value of 0.0000 indicates that the overall model is highly significant. Moreover, In the fixed effects model, $\rho = 0.9789$, which indicates a high degree of correlation within groups (individuals). This suggests that there are individual-specific effects that need to be accounted for. Ignoring these effects in an OLS model (with or without robust standard errors) would lead to biased and inconsistent estimates.

The suitability of a fixed or random effects model depends on whether the regressors are correlated with the error term, which is addressed in the next question.

Questions 5. Compare the random effect and the fixed effect model, which one is better?

```
use "../Data/mus08psidextract.dta", clear

xtreg lwage exp exp2 wks ed, fe
estimates store FE

xtreg lwage exp exp2 wks ed, re
estimates store RE

hausman FE RE
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))

note: ed omitted because of collinearity.

Fixed-effects (within) regression               Number of obs     =      4,165
Group variable: id                              Number of groups  =        595

R-squared:                                      Obs per group:
     Within  = 0.6566                                         min =          7
     Between = 0.0276                                         avg =        7.0
     Overall = 0.0476                                         max =          7

                                                F(3,3567)         =    2273.74
corr(u_i, Xb) = -0.9107                         Prob > F          =     0.0000

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         exp |   .1137879   .0024689    46.09   0.000     .1089473    .1186284
        exp2 |  -.0004244   .0000546    -7.77   0.000    -.0005315   -.0003173
         wks |   .0008359   .0005997     1.39   0.163    -.0003399    .0020116
          ed |          0  (omitted)
       _cons |   4.596396   .0389061   118.14   0.000     4.520116    4.672677
-------------+----------------------------------------------------------------
     sigma_u |  1.0362039
     sigma_e |  .15220316
         rho |  .97888036   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(594, 3567) = 53.12                  Prob > F = 0.0000



Random-effects GLS regression                   Number of obs     =      4,165
Group variable: id                              Number of groups  =        595

R-squared:                                      Obs per group:
     Within  = 0.6340                                         min =          7
     Between = 0.1716                                         avg =        7.0
     Overall = 0.1830                                         max =          7
```

```
                                                Wald chi2(4)      =    3012.45
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         exp |   .0888609   .0028178    31.54   0.000     .0833382    .0943837
        exp2 |  -.0007726   .0000623   -12.41   0.000    -.0008946   -.0006505
         wks |   .0009658   .0007433     1.30   0.194     -.000491    .0024226
          ed |   .1117099   .0060572    18.44   0.000     .0998381    .1235818
       _cons |   3.829366   .0936336    40.90   0.000     3.645848    4.012885
-------------+----------------------------------------------------------------
     sigma_u |  .31951859
     sigma_e |  .15220316
         rho |  .81505521   (fraction of variance due to u_i)
------------------------------------------------------------------------------

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       FE           RE         Difference       Std. err.
-------------+----------------------------------------------------------------
         exp |    .1137879     .0888609        .0249269               .
        exp2 |   -.0004244    -.0007726        .0003482               .
         wks |    .0008359     .0009658       -.0001299               .
------------------------------------------------------------------------------
                          b = Consistent under H0 and Ha; obtained from xtreg.
           B = Inconsistent under Ha, efficient under H0; obtained from xtreg.

Test of H0: Difference in coefficients not systematic

    chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            = 6191.43
Prob > chi2 =  0.0000
(V_b-V_B is not positive definite)
```

Based on the $\chi^2 = 6191.43$ and p-value $= 0.0000$, we can reject the null hypothesis that there is no correlation between the regressors and the error. This implies that,

$$\mathbb{E}[\epsilon_{it} \mid X_{i1}, X_{i2}, \ldots] \neq 0.$$

Questions 6. Export the above regression results to Excel, Word or Latex. (only need to output one).

The *outreg2* produces nicely formatted MS Word documents. However, *outreg2* produces csv files, they are not well formatted. Latex and text files look better, but they are difficult to read back into R Markdown and render in a pdf as a nicely formatted table. The best alternative is to save a table of regression coefficients using *putexcel*. The Excel file can then be read into R Markdown to display the results in the rendered pdf. However, some effort would be needed to format the table nicely.

