Panel Data Analysis Using Stata

Sebastian T. Braun
University of St Andrews

Course Objectives

- To provide a concise **introduction** to applied panel data analysis.
- To review core theoretical methods of panel data analysis and apply these methods hands-on.
- To learn how to analyze (microeconometric) panel data using the statistical software Stata.

Recommended Readings

The applied part of the course will draw heavily on Chapter 8 of

- Cameron, A. Colin and Pravin K. Trivedi (2010). *Applied Microeconometrics Using Stata*. Stata Press.

Recommended introductory textbooks that provide an introduction to panel data analysis are:

- Wooldridge, Jeffrey M. (2015). *Introductory Econometrics*. Cengage Learning Services, 5th edition.
- Kennedy, Peter (2008). *A Guide to Econometrics*. John Wiley & Sons, 6th edition.

Course Material

You find the slides on my homepage:

- https://sebastiantillbraun.wordpress.com/teaching/

Overview

1. Course Outline ✓
2. Introduction
3. Panel Data Management
4. Regression Analysis
5. Hypothesis Testing
6. Extensions
7. Outlook: Advanced Panel Data Analysis

What is Panel Data?

- A cross-section (of people, firms, countries, etc.) is observed over time.
- Panel data provides observations on the same units in several time periods (unlike independently pooled cross sections).
- Panel data often consist of a very large number of cross-sections over a small number of time periods.

What Advantages Do Panel Data Offer?

Panel data allows us to...

- ...examine issues that cannot be studied using either time series or cross-sectional data.
- ...deal with unobserved heterogeneity in the micro units.
- ...analyze dynamics with only a short time series.
- ...increase the efficiency of estimation.

What Advantages Do Panel Data Offer?

Panel data allows us to...

- ...examine issues that cannot be studied using either time series or cross-sectional data
- ...deal with unobserved heterogeneity in the micro units.
- ...analyze dynamics with only a short time series.
- ...increase the efficiency of estimation.

Getting Started...

We now consider data from the Panel Study of Income Dynamics.

- You can install the relevant files from within *Stata*. Type:
  . net from http://www.stata-press.com/data/mus
  . net install mus
  . net get mus
- You can also download the data from www.stata-press.com/data/mus.html.

The Dataset

- Open the data set:

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

- The data set contains information on 595 individuals (the cross-sectional units) over 7 years (1976-1982).
- The total number of observations is thus $595 \times 7 = 4165$.
- There are no missing observations (so the data set is balanced).

describe the Data

```
. use "mus08psidextract.dta", clear
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))

. describe

Contains data from mus08psidextract.dta
  obs:         4,165                          PSID wage data 1976-82 from
                                                Baltagi and Khanti-Akom (1990)
 vars:            22                          26 Nov 2008 17:15
 size:       295,715 (99.7% of memory free)   (_dta has notes)
-------------------------------------------------------------------------------
              storage  display     value
variable 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
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
lwage           float  %9.0g                  log wage
id              float  %9.0g
t               float  %9.0g
tdum1           byte   %8.0g                  t==    1.0000
exp2            float  %9.0g
-------------------------------------------------------------------------------
```

Panel Data Organization

- Panel data is usually organised in the so-called long form, with each observation a distinct individual-time pair.
- In our case, the cross-section (panel) and time variables are *id* and t, respectively.

Panel Data Organization (ctd.)

```
. * Organization of dataset
. list id t lwage exp union occ in 1/14, clean

        id   t     lwage   exp   union   occ
  1.     1   1   5.56068     3       0     0
  2.     1   2   5.72031     4       0     0
  3.     1   3   5.99645     5       0     0
  4.     1   4   5.99645     6       0     0
  5.     1   5   6.06146     7       0     0
  6.     1   6   6.17379     8       0     0
  7.     1   7   6.24417     9       0     0
  8.     2   1   6.16331    30       0     1
  9.     2   2   6.21461    31       0     1
 10.     2   3    6.2634    32       1     1
 11.     2   4   6.54391    33       0     1
 12.     2   5   6.69703    34       0     1
 13.     2   6   6.79122    35       0     1
 14.     2   7   6.81564    36       0     1
```

Panel Data Organization (ctd.)

- Inform *Stata* about the panel and time variables *id* and *t* by typing:
  . xtset id t
- You can now use the time-series operators of *Stata* (*L*., *D*.,...) and all the *xt* commands.

xtdescribe the data

```
. * Panel description of dataset
. xtdescribe

      id:  1, 2, ..., 595                                    n =        595
       t:  1, 2, ..., 7                                      T =          7
           Delta(t) = 1 unit
           Span(t)  = 7 periods
           (id*t uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         7       7       7         7         7       7       7

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
      595    100.00  100.00 |  1111111
 ---------------------------+---------
      595    100.00         |  XXXXXXX
```

Estimating the Union Wage Premium

- We now use the panel data to estimate the union wage premium.
- In our case, the premium measures the degree to which wages are higher if set by a union contract.
- In general, the vast empirical literature on the issue finds that union bargaining increase wages above the market rate.
- We will see how panel data can be used to overcome some of the difficulties associated with estimating the wage premium.
- We restrict the analysis to men (*drop if fem == 1*)!

The Basic Linear Panel Model

$$y_{it} = \alpha + x_{it}\beta + a_i + \epsilon_{it} \quad (1)$$

- i denotes the cross-sectional unit and t the time period.
- y_{it} is the dependent variable.
- α is a common intercept.
- x_{it} are explanatory variables.
- a_i are unobserved individual-specific (fixed) effects.
- ϵ_{it} is an error term.

The Basic Linear Panel Model

$$w_{it} = \alpha + x_{it}\beta + \gamma Union_{it} + a_i + \epsilon_{it} \quad (2)$$

- i denotes the cross-sectional unit and t the time period.
- w_{it} is the log of the hourly wage.
- α is a common intercept.
- $Union_{it}$ indicates whether wage is set by a union contract.
- a_i are unobserved individual-specific (fixed) effects.
- ϵ_{it} is an error term.

The Unobserved Fixed Effect

$$w_{it} = \alpha + x_{it}\beta + \gamma Union_{it} + a_i + \epsilon_{it} \tag{3}$$

- The a_i captures all **unobserved**, **time constant** factors that affect w_{it}.
- The unobserved fixed effect is specific to an individual and does not vary over time.
- The unobserved fixed effect is unknown to the researcher.
- Examples: ability, ambition...

