ECON60052: Cross Section Econometrics

Lecture 4: Advanced Panel data methods

Nicky L. Grant

Semester 2

The last lecture

Discussed estimation of treatment effects using panel data models

Difference-in-Difference estimator

Outlined the notion of heterogeneity bias and how to overcome this using panel data estimators

This lecture

We discuss 3 estimators:

- First Difference (FD) estimator
- Fixed Effects (FE) estimator
- Random Effects (RE) estimator

First Difference Estimator

Assume we have a balanced cross-section (all cross section units observed all T time periods)

$$y_{it} = \alpha_i + \alpha_2 d2_t + \alpha_3 d3_t + \cdots + \alpha_T dT_t + \boldsymbol{x}_{it}\boldsymbol{\beta} + u_{it} \quad i = 1, ..N; t = 1, .., T \tag{1}$$

Where dt_j is time dummy that equals one when $t = j$ and zero otherwise.

FD estimator is OLS applied to (1) in first differences

$$\Delta y_{it} = \alpha_2 \Delta d2_t + \alpha_3 \Delta d3_t + \cdots + \alpha_T \Delta dT_t + \Delta \boldsymbol{x}_{it}\boldsymbol{\beta} + \Delta u_{it}, \quad t = 2, 3, \ldots, T \tag{2}$$

Where $\Delta y_{it} \equiv y_{it} - y_{i,t-1}$ etc.

The first obs for each i is lost: we have $T - 1$ time periods on each unit i hence $N(T - 1)$ obs.

Data example

i	t	y_{it}	1	$d1_t \ldots$	dT_t	$\boldsymbol{x}_{it}$	Δy_{it}	$\Delta 1$	$\Delta d1_t \ldots$	ΔdT_t	$\Delta \boldsymbol{x}_{it}$
1	1	y_{11}	1	1 …	0	$\boldsymbol{x}_{11}$	.	.	.	.	.
1	2	y_{12}	1	0 …	0	$\boldsymbol{x}_{12}$	Δy_{12}	0	–1 …	0	$\Delta \boldsymbol{x}_{12}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	T	y_{1T}	1	0 …	1	$\boldsymbol{x}_{1T}$	Δy_{1T}	0	0 …	1	$\Delta \boldsymbol{x}_{1T}$
2	1	y_{21}	1	1 …	0	$\boldsymbol{x}_{21}$	.	.	.	.	.
2	2	y_{22}	1	0 …	0	$\boldsymbol{x}_{22}$	Δy_{22}	0	–1 …	0	$\Delta \boldsymbol{x}_{22}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	T	y_{2T}	1	0 …	1	$\boldsymbol{x}_{2T}$	Δy_{2T}	0	0 …	1	$\Delta \boldsymbol{x}_{2T}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	1	y_{N1}	1	1 …	0	$\boldsymbol{x}_{N1}$	.	.	.	.	.
N	2	y_{N2}	1	0 …	0	$\boldsymbol{x}_{N2}$	Δy_{N2}	0	–1 …	0	$\Delta \boldsymbol{x}_{N2}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	T	y_{NT}	1	0 …	1	$\boldsymbol{x}_{NT}$	Δy_{NT}	0	0 …	1	$\Delta \boldsymbol{x}_{NT}$

Issues in FD when T>2

For errors in (2) to be uncorrelated we need Δu_{it} are uncorrelated over time

Hence requires u_{it} is a Random Walk

Can test for serial correlation in Δu_{it} by estimating ρ by OLS in

$$\Delta u_{it} = \rho u_{i,t-1} + e_{it}$$

and testing $H_0 : \rho = 0$ against $H_A \rho \neq 0$ using the Dickey Fuller test

Possible if $T \geq 3$

Fixed Effects Estimator

For simplicity take the special case of (1):

$$y_{it} = a_i + \beta x_{it} + u_{it} \qquad i = 1, \ldots, N; \quad t = 1, \ldots, T \tag{3}$$

The panel is balanced and $T \geq 2$

There are N individuals

By averaging the data over T observations for each i, we get

$$\overline{y}_i = a_i + \beta \overline{x}_i + \overline{u}_i \tag{4}$$

where $\overline{y}_i \equiv \sum_t y_{it}/T$ ...