Consider another model:

$$lwage_{it} = \alpha + \beta_1 exp_{it} + \beta_2 exp^2_{it} + \beta_3 wks_{it} + \beta_4 ed_{it} + \beta_5 occ_{it} + \epsilon_{it}$$

```
use "../Data/mus08psidextract.dta", clear
cd "C:/Users/micha/MyDocuments/Brunel/Stata_R_Workshop/Programs"
quietly regress lwage exp exp2 wks occ ed
estimates store OLS

outreg2 using myreg.doc,replace ctitle (OLS)

quietly regress lwage exp exp2 wks occ ed
eret li

matrix coef = r(table)
mat li coef
quietly putexcel set regress.xlsx, replace

putexcel A1 = matrix(coef), names
quietly putexcel save

estimate table OLS, star(.05 .01 .001) b(%7.2f) stats(N r2_a)
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))

C:\Users\micha\MyDocuments\Brunel\Stata_R_Workshop\Programs



myreg.doc
dir : seeout



scalars:
                  e(N) =  4165
               e(df_m) =  5
               e(df_r) =  4159
                  e(F) =  336.9656093738905
                 e(r2) =  .2883089703113386
               e(rmse) =  .3895738644125261
                e(mss) =  255.7026390315164
                e(rss) =  631.2022628707333
               e(r2_a) =  .2874533667651873
                 e(ll) =  -1980.523861701613
               e(ll_0) =  -2688.805870567022
               e(rank) =  6

macros:
            e(cmdline) : "regress lwage exp exp2 wks occ ed"
              e(title) : "Linear regression"
          e(marginsok) : "XB default"
                e(vce) : "ols"
             e(depvar) : "lwage"
                e(cmd) : "regress"
         e(properties) : "b V"
            e(predict) : "regres_p"
```

```
             e(model) : "ols"
         e(estat_cmd) : "regress_estat"

matrices:
                 e(b) :  1 x 6
                 e(V) :  6 x 6

functions:
             e(sample)
```

```
coef[9,6]
              exp        exp2         wks         occ          ed       _cons
     b   .0442409  -.00071048   .00575302  -.08123869   .06684296   5.0770654
    se  .00238663   .00005263   .00117895   .01542234   .00282399   .07439699
     t   18.53699  -13.498413   4.8797787     -5.2676   23.669708   68.242887
pvalue  9.041e-74   1.135e-40   1.102e-06   1.452e-07   2.64e-116           0
    ll  .03956183  -.00081367   .00344164  -.11147472   .06130644   4.9312075
    ul  .04891997  -.00060729   .00806439  -.05100267   .07237949   5.2229233
    df       4159        4159        4159        4159        4159        4159
  crit  1.9605345   1.9605345   1.9605345   1.9605345   1.9605345   1.9605345
 eform          0           0           0           0           0           0
```

```
file regress.xlsx saved
```

```
---------------------------
    Variable |    OLS
-------------+-------------
         exp |    0.04***
        exp2 |   -0.00***
         wks |    0.01***
         occ |   -0.08***
          ed |    0.07***
       _cons |    5.08***
-------------+-------------
           N |    4165
        r2_a |    0.29
---------------------------
Legend: * p<.05; ** p<.01; *** p<.001
```

Here is the raw table of regression output after reading in the Excel file created in the previous step.

```r
x <- readxl::read_xlsx("C:/Users/micha/MyDocuments/Brunel/Stata_R_Workshop/Programs/regress.xlsx")
```

```
New names:
* `` -> `...1`
```

```r
names(x)[1] <- "variable"   # the first column needs a name
unlink("regress.xlsx")      # cleanup the output file
x
```

```
# A tibble: 9 x 7
  variable      exp      exp2           wks      occ        ed   `_cons`
  <chr>       <dbl>     <dbl>         <dbl>    <dbl>     <dbl>     <dbl>
1 b        4.42e- 2 -7.10e- 4    0.00575   -8.12e-2 6.68e-  2    5.08
2 se       2.39e- 3  5.26e- 5    0.00118    1.54e-2 2.82e-  3    0.0744
3 t        1.85e+ 1 -1.35e+ 1    4.88      -5.27e+0 2.37e+  1   68.2
4 pvalue   9.04e-74  1.14e-40    0.00000110  1.45e-7 2.64e-116    0
5 ll       3.96e- 2 -8.14e- 4    0.00344   -1.11e-1 6.13e-  2    4.93
6 ul       4.89e- 2 -6.07e- 4    0.00806   -5.10e-2 7.24e-  2    5.22
7 df       4.16e+ 3  4.16e+ 3 4159          4.16e+3 4.16e+  3 4159
8 crit     1.96e+ 0  1.96e+ 0    1.96       1.96e+0 1.96e+  0    1.96
9 eform    0         0           0          0       0            0
```