Pooled OLS

$$w_{it} = \alpha + x_{it}\beta + \gamma Union_{it} + a_i + \epsilon_{it}. \quad (4)$$

- How should one estimate the parameter of interest, γ, given our seven years of panel data?
- One possibility is just to 'pool' the data and use OLS.
- Do this in *Stata* using *exp*, *exp*2, *wks*, *ed*, *ind* and *occ* as additional controls.

Pooled OLS Estimates

```
. ******* 2. POOLED OLS
.
. * Pooled OLS with incorrect default standard errors
. regress lwage exp exp2 wks ed union ind occ

      Source |       SS       df       MS              Number of obs =    3696
-------------+------------------------------           F(  7,  3688) =  229.48
       Model |  215.291596     7  30.7559423           Prob > F      =  0.0000
    Residual |  494.284928  3688  .134025197           R-squared     =  0.3034
-------------+------------------------------           Adj R-squared =  0.3021
       Total |  709.576524  3695  .192036948           Root MSE      =  .36609

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .0384824    .002442    15.76   0.000     .0336945    .0432703
        exp2 |  -.0006084   .0000539   -11.29   0.000     -.000714   -.0005027
         wks |   .0047247   .0012331     3.83   0.000     .0023071    .0071422
          ed |   .0635993   .0028509    22.31   0.000     .0580098    .0691887
       union |   .1204051   .0138341     8.70   0.000     .0932818    .1475284
         ind |   .0431938   .0126986     3.40   0.001     .0182968    .0680908
         occ |   -.150339    .016286    -9.23   0.000    -.1822695   -.1184086
       _cons |    5.24959   .0780379    67.27   0.000     5.096589    5.402592
------------------------------------------------------------------------------
```

What is Wrong with Pooled OLS?

Let us re-write the linear basic panel model as follows:

$$w_{it} = \alpha + x_{it}\beta + \gamma Union_{it} + (a_i + \epsilon_{it}) \quad (5)$$
$$= \alpha + x_{it}\beta + \gamma Union_{it} + v_{it}. \quad (6)$$

where $v_{it} = a_i + \epsilon_{it}$ is referred to as the **composite error term**.

Problem 1: Serially Correlated Errors

- If w is overpredicted in one year for a given person, then it is likely to be overpredicted in other years.
- The composite error $v_{it} = a_i + \epsilon_{it}$ is serially correlated even if ϵ_{it} is $i.i.d.$ with a variance of σ_ϵ^2.
- A worker that is more able today will be more able tomorrow and we thus have:

$$Cor(v_{it}, v_{is}) \neq 0 \text{ for } t \neq s \text{ if } a_i \neq 0. \tag{7}$$

- We can easily show this in *Stata*...

Problem 1: Serially Correlated Errors (ctd.)

```
.
. * Autocorrelations of residual
. quietly regress lwage exp exp2 wks ed union ind occ

. predict uhat, residuals

. corr uhat L1.uhat
(obs=3168)

             |                 L.
             |     uhat     uhat
-------------+------------------
        uhat |
         --. |   1.0000
         L1. |   0.8599   1.0000
```

Problem 1: Serially Correlated Errors (ctd.)

- Each additional observation for a given person provides less than an independent piece of new information.
- With serially correlated errors, standard errors are thus biased.

Solution: Cluster-Robust Standard Errors

- Calculate cluster-robust standard errors that allow for correlation within clusters (cross-sections).
- Cluster-robust standard errors only require that errors are independent between cross-sections.
- Use the *vce*(*cluster*) option in *Stata*...

OLS with Cluster-Robust Standard Errors

```
. * Pooled OLS with cluster-robust standard errors
. regress lwage exp exp2 wks ed union ind occ, vce(cluster id)

Linear regression                                      Number of obs =    3696
                                                       F(  7,    527) =   46.65
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3034
                                                       Root MSE      =  .36609

                                  (Std. Err. adjusted for 528 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .0384824   .0047986     8.02   0.000     .0290556    .0479092
        exp2 |  -.0006084   .0001087    -5.60   0.000    -.0008219   -.0003948
         wks |   .0047247   .0018448     2.56   0.011     .0011005    .0083488
          ed |   .0635993   .0062134    10.24   0.000     .0513931    .0758054
       union |   .1204051    .027477     4.38   0.000     .0664272     .174383
         ind |   .0431938   .0254036     1.70   0.090     -.006711    .0930986
         occ |   -.150339   .0321478    -4.68   0.000    -.2134925   -.0871856
       _cons |    5.24959   .1434456    36.60   0.000     4.967795    5.531386
------------------------------------------------------------------------------
```

Problem 2: Omitted Variable Bias

$$w_{it} = \alpha + x_{it}\beta + \gamma Union_{it} + v_{it} \tag{8}$$

- We must assume that $v_{it} = a_i + \epsilon_{it}$ is uncorrelated with $Union_{it}$ for OLS to consistently estimate γ.
- So even if ϵ_{it} is uncorrelated with $Union_{it}$, pooled OLS is biased and inconsistent if a_i and $Union_{it}$ are correlated.
- The resulting **heterogeneity bias** is caused from omitting a time-constant variable.

Problem 2: Omitted Variable Bias (ctd.)

Why should a_i and $Union_{it}$ be correlated?

- Unobserved factors that affect wages may also affect workers' selection into the covered sector.
- Wage standardization policy of unions might be most appealing to workers with low underlying earnings potential.
- Unionised employers might pick workers from the queue, as not all workers who desire union employment can find union jobs.

$\Rightarrow$ $Union_{it}$ might be positively or negatively correlated with ability.

The Fixed Effects Model

- The fixed effects model allows the unobserved effects to be correlated with the explanatory variables.
- In fact, it uses a transformation to remove the unobserved effect prior to estimation.

The Fixed Effects Transformation

Consider our basic linear panel model:

$$y_{it} = \alpha + x_{it}\beta + a_i + \epsilon_{it}. \tag{9}$$

For each cross-section, average this equation over time:

$$\overline{y_i} = \alpha + \overline{x}_i\beta + a_i + \overline{\epsilon}_i, \tag{10}$$

where the average of some variable z is given by
$\overline{z}_i = T^{-1} \sum_{t=1}^{T} y_{it}$.

The Fixed Effects Transformation (ctd.)