These are **group means**

Fixed Effects Estimator

We remove ('sweep out') the fixed-effects by subtracting (2) from (1):

$$y_{it} - \overline{y}_i = \beta(x_{it} - \overline{x}_i) + u_{it} - \overline{u}_i \qquad i = 1, \ldots, N; \quad t = 1, \ldots, T$$

$$\ddot{y}_{it} = \beta \ddot{x}_{it} + \ddot{u}_{it} \qquad i = 1, \ldots, N; \quad t = 1, \ldots, T \tag{5}$$

where $\ddot{y}_{it} \equiv y_{it} - \overline{y}_i$ is the **time-demeaned data** on $y \ldots$

The data have been transformed so that only **within group** variation remains

where "group" refers to an individual and "within" refers to time

The FE (within group) estimator is OLS (5):

$$\widehat{\beta}_w = \frac{W_{xy}}{W_{xx}} \equiv \frac{\sum_i \sum_t (x_{it} - \overline{x}_i)(y_{it} - \overline{y}_i)}{\sum_i \sum_t (x_{it} - \overline{x}_i)^2}$$

$$\widehat{a}_i = \overline{y}_i - \widehat{\beta}_w \overline{x}_i \qquad i = 1, \ldots, N \tag{6}$$

Data Example

i	t	y_{it}	1	$d1_t \ldots$	dT_t	$\boldsymbol{x}_{it}$	$\overline{y}_i$	$\ddot{y}_{it}$	$\overline{y}_i - \overline{y}$
1	1	y_{11}	1	1 ...	0	$\boldsymbol{x}_{11}$	$\overline{y}_1$	$y_{11} - \overline{y}_1$	
1	2	y_{12}	1	0 ...	0	$\boldsymbol{x}_{12}$	$\overline{y}_1$	$y_{12} - \overline{y}_1$	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
1	T	y_{1T}	1	0 ...	1	$\boldsymbol{x}_{1T}$	$\overline{y}_1$	$y_{1T} - \overline{y}_1$	$\overline{y}_1 - \overline{y}$
2	1	y_{21}	1	1 ...	0	$\boldsymbol{x}_{21}$	$\overline{y}_2$	$y_{21} - \overline{y}_2$	
2	2	y_{22}	1	0 ...	0	$\boldsymbol{x}_{22}$	$\overline{y}_2$	$y_{22} - \overline{y}_2$	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
2	T	y_{2T}	1	0 ...	1	$\boldsymbol{x}_{2T}$	$\overline{y}_2$	$y_{2T} - \overline{y}_2$	$\overline{y}_2 - \overline{y}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
N	1	y_{N1}	1	1 ...	0	$\boldsymbol{x}_{N1}$	$\overline{y}_N$	$y_{N1} - \overline{y}_N$	
N	2	y_{N2}	1	0 ...	0	$\boldsymbol{x}_{N2}$	$\overline{y}_N$	$y_{N2} - \overline{y}_N$	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
N	T	y_{NT}	1	0 ...	1	$\boldsymbol{x}_{NT}$	$\overline{y}_N$	$y_{NT} - \overline{y}_N$	$\overline{y}_N - \overline{y}$
mean							$\overline{y}$	0	0

Fixed Effects and Stata

This the **fixed effects estimator** or **within estimator**

It is also called the **covariance estimator**, **least squares dummy variables estimator** (LSDV) (see later)

The LSDV estimator estimates each a_i for $i = 1, \ldots, N$ by including as regressors a dummy for an intercept for each i).

We can show the FE and LSDV are equivalent

Stata's command for fixed-effects/within estimator is

xtreg y x, fe i( )

Further comments

Note when (5) is estimated it appears $df = NT - k$

The correct df is $df = N(T-1) - k$

This is because implicitly the a_i is estimated for $i = 1, ..N$

If OLS is used on (4), $\widehat{\sigma}^2$ is incorrectly computed as

$$\text{SSR}/(NT - k)$$

and so need to correct standard errors by

$$\sqrt{\frac{NT-k}{N(T-1)-k}} \approx \sqrt{\frac{T}{T-1}}$$

Packages like Stata make this correction automatically ...