Questions 7. Consider an endogenous variable $\beta_5 occ_{it}$, and $south_{it}$ and fem_{it} as instrumental variables.

Since wages likely influence occupation, there is a high probability that *occ* is endogenous and, therefore, correlated with the error term.

```stata
use "../Data/mus08psidextract.dta", clear

* First-stage regression
regress occ south fem exp exp2 wks ed
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))

      Source |       SS           df       MS      Number of obs   =     4,165
-------------+----------------------------------   F(6, 4158)      =    455.51
       Model |   412.76633         6  68.7943884   Prob > F        =    0.0000
    Residual |  627.964522     4,158  .151025619   R-squared       =    0.3966
-------------+----------------------------------   Adj R-squared   =    0.3957
       Total |  1040.73085     4,164  .249935363   Root MSE        =    .38862

------------------------------------------------------------------------------
         occ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       south |  -.0382081   .0134306    -2.84   0.004    -.0645391    -.011877
         fem |  -.1449021   .0192327    -7.53   0.000    -.1826084   -.1071957
         exp |  -.0061557   .0023814    -2.58   0.010    -.0108246   -.0014868
        exp2 |   .0000694   .0000525     1.32   0.186    -.0000335    .0001723
         wks |  -.0015946   .0011809    -1.35   0.177    -.0039098    .0007206
          ed |  -.1144709   .0022352   -51.21   0.000    -.1188531   -.1100886
       _cons |   2.170147   .0678959    31.96   0.000     2.037035     2.30326
------------------------------------------------------------------------------
```

In the first-stage regression the endogenous variable *occ* is regressed on the instruments *south* and fem, along with any other exogenous variables. The coefficients of *south* and fem are relatively large compared to the coefficients for the other regressors and they statistically significant, which indicates that the instruments are correlated with the endogenous variable.

The F-statistic for the joint significance of the instruments (*south* and *fem*) and other exogenous variables is large and statistically significant. It is also worth noting that the R-squared value of the first-stage regression is 0.3966, which suggests that the instruments and other exogenous variables explain a substantial amount of the variation in *occ*.