Now subtract equation (10) from (9) to get rid of the fixed effect:

$$\begin{aligned} (y_{it} - \overline{y_i}) &= (\alpha - \alpha) + (x_{it} - \overline{x}_i)\beta + (a_i - a_i) + (\epsilon_{it} - \overline{\epsilon}_i) \\ &= (x_{it} - \overline{x}_i)\beta + (\epsilon_{it} - \overline{\epsilon}_i). \end{aligned} \tag{11}$$

Because a_i has been eliminated, OLS leads to consistent estimates of β even if x_{it} is correlated with a_i!

Re-Estimate the Union Wage Premium...

Use *xtreg*, *fe* in *STATA* to re-estimate the union wage premium using the fixed effects model:

$$(w_{it} - \overline{w_i}) = (x_{it} - \overline{x}_i)\beta + \gamma(Union_{it} - \overline{Union}_i) + (\epsilon_{it} - \overline{\epsilon}_i). \quad (12)$$

What does your estimate of γ suggests about the correlation between *Union* and ability?

Fixed Effects Estimates

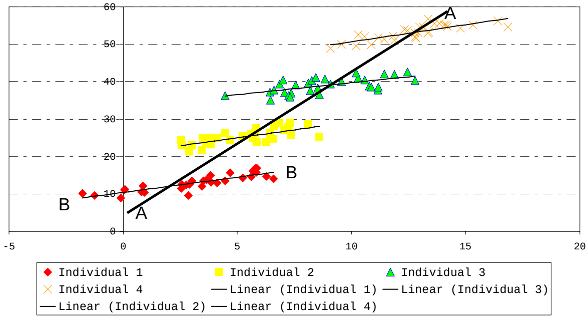
```
. ******* 3. FIXED EFFECTS ESTIMATOR (WITHIN ESTIMATOR)
.
. * Within or FE estimator
. xtreg lwage exp exp2 wks ed union ind occ, fe
note: ed omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =      3696
Group variable: id                              Number of groups   =       528

R-sq:  within  = 0.6558                         Obs per group: min =         7
       between = 0.0228                                        avg =       7.0
       overall = 0.0427                                        max =         7

                                                F(6,3162)          =   1004.25
corr(u_i, Xb)  = -0.9223                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .1149389   .0026801    42.89   0.000     .1096841    .1201938
        exp2 |  -.0004347   .0000584    -7.44   0.000    -.0005491   -.0003202
         wks |   .0004693   .0006576     0.71   0.476    -.0008201    .0017587
          ed |  (omitted)
       union |   .0316998   .0159769     1.98   0.047     .0003736     .063026
         ind |   .0182395   .0160431     1.14   0.256    -.0132165    .0496954
         occ |  -.0113013   .0146455    -0.77   0.440    -.0400169    .0174144
       _cons |   4.600501    .044283   103.89   0.000     4.513675    4.687328
-------------+----------------------------------------------------------------
     sigma_u |   1.047369
     sigma_e |   .1534777
         rho |  .97897847   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(527, 3162) =    33.82           Prob > F = 0.0000
```

Caveats of the Fixed-Effects Estimator

The fixed effects estimator uses the time variation in y and x *within* cross-sectional units only.

- It discards variation across cross-sections (*between* variation).
- It does not allow us to estimate the coefficients of time-invariant regressors (gender, education...).
- Differenced regressors may be more susceptible to measurement error.
- Does not solve the problem of time-varying omitted variables.

Fixed Effects Estimator (by David Bell)



Within- and Between-Variation

The STATA command *xtsum* decomposes the overall variation in a variable as follows (where $s_O^2 \approx s_W^2 + s_B^2$):

$$\text{Within: } s_W^2 = \frac{1}{NT-1} \sum_i \sum_t (x_{it} - \overline{x}_i)^2, \tag{13}$$

$$\text{Between: } s_B^2 = \frac{1}{N-1} \sum_i (\overline{x_i} - \overline{\overline{x}})^2, \tag{14}$$

$$\text{Overall: } s_O^2 = \frac{1}{NT-1} \sum_i \sum_t (x_{it} - \overline{\overline{x}})^2. \tag{15}$$

Use *xtsum* (and possibly *xttrans*) to assess the relative importance of between and within variation in the data.

xtsum the Data

```
. * Panel summary statistics: within and between variation
. * Notice: The min and max columns give the min and max of x_it for overall,
x^bar_i for
> between and x_it-x^bar_i+x^bar for within
. xtsum id t lwage exp wks ed union tdum1

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
id       overall |  297.7292   171.7455          1        594 |     N =    3696
         between |             171.8851          1        594 |     n =     528
         within  |                    0   297.7292   297.7292 |     T =       7
                 |                                            |
t        overall |         4   2.000271          1          7 |     N =    3696
         between |                    0          4          4 |     n =     528
         within  |             2.000271          1          7 |     T =       7
                 |                                            |
lwage    overall |  6.729774   .4382202    5.01728      8.537 |     N =    3696
         between |             .3656286   5.518704   7.813596 |     n =     528
         within  |             .2420119   4.835235   8.674519 |     T =       7
                 |                                            |
exp      overall |  20.21402   10.99381          1         51 |     N =    3696
         between |             10.81909          4         48 |     n =     528
         within  |             2.000271   17.21402   23.21402 |     T =       7
                 |                                            |
wks      overall |  46.96374   4.983413          5         52 |     N =    3696
         between |             3.165729   31.57143   51.57143 |     n =     528
         within  |             3.850824   12.39232   63.82089 |     T =       7
                 |                                            |
ed       overall |  12.84659   2.822298          4         17 |     N =    3696
         between |             2.824592          4         17 |     n =     528
         within  |                    0   12.84659   12.84659 |     T =       7
                 |                                            |
union    overall |  .3833874   .4862772          0          1 |     N =    3696
         between |             .4597488          0          1 |     n =     528
         within  |             .1594985  -.4737554    1.24053 |     T =       7
                 |                                            |
tdum1    overall |  .1428571   .3499745          0          1 |     N =    3696
         between |                    0   .1428571   .1428571 |     n =     528
         within  |             .3499745          0          1 |     T =       7
```