FE Assumptions

The strict exogeneity assumption in this model is

$$\mathsf{E}(u_{it}|\boldsymbol{x}_i, a_i) = \mathsf{E}(u_{it}|x_{i1}, \ldots, x_{iT}, a_i) = 0 \qquad t = 1, \ldots, T$$

$\Rrightarrow$ the fixed effects estimator is unbiased

$\hookrightarrow$ Strict exogeneity implies u_{it} is uncorrelated with $\boldsymbol{x}_i$ across *all* time periods

The FE estimator $\widehat{\boldsymbol{\beta}}_w$ is consistent providing *either* N or T "goes to infinity"

Advantages and Disadvantages of using FE

The fixed effects estimator allows for arbitrary correlations between a_i and $\boldsymbol{x}_i \equiv (x_{i1}, \ldots, x_{iT})$

Any variable that is constant is swept away by the fixed effects transformation: $\ddot{x}_{it} = 0$

BUT time-constant variables can be interacted with variables that do vary over time, in particular year dummies

When a full set of year dummies is included (less one), a variable whose change is constant cannot be included (such as age, experience)

i	t	m	$\ddot{m}$	a	$\ddot{a}$	$d1$	$\ddot{d1}$	$\ddot{d2}$	$\ddot{d3}$	$\ddot{d4}$	$\ddot{d5}$	$\ddot{md1}$
1	1	1	0	24	-2	1	4/5	-1/5	-1/5	-1/5	-1/5	4/5
1	2	1	0	25	-1	0	-1/5	4/5	-1/5	-1/5	-1/5	-1/5
1	3	1	0	26	0	0	-1/5	-1/5	4/5	-1/5	-1/5	-1/5
1	4	1	0	27	1	0	-1/5	-1/5	-1/5	4/5	-1/5	-1/5
1	5	1	0	28	2	0	-1/5	-1/5	-1/5	-1/5	4/5	-1/5
2	1	0	0	33	-2	1	4/5	-1/5	-1/5	-1/5	-1/5	0
2	2	0	0	34	-1	0	-1/5	4/5	-1/5	-1/5	-1/5	0
2	3	0	0	35	0	0	-1/5	-1/5	4/5	-1/5	-1/5	0
2	4	0	0	36	1	0	-1/5	-1/5	-1/5	4/5	-1/5	0
2	5	0	0	37	2	0	-1/5	-1/5	-1/5	-1/5	4/5	0

$$\ddot{a} = -2\ddot{d1} - \ddot{d2} + \ddot{d4} + 2\ddot{d5} \qquad 0 = \ddot{d1} + \ddot{d2} + \ddot{d3} + \ddot{d4} + \ddot{d5}$$

In this example, the coeff on age *a* cannot be identified together with the time dummies as changes in age are perfect linear relation to changes in time

Also, we cannot include all time dummy variables at a time $\Rightarrow$ the constant or a time dummy needs dropping

In the next example, we use the same BHPS1996/2006 data, but balance the data out

The estimates are identical to FD (because $T = 2$)

FE Example

```
. xtreg   lrhrpay  southeast  age age2 male married manual covered wave16 male16, robust i(pid) fe
note: male omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =      4290
Group variable: pid                             Number of groups   =      2145

R-sq:  within  = 0.4309                         Obs per group: min =         2
       between = 0.0428                                        avg =       2.0
       overall = 0.1463                                        max =         2

                                                F(8,2144)          =    193.77
corr(u_i, Xb)  = -0.0723                        Prob > F           =    0.0000

                                  (Std. Err. adjusted for 2145 clusters in pid)
------------------------------------------------------------------------------
             |               Robust
     lrhrpay |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   southeast |   .0767908   .0564227     1.36   0.174    -.0338581    .1874396
         age |    .055333   .0330886     1.67   0.095    -.0095561    .1202221
        age2 |  -.0008047   .0000626   -12.85   0.000    -.0009275   -.0006819
        male |           0  (omitted)
     married |    .050731   .0202659     2.50   0.012     .0109882    .0904738
      manual |  -.0258372   .0258402    -1.00   0.317    -.0765118    .0248373
     covered |   .1006201   .0229372     4.39   0.000     .0556386    .1456017
      wave16 |   .4661019   .3263089     1.43   0.153    -.1738132    1.106017
      male16 |   -.043198   .0195837    -2.21   0.028     -.081603   -.0047929
       _cons |   1.071068   1.170823     0.91   0.360       -1.225    3.367135
-------------+----------------------------------------------------------------
     sigma_u |  .46986452
     sigma_e |  .31928609
         rho |  .68410764   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

FD Example