Questions 8. Run 2SLS and GMM.

```
use "../Data/mus08psidextract.dta", clear

* 2SLS
ivregress 2sls lwage exp exp2 wks ed (occ = south fem)

* GMM
gmm (lwage - {b0} - {b1}*exp - {b2}*exp2 - {b3}*wks - {b4}*ed - {b5}*occ), ///
    instruments(exp exp2 wks ed south fem)

* IVGMM
ivregress gmm lwage exp exp2 wks ed (occ = south fem), vce(robust)
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))


Instrumental variables 2SLS regression            Number of obs   =      4,165
                                                  Wald chi2(5)    =     226.62
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =          .
                                                  Root MSE        =     1.2142

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         occ |   2.857199   .3816592     7.49   0.000     2.109161    3.605238
         exp |   .0599443    .007709     7.78   0.000      .044835    .0750536
        exp2 |  -.0008969   .0001658    -5.41   0.000    -.0012219    -.000572
         wks |   .0084283   .0036907     2.28   0.022     .0011946    .0156619
          ed |   .3995287   .0437611     9.13   0.000     .3137585     .485299
       _cons |   -1.03952   .8215311    -1.27   0.206    -2.649691    .5706511
------------------------------------------------------------------------------
Instrumented: occ
 Instruments: exp exp2 wks ed south fem


Step 1
Iteration 0:   GMM criterion Q(b) =  44.653836
Iteration 1:   GMM criterion Q(b) =  .00001417
Iteration 2:   GMM criterion Q(b) =  .00001417

Step 2
Iteration 0:   GMM criterion Q(b) =  9.394e-06
```

```
Iteration 1:   GMM criterion Q(b) =  9.327e-06

GMM estimation

Number of parameters =   6
Number of moments    =   7
Initial weight matrix: Unadjusted                 Number of obs   =      4,165
GMM weight matrix:     Robust

------------------------------------------------------------------------------
             |               Robust
             | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         /b0 |   -1.03706    .858783    -1.21   0.227    -2.720244    .6461234
         /b1 |    .059883   .0081703     7.33   0.000     .0438694    .0758965
         /b2 |  -.0008951   .0001803    -4.96   0.000    -.0012486   -.0005417
         /b3 |   .0084313   .0033893     2.49   0.013     .0017884    .0150741
         /b4 |   .3994096   .0457409     8.73   0.000     .3097591    .4890601
         /b5 |   2.855725   .3988726     7.16   0.000     2.073949    3.637501
------------------------------------------------------------------------------
Instruments for equation 1: exp exp2 wks ed south fem _cons


Instrumental variables GMM regression             Number of obs   =      4,165
                                                  Wald chi2(5)    =     284.25
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =          .
GMM weight matrix: Robust                         Root MSE        =     1.2137

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         occ |   2.855726   .3988726     7.16   0.000     2.073949    3.637502
         exp |    .059883   .0081703     7.33   0.000     .0438694    .0758965
        exp2 |  -.0008951   .0001803    -4.96   0.000    -.0012486   -.0005417
         wks |   .0084313   .0033893     2.49   0.013     .0017884    .0150741
          ed |   .3994096   .0457409     8.73   0.000     .3097591    .4890601
       _cons |   -1.03706    .858783    -1.21   0.227    -2.720244    .6461234
------------------------------------------------------------------------------
Instrumented: occ
 Instruments: exp exp2 wks ed south fem
```

The results for GMM and IVGMM are identical. The results for the 2SLS are nearly identical to the results from the GMM and IVGMM.

Questions 9. Test if occ_{it} is endogenous or not, and examine $south_{it}$, fem_{it} are valid instrumental variables.

```
use "../Data/mus08psidextract.dta", clear

regress lwage exp exp2 wks occ ed
```

```
estimates store ols

ivregress 2sls lwage exp exp2 wks ed (occ = south fem)
estimates store iv
hausman iv ols,constant

estat overid
estat endogenous occ
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))

      Source |       SS           df       MS      Number of obs   =     4,165
-------------+----------------------------------   F(5, 4159)      =    336.97
       Model |  255.702639         5  51.1405278   Prob > F        =    0.0000
    Residual |  631.202263     4,159  .151767796   R-squared       =    0.2883
-------------+----------------------------------   Adj R-squared   =    0.2875
       Total |  886.904902     4,164  .212993492   Root MSE        =    .38957

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         exp |   .0442409   .0023866    18.54   0.000     .0395618      .04892
        exp2 |  -.0007105   .0000526   -13.50   0.000    -.0008137   -.0006073
         wks |    .005753    .001179     4.88   0.000     .0034416    .0080644
         occ |  -.0812387   .0154223    -5.27   0.000    -.1114747   -.0510027
          ed |    .066843    .002824    23.67   0.000     .0613064    .0723795
       _cons |   5.077065    .074397    68.24   0.000     4.931208    5.222923
------------------------------------------------------------------------------



Instrumental variables 2SLS regression          Number of obs   =      4,165
                                                Wald chi2(5)    =     226.62
                                                Prob > chi2     =     0.0000
                                                R-squared       =          .
                                                Root MSE        =     1.2142

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         occ |   2.857199   .3816592     7.49   0.000     2.109161    3.605238
         exp |   .0599443    .007709     7.78   0.000      .044835    .0750536
        exp2 |  -.0008969   .0001658    -5.41   0.000    -.0012219    -.000572
         wks |   .0084283   .0036907     2.28   0.022     .0011946    .0156619
          ed |   .3995287   .0437611     9.13   0.000     .3137585     .485299
       _cons |   -1.03952   .8215311    -1.27   0.206    -2.649691    .5706511
------------------------------------------------------------------------------
Instrumented: occ
 Instruments: exp exp2 wks ed south fem



Note: the rank of the differenced variance matrix (5) does not equal the number
```

```
        of coefficients being tested (6); be sure this is what you expect, or
        there may be problems computing the test.  Examine the output of your
        estimators for anything unexpected and possibly consider scaling your
        variables so that the coefficients are on a similar scale.

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       iv          ols         Difference       Std. err.
-------------+----------------------------------------------------------------
         occ |    2.857199    -.0812387        2.938438        .3813475
         exp |    .0599443     .0442409        .0157034        .0073302
        exp2 |   -.0008969    -.0007105       -.0001865        .0001572
         wks |    .0084283      .005753        .0026752        .0034973
          ed |    .3995287      .066843        .3326857        .0436699
       _cons |    -1.03952     5.077065       -6.116586        .8181555
------------------------------------------------------------------------------
                          b = Consistent under H0 and Ha; obtained from ivregress.
         B = Inconsistent under Ha, efficient under H0; obtained from regress.

Test of H0: Difference in coefficients not systematic

    chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            =  59.37
Prob > chi2 = 0.0000


  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =  .040041  (p = 0.8414)
  Basmann chi2(1)        =  .039974  (p = 0.8415)


  Tests of endogeneity
  H0: Variables are exogenous

  Durbin (score) chi2(1)          =   585.97  (p = 0.0000)
  Wu-Hausman F(1,4158)            =  680.761  (p = 0.0000)
```

