Stata – Session 3

Tarjei Havnes

[1]ESOP and Department of Economics
University of Oslo

[2]Research department
Statistics Norway

ECON 4136, UiO, 2012

Preparation

Before we start:

1. Open StataIC 11 from kiosk.uio.no (Internet Explorer!), using your UIO user name
2. Follow the instructions on installing add-ons available on the course homepage
3. - findit estout - and install the estout-package (st0085_1)
4. Download the Wooldridge data

Learning goals

1. set up panel data: -reshape-, -merge-, -append-
2. work with panel data
   - -xtset- , -xtdata- [, -tsset-]
   - D. and L.-operators
   - i. and c.-operators (-c.- only in Stata 12)
3. descriptives in panel data:
   - -xtsum-, -xttab-, -xttrans-: decompose variation
   - graph data
4. regression in panel data
   - fixed effects
     - ★ First-difference
     - ★ OLS with dummy variables
     - ★ -xtreg, fe-
   - random effects
     - ★ -xtreg, re-
   - compare FE and RE: -hausman-
   - IV-regression: -xtivreg-

Reshape

(wide form)

id	sex	inc80	inc81	inc82
1	0	5000	5500	6000
2	1	2000	2200	3300
3	0	3000	2000	1000

(long form)

id	year	sex	inc
1	80	0	5000
1	81	0	5500
1	82	0	6000
2	80	1	2000
2	81	1	2200
2	82	1	3300
3	80	0	3000
3	81	0	2000
3	82	0	1000

You can move from wide to long

- `reshape long inc, i(id sex) j(year)`

or from long to wide

- `reshape wide inc, i(id sex) j(year)`

(try it with country2.dta)

Combining datasets vertically (append)

```
. use a
. append using b
```

(a.dta)

x	y
1	1.2
2	2.3
3	0.5

(b.dta)

x	z
6	0.03
12	0.01

(b appended to a)

x	y	z
1	1.2	.
2	2.3	.
3	0.5	.
6	.	0.03
12	.	0.01

Combining datasets horizontally (merge)

```
. use c
. sort id
. merge id using d
```

(c.dta)

id	y
1	1.2
2	2.3
3	0.5

(d.dta)

id	x
1	3.5
2	1.0
6	0.1

(d merged to c)

id	y	x	_merge
1	1.2	3.5	3
2	2.3	1.0	3
3	0.5	.	1
6	.	0.1	2

_merge==1 observation in master only
_merge==2 observation in using only
_merge==3 observation in both master and using

Merge requires both datasets to be sorted on the merge vars

Declaring panel data

Stata has many built-in commands that can be used with panel data

- require that Stata knows what variables denote the panel
- declare using -xtset *idvar timevar*-

```
. loc urlpath http://www.uio.no/studier/emner/sv/oekonomi/ECON4136/h12/undervisningsmateriale

. use  `urlpath'/mus08psidextract.dta

. describe
[output omitted]

. xtset id t
       panel variable:  id (strongly balanced)
        time variable:  t, 1 to 7
                delta:  1 unit
```

Declaring panel data

The prefix -L.- and -D.- tells Stata to consider lagged or differenced variables

- can be combined, e.g. -LD.-
- allow different lag length, e.g. -L2D.-

```
. corr lwage D.lwage wks D.wks
(obs=3570)
             |                   D.                  D.
             |    lwage     lwage       wks       wks
-------------+------------------------------------
       lwage |
         --. |   1.0000
         D1. |   0.2240    1.0000
         wks |
         --. |   0.0350    0.0097    1.0000
         D1. |  -0.0418   -0.0057    0.5090    1.0000


. corr lwage D.lwage LD.lwage
(obs=2975)
             |                   D.       LD.
             |    lwage     lwage     lwage
-------------+----------------------------
       lwage |
         --. |   1.0000
         D1. |   0.2560    1.0000
         LD. |   0.0747   -0.3521    1.0000
```

Working with panel data

When working with panel data, you will find great use for some of the commands we have discussed previously

- -collapse-, -bysort-, -tabstat-

```
. gen wage = exp(lwage)/wks

. preserve

. collapse (mean) wmean=wage (sd) wsd=wage (p50) wp50=wage , by(t)

. order t

. cl
             t       wmean         wsd        wp50
  1.         1    13.99719    6.524022    13.24006
  2.         2    14.76781    5.358874     14.6452
  3.         3    17.47416    9.978057    15.97222
  4.         4    19.12041    9.701109    17.48892
  5.         5    20.86073     9.36695    19.38772
  6.         6    22.72049    10.50036    21.16269
  7.         7    25.26531    13.65805    23.36735

. restore


. tabstat wage, s(mean sd p50) by(t)
Summary for variables: wage
     by categories of: t
       t |      mean         sd        p50
---------+------------------------------
       1 |   13.99719   6.524022   13.24006
```

Summary statistics: Decomposing variation

To describe panel data, we are often interested in summarizing the dimensions separately