xttrans union

```
. xttrans union, freq

   if wage |
  set be a |   if wage set be a
     union |    union contract
  contract |         0          1 |     Total
-----------+----------------------+----------
         0 |     1,890         66 |     1,956
           |     96.63       3.37 |    100.00
-----------+----------------------+----------
         1 |        63      1,149 |     1,212
           |      5.20      94.80 |    100.00
-----------+----------------------+----------
     Total |     1,953      1,215 |     3,168
           |     61.65      38.35 |    100.00
```

xttrans ed

```
. * Transition probabilities for a variable
. xttrans ed if ed>=12, freq

  years of |                        years of education
 education |        12         13         14         15         16         17 |     Total
-----------+------------------------------------------------------------------+----------
        12 |     1,116          0          0          0          0          0 |     1,116
           |    100.00       0.00       0.00       0.00       0.00       0.00 |    100.00
-----------+------------------------------------------------------------------+----------
        13 |         0        144          0          0          0          0 |       144
           |      0.00     100.00       0.00       0.00       0.00       0.00 |    100.00
-----------+------------------------------------------------------------------+----------
        14 |         0          0        264          0          0          0 |       264
           |      0.00       0.00     100.00       0.00       0.00       0.00 |    100.00
-----------+------------------------------------------------------------------+----------
        15 |         0          0          0         66          0          0 |        66
           |      0.00       0.00       0.00     100.00       0.00       0.00 |    100.00
-----------+------------------------------------------------------------------+----------
        16 |         0          0          0          0        480          0 |       480
           |      0.00       0.00       0.00       0.00     100.00       0.00 |    100.00
-----------+------------------------------------------------------------------+----------
        17 |         0          0          0          0          0        396 |       396
           |      0.00       0.00       0.00       0.00       0.00     100.00 |    100.00
-----------+------------------------------------------------------------------+----------
     Total |     1,116        144        264         66        480        396 |     2,466
           |     45.26       5.84      10.71       2.68      19.46      16.06 |    100.00
```

Within and Between R^2

Stata's *xtreg* command calculates the following three R^2 measures:

$$\text{Within } R^2: \quad \rho^2\left\{(y_{it}-\overline{y}_i),(x_{it}-\overline{x}_i)\hat{\beta}\right\}, \tag{16}$$

$$\text{Between } R^2: \quad \rho^2\left\{\overline{y}_i,\overline{x}_i\hat{\beta}\right\}, \tag{17}$$

$$\text{Overall } R^2: \quad \rho^2\left\{y_{it},x_{it}\hat{\beta}\right\}, \tag{18}$$

where $\rho^2(x, y)$ denotes the squared correlation between x and y.

LSDV and First-Difference Estimators

There are two other estimators that also allow the unobserved fixed-effect to be correlated with the regressors:

1. Least-squares dummy variables (LSDV) estimator
2. First-difference (FD) estimator

Both estimators are also widely used in practice but share the caveats of the fixed effects estimator.

The Dummy Variables Regression

- The LSDV regression considers the unobserved effects a_i as parameters to be estimated.
- It directly estimates $y_{it} = \alpha + x_{it}\beta + a_i + \epsilon_{it}$ adding a dummy for each cross-sectional unit i.

The Dummy Variables Regression

- The LSDV regression gives us exactly the same estimate of β as the fixed-effects estimator.
- It does not allow us to estimate the coefficients of time-invariant regressors (why?).

Use *areg* or *reg* to estimate the union wage premium using the LSDV regression!

The LSDV Estimates

```
. * LSDV model fitted using areg
. areg lwage exp exp2 wks ed union ind occ, absorb(id)
note: ed omitted because of collinearity

Linear regression, absorbing indicators                  Number of obs =    3696
                                                          F(   6,  3162) = 1004.25
                                                          Prob > F      =  0.0000
                                                          R-squared     =  0.8950
                                                          Adj R-squared =  0.8773
                                                          Root MSE      =  .15348

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .1149389   .0026801    42.89   0.000     .1096841    .1201938
        exp2 |  -.0004347   .0000584    -7.44   0.000    -.0005491   -.0003202
         wks |   .0004693   .0006576     0.71   0.476    -.0008201    .0017587
          ed |  (omitted)
       union |   .0316998   .0159769     1.98   0.047     .0003736     .063026
         ind |   .0182395   .0160431     1.14   0.256    -.0132165    .0496954
         occ |  -.0113013   .0146455    -0.77   0.440    -.0400169    .0174144
       _cons |   4.600501    .044283   103.89   0.000     4.513675    4.687328
-------------+----------------------------------------------------------------
          id |      F(527, 3162) =     39.191   0.000         (528 categories)
```

First-difference Estimator

Consider again our basic linear panel model

$$y_{it} = \alpha + x_{it}\beta + a_i + \epsilon_{it}. \tag{19}$$

and lag it one period:

$$y_{i,t-1} = \alpha + x_{i,t-1}\beta + a_i + \epsilon_{i,t-1}. \tag{20}$$

First-difference Estimator (ctd.)

Now subtract equation (20) from (19) to obtain the first-differenced equation...

$$\begin{aligned}(y_{it} - y_{i,t-1}) &= (\alpha - \alpha) + (x_{it} - x_{i,t-1})\beta + (a_i - a_i) \\ &\quad +(\epsilon_{it} - \epsilon_{i,t-1}) \\ &= (x_{it} - x_{i,t-1})\beta + (\epsilon_{it} - \epsilon_{i,t-1}) \qquad (21)\end{aligned}$$

Because a_i has been eliminated, OLS leads to consistent estimates of β even if x_{it} is correlated with a_i!

Re-Estimate the Union Wage Premium...

Now use *Stata* to re-estimate the union wage premium using the model in first differences:

$$
\begin{aligned}
(w_{it} - \overline{w_i}) &= (x_{it} - x_{i,t-1})\beta + \gamma(Union_{it} - Union_{i,t-1}) \\
&\quad +(\epsilon_{it} - \epsilon_{i,t-1}). \qquad (22)
\end{aligned}
$$

You should use the time-series operator for differences *D*...