The Hausman test compares the estimates from the OLS model with those from an IV model. There are significant differences between these estimates suggesting that the OLS suffers from endogeneity due to *occ*.

Questions 10. Store OLS, 2SLS and GMM regression results in Stata.

```
use "../Data/mus08psidextract.dta", clear

* OLS
regress lwage exp exp2 wks occ ed
estimate store OLS

* 2SLS
ivregress 2sls lwage exp exp2 wks ed (occ = south fem)
estimate store TSLS
```

```
* GMM
gmm (lwage - {b0} - {b1}*exp - {b2}*exp2 - {b3}*wks - {b4}*ed - {b5}*occ), ///
    instruments(exp exp2 wks ed south fem)
estimate store GMM

* IVGMM
ivregress gmm lwage exp exp2 wks ed (occ = south fem), vce(robust)
estimate store IVGMM

estimate table OLS TSLS GMM IVGMM, star(.05 .01 .001) b(%7.2f) stats(N r2_a)
estimate table OLS TSLS GMM IVGMM, b(%7.2f) se(%7.2f) p(%7.2f)  stats(N r2_a)
```

```
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))

      Source |       SS           df       MS      Number of obs   =     4,165
-------------+----------------------------------   F(5, 4159)      =    336.97
       Model |  255.702639         5  51.1405278   Prob > F        =    0.0000
    Residual |  631.202263     4,159  .151767796   R-squared       =    0.2883
-------------+----------------------------------   Adj R-squared   =    0.2875
       Total |  886.904902     4,164  .212993492   Root MSE        =    .38957

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         exp |   .0442409   .0023866    18.54   0.000     .0395618      .04892
        exp2 |  -.0007105   .0000526   -13.50   0.000    -.0008137   -.0006073
         wks |    .005753    .001179     4.88   0.000     .0034416    .0080644
         occ |  -.0812387   .0154223    -5.27   0.000    -.1114747   -.0510027
          ed |    .066843    .002824    23.67   0.000     .0613064    .0723795
       _cons |   5.077065    .074397    68.24   0.000     4.931208    5.222923
------------------------------------------------------------------------------



Instrumental variables 2SLS regression            Number of obs   =      4,165
                                                  Wald chi2(5)    =     226.62
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =          .
                                                  Root MSE        =     1.2142

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         occ |   2.857199   .3816592     7.49   0.000     2.109161    3.605238
         exp |   .0599443    .007709     7.78   0.000      .044835    .0750536
        exp2 |  -.0008969   .0001658    -5.41   0.000    -.0012219    -.000572
         wks |   .0084283   .0036907     2.28   0.022     .0011946    .0156619
          ed |   .3995287   .0437611     9.13   0.000     .3137585     .485299
       _cons |   -1.03952   .8215311    -1.27   0.206    -2.649691    .5706511
------------------------------------------------------------------------------
Instrumented: occ
 Instruments: exp exp2 wks ed south fem
```

```
Step 1
Iteration 0:   GMM criterion Q(b) =  44.653836
Iteration 1:   GMM criterion Q(b) =  .00001417
Iteration 2:   GMM criterion Q(b) =  .00001417

Step 2
Iteration 0:   GMM criterion Q(b) =  9.394e-06
Iteration 1:   GMM criterion Q(b) =  9.327e-06

GMM estimation

Number of parameters =   6
Number of moments    =   7
Initial weight matrix: Unadjusted                 Number of obs   =      4,165
GMM weight matrix:     Robust

------------------------------------------------------------------------------
             |               Robust
             | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         /b0 |   -1.03706    .858783    -1.21   0.227    -2.720244    .6461234
         /b1 |    .059883   .0081703     7.33   0.000     .0438694    .0758965
         /b2 |  -.0008951   .0001803    -4.96   0.000    -.0012486   -.0005417
         /b3 |   .0084313   .0033893     2.49   0.013     .0017884    .0150741
         /b4 |   .3994096   .0457409     8.73   0.000     .3097591    .4890601
         /b5 |   2.855725   .3988726     7.16   0.000     2.073949    3.637501
------------------------------------------------------------------------------
Instruments for equation 1: exp exp2 wks ed south fem _cons


Instrumental variables GMM regression             Number of obs   =      4,165
                                                  Wald chi2(5)    =     284.25
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =          .
GMM weight matrix: Robust                         Root MSE        =     1.2137

------------------------------------------------------------------------------
             |               Robust
       lwage | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         occ |   2.855726   .3988726     7.16   0.000     2.073949    3.637502
         exp |    .059883   .0081703     7.33   0.000     .0438694    .0758965
        exp2 |  -.0008951   .0001803    -4.96   0.000    -.0012486   -.0005417
         wks |   .0084313   .0033893     2.49   0.013     .0017884    .0150741
          ed |   .3994096   .0457409     8.73   0.000     .3097591    .4890601
       _cons |   -1.03706    .858783    -1.21   0.227    -2.720244    .6461234
------------------------------------------------------------------------------
Instrumented: occ
 Instruments: exp exp2 wks ed south fem
```