```
. reg Dlrhrpay  Dsoutheast  Dage Dage2 Dmale Dmarried Dmanual Dcovered Dmale16, robust
note: Dmale omitted because of collinearity

Linear regression                                      Number of obs =    2145
                                                       F(  7,  2137) =   30.70
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1317
                                                       Root MSE      =  .45154

------------------------------------------------------------------------------
             |               Robust
    Dlrhrpay |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Dsoutheast |   .0767908   .0564623     1.36   0.174    -.0339359    .1875175
       Dage |    .055333   .0331118     1.67   0.095    -.0096018    .1202677
      Dage2 |  -.0008047   .0000627   -12.84   0.000    -.0009276   -.0006818
      Dmale |          0  (omitted)
   Dmarried |    .050731   .0202801     2.50   0.012     .0109602    .0905018
    Dmanual |  -.0258372   .0258584    -1.00   0.318    -.0765475     .024873
   Dcovered |   .1006201   .0229533     4.38   0.000     .0556069    .1456334
    Dmale16 |   -.043198   .0195975    -2.20   0.028      -.08163   -.0047659
      _cons |   .4661019    .326538     1.43   0.154    -.1742635    1.106467
------------------------------------------------------------------------------
```

Pooled OLS

When the model constrains all the fixed-effects equal, OLS is applied to

$$y_{it} = a + \beta x_{it} + u_{it} \qquad t = 1, \ldots, T \tag{7}$$

giving

$$\widehat{\beta} = \frac{T_{xy}}{T_{xx}} \equiv \frac{\sum_i \sum_t (x_{it} - \overline{x})(y_{it} - \overline{y})}{\sum_i \sum_t (x_{it} - \overline{x})^2}$$
$$\widehat{a} = \overline{y} - \widehat{\beta}\overline{x}.$$

Here both between and within variation are used.

↪ Consistency of Pooled OLS requires

$Cov(a_i, x_{it}) = 0$
$Cov(x_{it}, u_{it}) = 0$

```
. reg   lrhrpay  southeast  age age2 male married manual covered wave16 male16 if N==2, robust

Linear regression                                       Number of obs =    4290
                                                        F(  9,  4280) =  230.56
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.3427
                                                        Root MSE      =  .45671