The First-difference Estimates

```
. ******* 5. FIRST DIFFERENCE ESTIMATOR
.
. sort id t

. * First-differences estimator
. regress D.(lwage exp exp2 wks ed union ind occ), noconstant
note: _delete omitted because of collinearity

      Source |       SS       df       MS              Number of obs =    3168
-------------+------------------------------           F(  6,  3162) =  149.83
       Model |  30.2606538     6  5.04344229           Prob > F      =  0.0000
    Residual |  106.438928  3162    .0336619           R-squared     =  0.2214
-------------+------------------------------           Adj R-squared =  0.2199
       Total |  136.699582  3168   .04315012           Root MSE      =  .18347

------------------------------------------------------------------------------
     D.lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |
         D1. |   .1192783    .006853    17.41   0.000     .1058416    .1327151
             |
        exp2 |
         D1. |  -.0005583   .0001491    -3.75   0.000    -.0008506    -.000266
             |
         wks |
         D1. |  -.0003622   .0006187    -0.59   0.558    -.0015752    .0008508
             |
          ed |
         D1. |  (omitted)
             |
       union |
         D1. |   .0162718   .0162282     1.00   0.316    -.0155471    .0480907
             |
         ind |
         D1. |   .0144281   .0168265     0.86   0.391    -.0185639    .0474201
             |
         occ |
         D1. |  -.0158051   .0147561    -1.07   0.284    -.0447376    .0131274
------------------------------------------------------------------------------
```

Excursion: The Between Estimator

- The antipode to the within estimator is the between estimator.
- The between estimator uses only the cross-section variation in the data.
- To obtain the between model, average the basic linear panel model:

$$\bar{y}_i = \alpha + \bar{x}_i \beta + a_i + \bar{\epsilon}_i. \tag{23}$$

- The between estimator is simply the OLS estimator in this model (*xtreg, be* in *Stata*).

Excursion: The Between Estimator (ctd.)

- The between estimator is only consistent if the error $a_i + \bar{\epsilon}_i$ is uncorrelated with $\overline{x}_i$.
- Even if the between estimator is consistent, we have more efficient estimators at hand (pooled OLS and RE).
- The between estimator is rarely used in practice but is actually an input into the RE estimator that we study now...

The Random Effects Estimator

- The FE / LSDV / FD estimators wipe out *between variation* (i.e., variation across cross-sections).
- If we believe that the a_i's are not correlated with the regressors, we should resort to more efficient estimators.
- One possibility is pooled OLS with clustered-robust standard errors.
- A more efficient estimator is the Random Effects (RE) estimator.

The Random Effects Estimator (ctd.)

Consider again our basic linear panel model:

$$y_{it} = \alpha + x_{it}\beta + a_i + \epsilon_{it}. \tag{24}$$

- Now suppose that a_i is *purely random* with mean 0 and variance σ_α^2.
- Thus, a_i is assumed to be **uncorrelated** with the regressors.

The Random Effects Estimator (ctd.)

- Yet, the composite error $v_{it} = a_i + \epsilon_{it}$ is still serially correlated.
- Under the random effects assumption, $Var(v_{it}) = \sigma_\alpha^2 + \sigma_\epsilon^2$ and $Cov(u_{it}, u_{is}) = \sigma_\alpha^2,\ s \neq t$.
- The intraclass correlation of the error is thus given by

$$\rho_v = Cor(v_{it}, v_{is}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}. \tag{25}$$

The Random Effects Estimator (ctd.)

- The Random Effects Model is the OLS estimator in a model transformed to have serially uncorrelated errors.
- The transformed model is given by (see Wooldridge, 2002):

$$\begin{aligned}(y_{it} - \hat{\theta}_i \overline{y_i}) &= (1-\hat{\theta}_i)\alpha + (x_{it} - \hat{\theta}_i \overline{x}_i)\beta + (1-\hat{\theta}_i)a_i \\ &+ (\epsilon_{it} - \hat{\theta}_i \overline{\epsilon}_i) \end{aligned} \tag{26}$$

where $\hat{\theta}_i$ is a consistent estimator of

$$\theta = 1 - \sqrt{\sigma_\alpha^2 / (T\sigma_\alpha^2 + \sigma_\epsilon^2)}, \tag{27}$$

which is bounded by 0 and 1.

The Random Effects Estimator (ctd.)

It is worth noting that the RE model is a *weighted* average of the within and the between model:

$$
\begin{aligned}
(y_{it} - \overline{y_i}) + (1 - \hat{\theta}_i)\overline{y}_i &= (x_{it} - \overline{x}_i)\beta + (\epsilon_{it} - \overline{\epsilon}_i) \\
&\quad + (1 - \hat{\theta}_i)(\alpha + \overline{x}_i\beta + a_i + \overline{\epsilon}_i). \quad (28)
\end{aligned}
$$

The Random Effects Estimator (ctd.)

Before you apply the RE estimator in *Stata* using *xtreg*, *re* notice that...

- ...for $\theta = 0$, we obtain pooled OLS (which is an *unweighted* average of the between and within model):

$$y_{it} = \alpha + x_{it}\beta + a_i + \epsilon_{it}. \tag{29}$$

- ...for $\theta = 1$, we obtain the within model:

$$(y_{it} - \overline{y_i}) = (x_{it} - \overline{x}_i)\beta + (\epsilon_{it} - \overline{\epsilon}_i). \tag{30}$$