Variable	OLS	TSLS	GMM	IVGMM
_				
exp	0.04***	0.06***		0.06***
exp2	-0.00***	-0.00***		-0.00***
wks	0.01***	0.01*		0.01*
occ	-0.08***	2.86***		2.86***
ed	0.07***	0.40***		0.40***
_cons	5.08***	-1.04		-1.04
b0				
_cons			-1.04	
b1				
_cons			0.06***	
b2				
_cons			-0.00***	
b3				
_cons			0.01*	
b4				
_cons			0.40***	
b5				
_cons			2.86***	
Statistics				
N	4165	4165	4165	4165
r2_a	0.29	.		.

Legend: * p<.05; ** p<.01; *** p<.001

Variable	OLS	TSLS	GMM	IVGMM
_				
exp	0.04	0.06		0.06
	0.00	0.01		0.01
	0.00	0.00		0.00
exp2	-0.00	-0.00		-0.00
	0.00	0.00		0.00
	0.00	0.00		0.00
wks	0.01	0.01		0.01
	0.00	0.00		0.00
	0.00	0.02		0.01
occ	-0.08	2.86		2.86
	0.02	0.38		0.40
	0.00	0.00		0.00
ed	0.07	0.40		0.40

	0.00	0.04		0.05
	0.00	0.00		0.00
_cons	5.08	-1.04		-1.04
	0.07	0.82		0.86
	0.00	0.21		0.23
b0				
_cons			-1.04	
			0.86	
			0.23	
b1				
_cons			0.06	
			0.01	
			0.00	
b2				
_cons			-0.00	
			0.00	
			0.00	
b3				
_cons			0.01	
			0.00	
			0.01	
b4				
_cons			0.40	
			0.05	
			0.00	
b5				
_cons			2.86	
			0.40	
			0.00	
Statistics				
N	4165	4165	4165	4165
r2_a	0.29	.		.

Legend: b/se/p