1. Overall variance: $s_O^2 = \frac{1}{NT-1}\sum_i\sum_t (x_{it} - \bar{x})^2$
2. Within variance:
   $s_W^2 = \frac{1}{NT-1}\sum_i\sum_t (x_{it} - \bar{x}_i)^2 = \frac{1}{NT-1}\sum_t\sum_i (x_{it} - \bar{x}_i + \bar{x})^2$
3. Between variance: $s_W^2 = \frac{1}{N-1}\sum_i (\bar{x}_i - \bar{x})^2$

```
. xtsum id t lwage ed exp wks south

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
id       overall |       298   171.7821         1        595 |     N =     4165
         between |             171.906         1        595 |     n =      595
         within  |                 0         298        298 |     T =        7
                 |                                          |
t        overall |         4    2.00024         1          7 |     N =     4165
         between |                 0         4          4 |     n =      595
         within  |             2.00024         1          7 |     T =        7
                 |                                          |
lwage    overall |  6.676346   .4615122     4.60517      8.537 |     N =     4165
         between |             .3942387      5.3364   7.813596 |     n =      595
         within  |             .2404023    4.781808   8.621092 |     T =        7
                 |                                          |
ed       overall |  12.84538   2.787995         4         17 |     N =     4165
         between |             2.790006         4         17 |     n =      595
         within  |                 0    12.84538   12.84538 |     T =        7
                 |                                          |
exp      overall |  19.85378   10.96637         1         51 |     N =     4165
```

Summary statistics: Decomposing variation

```
. xttab south
                    Overall              Between              Within
    south |     Freq.  Percent       Freq.  Percent         Percent
----------+-----------------------------------------------------------
        0 |     2956     70.97         428     71.93           98.66
        1 |     1209     29.03         182     30.59           94.90
----------+-----------------------------------------------------------
    Total |     4165    100.00         610    102.52           97.54
                                (n = 595)

. xttrans south, freq

residence; |
  south==1 |  residence; south==1
 if in the | if in the South area
South area |         0           1 |     Total
-----------+----------------------+----------
         0 |     2,527           8 |     2,535
           |     99.68        0.32 |    100.00
-----------+----------------------+----------
         1 |         8       1,027 |     1,035
           |      0.77       99.23 |    100.00
-----------+----------------------+----------
     Total |     2,535       1,035 |     3,570
           |     71.01       28.99 |    100.00h
```

Graphing panel data

It is often useful to graph separate time series plots for some or all units

```
. xtline lwage if id<= 20

. qui xtline lwage if id<=20, overlay legend(off) saving(lwage, replace)

. qui xtline wks if id<=20, overlay legend(off) saving(wks, replace)

. graph combine lwage.gph wks.gph, iscale(1)
```

Fixed effect, first difference

Remember that the FE-model can be rephrased in first-differences.

```
. reg D.(lwage exp exp2 wks ed), vce(cluster id)
note: _delete omitted because of collinearity
note: _delete omitted because of collinearity

Linear regression                                       Number of obs =    3570
                                                        F(  2,   594) =   22.66
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.0041
                                                        Root MSE      =  .18156
                                      (Std. Err. adjusted for 595 clusters in id)
------------------------------------------------------------------------------
             |               Robust
     D.lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |
          D1.|  (omitted)
             |
        exp2 |
          D1.|  -.0005321   .0000808    -6.58   0.000    -.0006908   -.0003734
             |
         wks |
          D1.|  -.0002683   .0011783    -0.23   0.820    -.0025824    .0020459
             |
          ed |
          D1.|  (omitted)
             |
       _cons |   .1170654   .0040974    28.57   0.000     .1090182    .1251126
------------------------------------------------------------------------------

. est sto fd
```

Fixed effect, dummy variables

Or, the FE-model can be phrased with individual dummy variables

```
. xi: reg lwage exp exp2 wks ed i.id, vce(cluster id)
i.id              _Iid_1-595          (naturally coded; _Iid_1 omitted)
note: _Iid_468 omitted because of collinearity

Linear regression                                      Number of obs =    4165
                                                       F(  2,   594) =       .
                                                       Prob > F      =       .
                                                       R-squared     =  0.9068
                                                       Root MSE      =   .1522
                                  (Std. Err. adjusted for 595 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .1137879   .0043514    26.15   0.000     .1052418    .1223339
        exp2 |  -.0004244   .0000888    -4.78   0.000    -.0005988     -.00025
         wks |   .0008359   .0009393     0.89   0.374    -.0010089    .0026806
          ed |  -.2749652   .0087782   -31.32   0.000    -.2922053    -.257725
      _Iid_2 |  -1.536779   .0375552   -40.92   0.000    -1.610536   -1.463021
      _Iid_3 |   1.037793   .0249882    41.53   0.000     .9887171    1.086869
      _Iid_4 |  -2.022192   .0453656   -44.58   0.000    -2.111288   -1.933095
[output omitted]

. est sto dum

// you could also generate dummies actively: -qui tab id, gen(iddum)-
// in Stata 12, you should avoid -xi:- in favor of factor variables
```