Random Effects Estimates

```
. ******* 6. RANDOM EFFECTS ESTIMATORS
.
. * Random-effects estimator
. xtreg lwage exp exp2 wks ed union ind occ, re theta

Random-effects GLS regression                   Number of obs      =      3696
Group variable: id                              Number of groups   =       528

R-sq:  within  = 0.6259                         Obs per group: min =         7
       between = 0.1823                                        avg =       7.0
       overall = 0.1887                                        max =         7

Random effects u_i ~ Gaussian                   Wald chi2(7)       =   2332.16
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000
theta              = .80072752

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .0825226   .0030822    26.77   0.000     .0764816    .0885636
        exp2 |  -.0007411   .0000676   -10.96   0.000    -.0008737   -.0006086
         wks |   .0006109   .0008368     0.73   0.465    -.0010292     .002251
          ed |   .1057496   .0062768    16.85   0.000     .0934474    .1180518
       union |   .0646388   .0181938     3.55   0.000     .0289796     .100298
         ind |   .0013936   .0179769     0.08   0.938    -.0338405    .0366276
         occ |  -.0348985   .0176442    -1.98   0.048    -.0694806   -.0003165
       _cons |   4.059829   .1030828    39.38   0.000     3.857791    4.261868
-------------+----------------------------------------------------------------
     sigma_u |  .28526616
     sigma_e |   .1534777
         rho |  .77551783   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Estimates of Variance Components

- The composite error $a_i + \epsilon_{it}$ is referred to as $u_i + e_{it}$ in the *Stata* output.
- *Stata* provides estimates of the std. dev. of a_i (denoted *sigma_u*) and ϵ_{it} (denoted *sigma_e*).
- The output *rho* is an estimate of the intraclass correlation $\rho_v = Cor(v_{it}, v_{is}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}$.
- For the RE model, the estimated $\hat{\theta}_i$ can be obtained using the *theta* option.

Comparison of Estimates

- Now recap and store the different estimates by typing:

  . quietly regress lwage exp exp2 wks ed union ind occ, vce(cluster id)
  . estimates store OLS_rob
  . quietly xtreg lwage exp exp2 wks ed union ind occ, fe
  . estimates store FE
  . quietly xtreg lwage exp exp2 wks ed union ind occ, re
  . estimates store RE
  . estimates table OLS_rob FE RE, b se stats(N r2 $r2_o$ $r2_b$ $r2_w$)

Comparison of Estimates (ctd.)

Variable	OLS_rob	FE	RE
exp	0.0415	0.1135	0.0869
	0.0050	0.0025	0.0028
exp2	-0.0007	-0.0004	-0.0008
	0.0001	0.0001	0.0001
wks	0.0076	0.0008	0.0010
	0.0018	0.0006	0.0008
ed	0.0704	(omitted)	0.1086
	0.0060		0.0062
union	0.1517	0.0321	0.0615
	0.0275	0.0149	0.0170
ind	0.0966	0.0182	0.0124
	0.0259	0.0155	0.0172
occ	-0.1447	-0.0214	-0.0411
	0.0318	0.0137	0.0164
_cons	4.9149	4.5930	3.9015
	0.1413	0.0400	0.0983
N	4165	4165	4165
r2	0.3205	0.6574	
r2_o		0.0491	0.1933
r2_b		0.0289	0.1833
r2_w		0.6574	0.6287

legend: b/se

Hypothesis Testing

In applied work, you may want to test whether...

- ...the cross-sectional intercepts differ at all.
- ...to use the random or fixed effects estimators in case intercepts differ.

RE/FE or Pooled OLS?

- There is an easy way to test whether the cross-sectional intercepts are different from each other.
- Just run *xtreg*, *fe* and calculate the corresponding dummy variable estimates.
- Do a simple F test to test whether the coefficients are identical.
- *Stata* actually reports this test below the regression output of *xtreg*, *fe*.
- If the coefficients do not differ, you can use pooled OLS.

RE/FE or Pooled OLS? (ctd.)

```
. * Within or FE estimator
. xtreg lwage exp exp2 wks ed union ind occ, fe
note: ed omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =      3696
Group variable: id                              Number of groups   =       528

R-sq:  within  = 0.6558                         Obs per group: min =         7
       between = 0.0228                                        avg =       7.0
       overall = 0.0427                                        max =         7

                                                F(6,3162)          =   1004.25
corr(u_i, Xb)  = -0.9223                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .1149389   .0026801    42.89   0.000     .1096841    .1201938
        exp2 |  -.0004347   .0000584    -7.44   0.000    -.0005491   -.0003202
         wks |   .0004693   .0006576     0.71   0.476    -.0008201    .0017587
          ed |  (omitted)
       union |   .0316998   .0159769     1.98   0.047     .0003736     .063026
         ind |   .0182395   .0160431     1.14   0.256    -.0132165    .0496954
         occ |  -.0113013   .0146455    -0.77   0.440    -.0400169    .0174144
       _cons |   4.600501    .044283   103.89   0.000     4.513675    4.687328
-------------+----------------------------------------------------------------
     sigma_u |   1.047369
     sigma_e |   .1534777
         rho |  .97897847   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(527, 3162) =    33.82           Prob > F = 0.0000
```

Random Effects or Fixed Effects?

- If the a_i's are correlated with the regressors, only the FE estimator is consistent.
- If the a_i's are uncorrelated with the regressors, both the FE and the RE estimators are consistent but the RE estimator is more efficient.
- The reason is simply that the RE estimator uses both the within and the between variation of the data.

Random Effects or Fixed Effects?

- The Hausman test uses the fact that both estimators are consistent under the null hypothesis of no correlation.
- Under the null hypothesis, the two estimates of β should thus not differ systematically.
- We can implement the test in *Stata* using *hausman*.
- For the technical details of the test, you may consult Greene (2008).

Random Effects or Fixed Effects? (ctd.)

```
. ******* 8. HAUSMAN TEST
.
. * Hausman test assuming RE estimator is fully efficient under null hypothesis
. hausman FE RE, sigmamore

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       FE           RE         Difference          S.E.
-------------+----------------------------------------------------------------
         exp |    .1134669      .086851         .026616        .0013417
        exp2 |   -.0004198    -.0007766        .0003567        .0000299
         wks |    .0008245     .0010262       -.0002017          .00012
       union |    .0321123     .0615498       -.0294375        .0084284
         ind |    .0182378     .0124311        .0058066        .0093685
         occ |   -.0214375    -.0410889        .0196515        .0059918
------------------------------------------------------------------------------
                           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(6) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =     1578.13
                Prob>chi2 =      0.0000
```

Fixed Effects or First Differencing?

- For $T = 2$, the two estimators are the same.
- For $T \geq 2$, the estimators are not the same but both are unbiased and consistent.
- For large N, and small T, FE is more efficient **if** the idiosyncratic error ϵ_{it} is serially uncorrelated.
- Yet, there is no easy way to choose between the two estimators and both are widely used in practice.

Extension: Two-way-effects Model

A standard extension of the basic linear panel model is to allow the intercept to vary not only over individuals but also over time:

$$y_{it} = \alpha + x_{it}\beta + \lambda_t + a_i + \epsilon_{it}. \tag{31}$$

- The time effects λ_t shift the intercept over time and affect all micro-units uniformly.
- Examples: business cycle movements, common trend in wages...
- In short panels, we usually include a full set of time dummies among the x_{it}'s.

Extension: Two-way-effects Model (ctd.)

- Re-estimate the wage equation including a full set of time dummies.
- You can either generate the dummies by hand or use the *i.* option of *Stata*.

FE Estimates with Time Dummies