------------------------------------------------------------------------------
             |               Robust
     lrhrpay |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   southeast |   .0840641   .0184034     4.57   0.000     .0479839    .1201443
         age |   .0767619    .004667    16.45   0.000     .0676123    .0859116
        age2 |  -.0009024    .000055   -16.42   0.000    -.0010102   -.0007947
        male |   .4276237   .019978    21.40   0.000     .3884565    .4667908
     married |   .0510472   .0157437     3.24   0.001     .0201814    .0819131
      manual |  -.3396322   .0156081   -21.76   0.000    -.3702321   -.3090323
     covered |   .1742004   .0142422    12.23   0.000     .1462783    .2021224
      wave16 |   .3120638   .0212828    14.66   0.000     .2703384    .3537892
      male16 |  -.0393342   .0278404    -1.41   0.158    -.0939158    .0152473
       _cons |   .2857594    .092003     3.11   0.002     .1053858     .466133
------------------------------------------------------------------------------
```

The dummy variables regression

Suppose we observe each individual twice. Now add N dummy variables for each individual:

i	t	y_{it}	1	$d1_t$	$d2_t$	$\boldsymbol{x}_{it}$	indiv dummies			
1	1	y_{11}	1	1	0	$\boldsymbol{x}_{11}$	1	0	...	0
2	1	y_{21}	1	1	0	$\boldsymbol{x}_{21}$	0	1	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
N	1	y_{N1}	1	1	0	$\boldsymbol{x}_{N1}$	0	0	...	1
1	2	y_{12}	1	0	1	$\boldsymbol{x}_{12}$	1	0	...	0
2	2	y_{22}	1	0	1	$\boldsymbol{x}_{22}$	0	1	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
N	2	y_{N2}	1	0	1	$\boldsymbol{x}_{N2}$	0	0	...	1

Least Squares Dummy Variable Regression

These dummies can be added to the pooled panel data model

If we estimate by OLS (dropping the constant), we get *exactly the same* estimates, standard errors, SSR, degrees of freedom, as the fixed effects estimator

FE is theoretically equivalent to LSDV

Note that we must drop the constant because

$$\sum_{1}^{n}(\text{indiv dummies})_{it} = 1, \qquad i = 1, \ldots, N;\ t = 1, 2.$$

There are not many advantages to doing LSDV, and it is not possible when N dummies are too many for the regression package

One does get direct estimates of $a_i \ldots$

Fixed Effects or Differencing?

When $T = 2$, FE and FD estimators are the same

When $T \geq 3$, the FE and FD estimators are different, but often close

Assuming both are unbiased the choice depends on the relative efficiency of the estimators:

- if the u_{it} are serially uncorrelated, in general we cannot tell if either FD or FE is more efficient
- If Δu_{it} are serially uncorrelated (i.e. u_{it} is a random walk), use FD, because OLS applied to differences is BLUE

We cannot test whether the u_{it} are serially uncorrelated, because we estimate $\ddot{u}_{it}$

For the FD estimate, we can test whether $\Delta\widehat{u}_{it}$ are serially uncorrelated

FE often preferred in unbalanced panels...

Fixed Effects with Unbalanced Panels

Individuals disappear from panels for various reasons (called **attrition**)

With an **Unbalanced Panel** the mechanics of FE estimation are much the same

Each individual has T_i obs, and

$$\overline{y}_i \equiv \sum_t y_{it}/T_i$$

For example, $T = 5$ but $T_i = 3$:

$$\begin{aligned}
\ddot{y}_{i1} &= y_{i1} - \overline{y}_i & \Delta y_{i1} &= \text{mv} \\
\ddot{y}_{i2} &= y_{i2} - \overline{y}_i & \Delta y_{i2} &= y_{i2} - y_{i1} \\
\ddot{y}_{i3} &= \text{mv} & \Delta y_{i3} &= \text{mv} \\
\ddot{y}_{i4} &= \text{mv} & \Delta y_{i4} &= \text{mv} \\
\ddot{y}_{i5} &= y_{i5} - \overline{y}_i & \Delta y_{i5} &= \text{mv}
\end{aligned}$$

where $\overline{y}_i = (y_{i1} + y_{i2} + y_{i5})/3$ and "mv" means missing

Fixed Effects with Unbalanced Panels ctd

Hence data loss often larger with FD

Can easily see that anybody who appears just once gets dropped as we cannot form the differenced/demean data

The more important point is deciding *why* the panel is unbalanced. If it is not for random reasons, FE is biased and inconsistent.

Random Effects Models

In the **random effects model** (RE), consistency requires unobserved effect is (assumed) uncorrelated with each explanatory variable

$$\text{Cov}(\boldsymbol{x}_{it}, a_i) = 0 \quad i = 1, ..N; \quad t = 1, \ldots, T \tag{8}$$

$\Rightarrow$ If this is not true use FE/FD

With this extra assumption, one could use a single cross-section, average the data (the between estimator), or use pooled OLS, as all are consistent. . . but . . .

Define the **composite error term** $v_{it} \equiv a_i + u_{it}$, then we can re-write the model

$$y_{it} = \beta_0 + \beta x_{it} + v_{it} \qquad i = 1, ..N; \quad t = 1, \ldots, T \tag{9}$$

Serial Correlation from Pooling Observations

Consider $v_{it} \equiv a_i + u_{it}$, $v_{is} \equiv a_i + u_{is}$ for the same individual :

$$\mathsf{Cov}(v_{it}, v_{is}) = \mathsf{E}[(a_i + u_{it})(a_i + u_{is})] = \mathsf{E}(a_i^2) = \sigma_a^2$$

$$\mathsf{Var}(v_{it}) = \mathsf{E}[(a_i + u_{it})^2] = \mathsf{E}(a_i^2 + u_{it}^2) = \sigma_u^2 + \sigma_a^2 = \mathsf{Var}(v_{is})$$

where $\sigma_a^2 = \mathsf{Var}(a_i)$ and $\sigma_u^2 = \mathsf{Var}(u_{it})$

Hence the v_{it} are positively serially correlated for the same individual across time:

$$\mathsf{Corr}(v_{it}, v_{is}) = \frac{\sigma_a^2}{\sqrt{\mathsf{Var}(v_{it})\mathsf{Var}(v_{is})}}$$

$$\mathsf{Corr}(v_{it}, v_{is}) = \frac{\sigma_a^2}{\sigma_u^2 + \sigma_a^2} \quad \text{for } t \neq s$$

It occurs because a_i is in the composite error in each time period. Stata calls this **rho**.

Since the errors are (positively) correlated, we must use **generalised least squares** (GLS). The estimator needs large N, relatively small T.

RE Estimator

The matrix algebra is found in the advanced texts (Wooldridge xsection, Chapter 10 or Baltagi, Chapter 1).

The GLS transformation turns out to be

$$\lambda = 1 - \sqrt{\sigma_u^2/(\sigma_u^2 + T\sigma_a^2)} = 1 - \sqrt{\theta} \tag{10}$$

where $0 \leq \lambda \leq 1$. The transformed equation is

$$y_{it} - \lambda \overline{y}_i = (1-\lambda)\beta_0 + \beta(x_{it} - \lambda \overline{x}_i) + (v_{it} - \lambda \overline{v}_i). \tag{11}$$

So,we have **quasi-demeaned data** and a_i goes into the error term:

$$v_{it} - \lambda \overline{v}_i = (1-\lambda)a_i + u_{it} - \lambda \overline{u}_i \tag{12}$$

If $\lambda = 1$, $\theta = 0$, GLS and the fixed effects estimator coincide. The a_i disappear completely.

But the random effects transformation only subtracts a proportion λ of the time average, where the fraction depends on σ_u^2, T, and σ_a^2

As T gets larger, GLS and FE get closer

If $\lambda = 0$, $\theta = 1$, GLS and OLS coincide. This is because $\sigma_a^2 = 0$, ie the random effects are no longer present. We just have pooled OLS.

Unless $\lambda = 1$, the transformation above allows for non-time-varying data. In which case, the effects of gender, ethnicity, education can be estimated.

Random Effects Models in Stata

Stata's command for random effects estimator is

xtreg y x, re i( )

and is computed by applying OLS to the quasi-demeaned equation

Consider the quasi-demeaned-error, $v_{it} - \lambda \overline{v}_i$, in (12)

RE is inconsistent if there is any correlation between a_i and any x_{it}.

However, as $\lambda \to 1$, the asymptotic bias goes to zero, and RE tends to FE

The GLS transformation also works on unbalanced data

Random Effects example