Fixed effect -xtreg-

You should always estimate FE-models using -xtreg- (except specification searching)

```
. xtreg lwage exp exp2 wks ed , i(id) fe
note: ed omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =      4165
Group variable: id                              Number of groups   =       595

R-sq:  within  = 0.6566                         Obs per group: 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 |      Coef.   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 |  (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) =    40.17                 Prob > F = 0.0000

. est sto fe
```

Random effects

We can also use -xtreg, re- to estimate models with random effects, rather than fixed effects

$$y_i = \bar{\alpha} + \mathbf{x}_i'\beta + (\alpha_i - \bar{\alpha} + u_i)$$

```
. xtreg lwage exp exp2 wks ed , i(id) re

Random-effects GLS regression                   Number of obs      =      4165

Group variable: id                              Number of groups   =       595
R-sq:  within  = 0.6340                         Obs per group: min =         7
       between = 0.1716                                        avg =       7.0
       overall = 0.1830                                        max =         7

Random effects u_i ~ Gaussian                   Wald chi2(4)       =   3012.45
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000
------------------------------------------------------------------------------
       lwage |      Coef.   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)
------------------------------------------------------------------------------

. est sto re
```

Compare RE and FE – Hausman test

Remember that the FE-model is consistent under the RE-model, but less efficient.

- But the RE-model is *not* consistent under the FE-model ($cov(\alpha_i, \mathbf{x}_i) \neq 0$).
- Test the RE-model by comparing the coefficients – Hausman test.

$$y_i = \bar{\alpha} + \mathbf{x}_i'\beta + (\alpha_i - \bar{\alpha} + u_i)$$

```
. hausman fe re

                ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       fe           re         Difference          S.E.
-------------+----------------------------------------------------------------
         exp |    .1137879     .0888609        .0249269                .
        exp2 |   -.0004244    -.0007726        .0003482                .
         wks |    .0008359     .0009658       -.0001299                .
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  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)
```

IV with FE

```
. xtivreg lwage exp exp2 (wks=occ), i(id) fe

Fixed-effects (within) IV regression          Number of obs      =         4165
Group variable: id                            Number of groups   =          595

R-sq:  within  = 0.5581                       Obs per group: min =            7
       between = 0.0261                                      avg =          7.0
       overall = 0.0456                                      max =            7

                                              Wald chi2(3)       =     6.23e+06
corr(u_i, Xb)  = -0.9084                      Prob > chi2        =       0.0000
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wks |  -.0183475   .0163986    -1.12   0.263    -.0504883    .0137932
         exp |    .118264   .0047392    24.95   0.000     .1089754    .1275526
        exp2 |  -.0005397   .0001164    -4.64   0.000    -.0007678   -.0003116
       _cons |   5.464863   .7430685     7.35   0.000     4.008475     6.92125
-------------+----------------------------------------------------------------
     sigma_u |  1.0368799
     sigma_e |  .17266125
         rho |  .97301924   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F  test that all u_i=0:     F(594,3567) =    41.11           Prob > F    = 0.0000
------------------------------------------------------------------------------
Instrumented:   wks
Instruments:    exp exp2 occ
------------------------------------------------------------------------------

. est sto iv
```

IV with FE

```
. xtreg wks occ exp exp2 , i(id) fe

Fixed-effects (within) regression               Number of obs      =      4165
Group variable: id                              Number of groups   =       595

R-sq:  within  = 0.0061                         Obs per group: min =         7
       between = 0.0014                                        avg =       7.0
       overall = 0.0019                                        max =         7

                                                F(3,3567)          =      7.27
corr(u_i, Xb)  = -0.1963                        Prob > F           =    0.0001
------------------------------------------------------------------------------
         wks |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         occ |   .9435816   .3805191     2.48   0.013     .1975247    1.689638
         exp |   .2411641   .0688367     3.50   0.000     .1062009    .3761274
        exp2 |  -.0061183   .0015213    -4.02   0.000    -.0091011   -.0031355
       _cons |   44.68845   .8122636    55.02   0.000      43.0959    46.28099
-------------+----------------------------------------------------------------
     sigma_u |    3.35623
     sigma_e |  4.2460376
         rho |  .38453678   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(594, 3567) =     4.19                 Prob > F = 0.0000

. est sto fs
```

IV with FE

```
. esttab dum fe re iv, keep(exp exp2 wks ed) order(exp exp2 wks ed) stat(N hausman) se mtit

------------------------------------------------------------------------------
                      (1)             (2)             (3)             (4)
                      dum              fe              re              iv
------------------------------------------------------------------------------

exp                 0.114***        0.114***        0.0889***        0.118***
                  (0.00435)       (0.00247)       (0.00282)       (0.00474)

exp2             -0.000424***    -0.000424***    -0.000773***    -0.000540***
                (0.0000888)     (0.0000546)     (0.0000623)      (0.000116)

wks                 0.000836        0.000836        0.000966         -0.0183
                  (0.000939)      (0.000600)      (0.000743)        (0.0164)

ed                  -0.275***               .           0.112***
                  (0.00878)               .         (0.00606)

------------------------------------------------------------------------------
N                      4165            4165            4165            4165
hausman                                                6191.4
------------------------------------------------------------------------------
```