```
. ******* 9. TWO-WAY-EFFECTS MODEL
.
. * Within or FE estimator with time dummies
. xtreg lwage exp exp2 wks ed union ind occ i.t, fe
note: ed omitted because of collinearity
note: 7.t omitted because of collinearity

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

R-sq:  within  = 0.6605                         Obs per group: min =         7
       between = 0.0287                                        avg =       7.0
       overall = 0.0493                                        max =         7

                                                F(11,3559)         =    629.58
corr(u_i, Xb)  = -0.9089                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .1117254   .0026174    42.69   0.000     .1065936    .1168571
        exp2 |   -.000401   .0000546    -7.35   0.000     -.000508   -.000294
         wks |   .0006692   .0005994     1.12   0.264     -.000506    .0018444
          ed |  (omitted)
       union |   .0288605   .0148838     1.94   0.053    -.0003212    .0580422
         ind |   .0198261   .0154044     1.29   0.198    -.0103763    .0500285
         occ |  -.0193217   .0137094    -1.41   0.159    -.0462008    .0075574
             |
           t |
          2  |  -.0081074   .0081694    -0.99   0.321    -.0241246    .0079098
          3  |   .0258101   .0077694     3.32   0.001     .0105772     .041043
          4  |   .0286159   .0076429     3.74   0.000      .013631    .0436009
          5  |   .0235754   .0077717     3.03   0.002     .0083381    .0388128
          6  |    .007083   .0081633     0.87   0.386    -.0089221    .0230881
          7  |  (omitted)
             |
       _cons |   4.613659   .0431638   106.89   0.000      4.52903    4.698287
-------------+----------------------------------------------------------------
     sigma_u |  1.0261585
     sigma_e |  .15150311
         rho |  .97866713   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(594, 3559) =    29.82           Prob > F = 0.0000
```

RE Estimates with Time Dummies

```
. * RE estimator with time dummies
. xtreg lwage exp exp2 wks ed union ind occ i.t, re

Random-effects GLS regression                   Number of obs      =      4165
Group variable: id                              Number of groups   =       595

R-sq:  within  = 0.6600                         Obs per group: min =         7
       between = 0.3420                                        avg =       7.0
       overall = 0.4279                                        max =         7

Random effects u_i ~ Gaussian                   Wald chi2(13)      =   7163.49
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .0289861   .0025233    11.49   0.000     .0240404    .0339317
        exp2 |  -.0004355     .00005    -8.72   0.000    -.0005334   -.0003375
         wks |   .0009599   .0005954     1.61   0.107    -.0002072    .0021269
          ed |   .0735711   .0049521    14.86   0.000     .0638651    .0832771
       union |   .0500422   .0134164     3.73   0.000     .0237465    .0763379
         ind |   .0409469   .0136322     3.00   0.003     .0142283    .0676656
         occ |  -.0339167   .0129552    -2.62   0.009    -.0593084   -.0085249
             |
           t |
          2  |   .0756715   .0089433     8.46   0.000      .058143    .0931999
          3  |   .1931833   .0092131    20.97   0.000      .175126    .2112406
          4  |    .279892   .0096431    29.03   0.000      .260992    .2987921
          5  |   .3588786   .0101782    35.26   0.000     .3389297    .3788275
          6  |   .4269871   .0108193    39.47   0.000     .4057817    .4481926
          7  |   .5040351   .0115335    43.70   0.000     .4814297    .5266404
             |
       _cons |   5.055152   .0812668    62.20   0.000     4.895871    5.214432
-------------+----------------------------------------------------------------
     sigma_u |  .30410708
     sigma_e |  .15150311
         rho |  .80115798   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Outlook: Advanced Panel Data Analysis

Two major topics in applied panel data analysis were not covered:

1. Panel IV estimation
2. Dynamic panel estimation

Panel IV estimation

The fixed effects estimator (and related estimators) has some major shortcomings:

- It does not allow us to estimate the coefficients of time-invariant regressors.
- It does not solve the problem of **time-varying** omitted variables.

Panel IV estimation (ctd.)

Suppose you want to estimate the effect of job training on worker productivity:

$$lscrap_{it} = \alpha + \gamma hrsemp_{it} + a_i + \epsilon_{it}, \quad (32)$$

using the dataset *scrap.dta*.

- i now refers to a single firm.
- *lscrap* is the log of the scrap rate of a firm (a productivity measure).
- *hrsemp* is hours of job training per employee.

Panel IV estimation (ctd.)

- The unobserved fixed effect may capture, e.g., the (average) ability of a firm's workforce.
- As $hrsemp_{it}$ may well be correlated with a_i (why?), we estimate the fixed effects model:

$$(lscrap_{it} - \overline{lscrap}_i) = \gamma(hrsemp_{it} - \overline{hrsemp}_i) + (\epsilon_{it} - \bar{\epsilon}_i). \quad (33)$$

xtset and *xtdescribe* the Data

```
. use "scrap.dta", clear

. xtset  fcode year
       panel variable:  fcode (unbalanced)
        time variable:  year, 1987 to 1988
                delta:  1 unit

. xtdescribe

   fcode:  410523, 410563, ..., 419483                       n =         47
    year:  1987, 1988, ..., 1988                             T =          2
           Delta(year) = 1 unit
           Span(year)  = 2 periods
           (fcode*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       2       2         2         2       2       2

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
       45     95.74   95.74 |  11
        1      2.13   97.87 |  .1
        1      2.13  100.00 |  1.
 ---------------------------+---------
       47    100.00         |  XX
```

FE Estimates

```
. xtreg  lscrap hrsemp, fe

Fixed-effects (within) regression               Number of obs      =        92
Group variable: fcode                           Number of groups   =        47

R-sq:  within  = 0.1193                         Obs per group: min =         1
       between = 0.0160                                        avg =       2.0
       overall = 0.0243                                        max =         2

                                                F(1,44)            =      5.96
corr(u_i, Xb)  = 0.0294                         Prob > F           =    0.0187

------------------------------------------------------------------------------
      lscrap |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hrsemp |  -.0097174   .0039812    -2.44   0.019     -.017741   -.0016937
       _cons |   .6737459    .064658    10.42   0.000     .5434363    .8040555
-------------+----------------------------------------------------------------
     sigma_u |  1.4400308
     sigma_e |  .43425379
         rho |  .91664268   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(46, 44) =    20.80              Prob > F = 0.0000
```

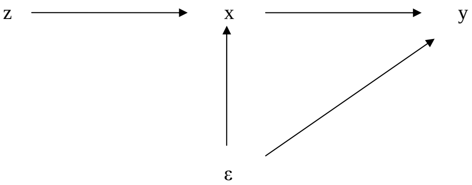
Panel IV estimation (ctd.)