```
. xtreg lrhrpay  southeast  age age2 male married manual covered wave16 male16 if N==2, re i(pid)
>  robust

Random-effects GLS regression                   Number of obs      =      4290
Group variable: pid                             Number of groups   =      2145

R-sq:  within  = 0.4107                         Obs per group: min =         2
       between = 0.3179                                        avg =       2.0
       overall = 0.3389                                        max =         2

                                                Wald chi2(9)       =   2177.34
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

                               (Std. Err. adjusted for 2145 clusters in pid)
------------------------------------------------------------------------------
             |               Robust
     lrhrpay |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   southeast |   .0888458    .021696     4.10   0.000     .0463223    .1313692
         age |   .0728763   .0044007    16.56   0.000     .0642511    .0815014
        age2 |  -.0008568   .0000505   -16.98   0.000    -.0009557   -.0007578
        male |   .4023984   .0200834    20.04   0.000     .3630356    .4417612
     married |   .0516603   .0153716     3.36   0.001     .0215325    .0817881
      manual |  -.2502279    .016773   -14.92   0.000    -.2831024   -.2173534
     covered |   .1527357   .0154441     9.89   0.000     .1224657    .1830056
      wave16 |   .3190984   .0160643    19.86   0.000     .2876129    .3505839
      male16 |  -.0403212   .0199205    -2.02   0.043    -.0793647   -.0012777
       _cons |   .3571799   .0896432     3.98   0.000     .1814823    .5328774
-------------+----------------------------------------------------------------
     sigma_u |  .31899204
     sigma_e |  .31928609
         rho |  .49953931   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Random Effects or Fixed effects?