- The fixed effects estimator will still be biased if $hrsemp_{it} - \overline{hrsemp}_t$ is correlated with the time-varying error ϵ_{it}.
- A firm might, for instance, increase productivity by hiring more skilled workers and simultaneously reduce job training.
- In that case, we have to resort to Panel IV estimation...

The IV Idea

We have to find an instrument z that is...

1. ...correlated with the endogenous variable x $(hrsemp_{it} - \overline{hrsemp}_i)$.
2. ...uncorrelated with the error term ϵ (ϵ_{it}).



Panel IV Estimation in Stata

- In our data, the dummy *grant* that indicates whether a firm received a job training grant by the state may provide a valid instrument (under which assumptions?).
- The *xtivreg* command allows us to combine the fixed effect transformation with IV estimation.
- Replace *hrsempl* with *(hrsempl = grant)* to instruct Stata that *hrsempl* should be instrumented by *grant*.

Correlation between *grant* and *hrsemp*

```
. xtreg  hrsemp grant, fe

Fixed-effects (within) regression               Number of obs      =        92
Group variable: fcode                           Number of groups   =        47

R-sq:  within  = 0.4836                         Obs per group: min =         1
       between = 0.0801                                        avg =       2.0
       overall = 0.2143                                        max =         2

                                                F(1,44)            =     41.21
corr(u_i, Xb)  = -0.0875                        Prob > F           =    0.0000

------------------------------------------------------------------------------
      hrsemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grant |   26.01751   4.053008     6.42   0.000     17.84921    34.18581
       _cons |   6.787258   1.441732     4.71   0.000     3.881638    9.692877
-------------+----------------------------------------------------------------
     sigma_u |  14.822833
     sigma_e |  11.816449
         rho |  .61143602   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(46, 44) =     3.11             Prob > F = 0.0001
```

Panel IV Estimation Using *xtivreg*

```
. xtivreg  lscrap (hrsemp=grant), fe

Fixed-effects (within) IV regression         Number of obs      =        92
Group variable: fcode                        Number of groups   =        47

R-sq:  within  = 0.0783                      Obs per group: min =         1
       between = 0.0160                                     avg =       2.0
       overall = 0.0243                                     max =         2

                                             Wald chi2(1)       =    153.69
corr(u_i, Xb)  = -0.0202                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      lscrap |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hrsemp |  -.0154088   .0058563    -2.63   0.009     -.026887   -.0039306
       _cons |   .7397372   .0821938     9.00   0.000     .5786403    .9008341
-------------+----------------------------------------------------------------
     sigma_u |  1.4405516
     sigma_e |  .44422418
         rho |  .91316478   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F  test that all u_i=0:     F(46,44) =    19.87           Prob > F    = 0.0000
------------------------------------------------------------------------------
Instrumented:   hrsemp
Instruments:    grant
------------------------------------------------------------------------------
```

Dynamic panel estimation

Panel data enables us to estimate parameters of dynamic models with lagged dependent variables such as:

$$y_{it} = \alpha + \upsilon y_{i,t-1} + x_{it}\beta + a_i + \epsilon_{it}. \quad (34)$$

Dynamic panel estimation (ctd.)

Dynamic models are usually estimated in first-differences so as to erase the unobserved effect a_i:

$$(y_{it} - y_{i,t-1}) = (x_{it} - x_{i,t-1})\beta + \upsilon(y_{i,t-1} - y_{i,t-2}) + (\epsilon_{it} - \epsilon_{i,t-1}). \quad (35)$$

As the lagged dependent variable is still correlated with the error term, IV estimation is required to obtain consistent estimates.

Dynamic panel estimation (ctd.)

- In practice, appropriate lags of the dependent variable are used as instruments.
- In our example, $y_{i,t-2} - y_{i,t-3}$ might be an appropriate instrument for $y_{i,t-1} - y_{i,t-2}$.
- A widely used estimator for dynamic panel models is the Arellano-Bond estimator (*xtabond* in *Stata*).

Omitted Variable Bias: An Example

Suppose that a person's wage w_i is a function of his education ed_i and his IQ a_i:

$$w_i = \alpha + \beta_1 ed_i + \beta_2 a_i + u_i. \qquad (36)$$

As you do not have on IQ, you instead estimate:

$$w_i = \alpha + \beta_1 ed_i + \tilde{u}_i. \qquad (37)$$

Omitted Variable Bias: An Example (ctd.)

Now suppose that IQ is related to education through the following model:

$$a_i = \gamma + \delta_1 ed_i + \epsilon_i. \tag{38}$$

Then the regression that you actually run can be written as:

$$\begin{aligned} w_i &= \alpha + \beta_1 ed_i + \beta_2(\gamma + \delta_1 ed_i + \epsilon_i) + \tilde{u}_i \\ &= (\alpha + \beta_2\gamma) + (\beta_1 + \beta_2\delta_1)ed_i + (\tilde{u}_i + \beta_2\epsilon_i) \end{aligned} \tag{39}$$

Omitted Variable Bias: An Example (ctd.)

- The estimated effect of education on wages is thus $\beta_1 + \beta_2\delta_1$.
- Education and IQ are usually positively correlated, i.e., $\delta_1 > 0$.
- IQ should also have a positive effect on wages, i.e., $\beta_2 > 0$.
- It thus follows that our estimated effect of education is too large as $\beta_1 + \beta_2\delta_1 > \beta_1$.

Why does *xtreg*, *fe* Report an Intercept?

Stata actually fits the model:

$$(y_{it} - \overline{y_i} + \overline{\overline{y}}) = \alpha + (x_{it} - \overline{x}_i + \overline{\overline{x}})\beta + (\epsilon_{it} - \overline{\epsilon}_i + \overline{a} + \overline{\overline{\epsilon}}), \quad (40)$$

where $\overline{\overline{z}} = N^{-1}\overline{z}$ is the 'grand' mean of some variable z and *Stata* imposes the constraint $\overline{a} = \frac{1}{N}\sum_{i=1}^{N} a_i = 0$.

Notice that the slope estimate β is not affected by the 'transformation'.