Crucial assumption for RE requires a_i uncorrelated with $\boldsymbol{x}_{it}$

For example, in the wage/schooling model, "motivation" (a_i) and actual school grades ($\boldsymbol{x}_{it}$) *must* be correlated

In this case RE is biased but FE is not

To test whether RE vs. FE is consistent we need a test for (i.e. $\text{Cov}(\boldsymbol{x}_{it}, a_i) = 0$)

Hausman's test

H_0 : a_i *are not correlated with* $\boldsymbol{x}_{it}$
H_a : a_i *are correlated with* $\boldsymbol{x}_{it}$

Under H_0, the GLS (random effects) estimator is consistent and efficient. The fixed effects estimator is always consistent, irrespective of whether H_0, is true.

So if H_0 is true, use GLS (as it is 'best' as in BLUE)

Under H_0, the test statistic

$$(\widehat{\boldsymbol{\beta}}_w - \widehat{\boldsymbol{\beta}}_g)'[\boldsymbol{V}(\widehat{\boldsymbol{\beta}}_w) - \boldsymbol{V}(\widehat{\boldsymbol{\beta}}_g)]^{-1}(\widehat{\boldsymbol{\beta}}_w - \widehat{\boldsymbol{\beta}}_g) \sim \chi^2(k),$$

where k is the number of elements in $\widehat{\boldsymbol{\beta}}_w$ and $\widehat{\boldsymbol{\beta}}_g$

Under the null we can show $[\boldsymbol{V}(\widehat{\boldsymbol{\beta}}_w) - \boldsymbol{V}(\widehat{\boldsymbol{\beta}}_g)] > 0$, then any systematic differences leads to a positive test statistic

If H_0 is true, $\widehat{\boldsymbol{\beta}}_w - \widehat{\boldsymbol{\beta}}_g$ is small

Hence, if H_0 is true, the test statistic is small and positive

In Stata we can use the **hausman** command

```
quietly xtreg lrhrpay  southeast  age age2 male married manual covered wave16 male16 if N==2,   i(pid) fe
estimates store FE1

quietly xtreg lrhrpay  southeast  age age2 male married manual covered wave16 male16 if N==2,   i(pid) re
estimates store RE1

 hausman FE1 RE1

Note: the rank of the differenced variance matrix (7) does not equal the number of coefficients
        being tested (8); 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))
              |      FE1          RE1         Difference          S.E.
-------------+------------------------------------------------------------
   southeast |    .0767908     .0888458        -.012055        .0428738
         age |     .055333     .0728763       -.0175433        .0370024
        age2 |   -.0008047    -.0008568         .000052        .0000279
     married |     .050731     .0516603       -.0009293        .0150408
      manual |   -.0258372    -.2502279        .2243907        .0189325
     covered |    .1006201     .1527357       -.0521155         .013627
      wave16 |    .4661019     .3190984        .1470035        .3684971
      male16 |    -.043198    -.0403212       -.0028768               .
------------------------------------------------------------------------------
                           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(7) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =      157.74
                Prob>chi2 =      0.0000
                (V_b-V_B is not positive definite)
```

Correlated Random Effects (CRE)

CRE assumes a_i is a linear function of the average level of x_{it}:

$$a_i = \alpha + \gamma \bar{x}_i + r_i, \tag{13}$$

where r_i is uncorrelated with each x_{it}

Substituting (13) into (14):

$$y_{it} = \alpha + \beta x_{it} + \gamma \bar{x}_i + r_i + u_{it} \qquad t = 1, \ldots, T \tag{14}$$

CRE applies GLS (RE) to (14) [because of the error term $r_i + u_{it}$]

The only difference to RE is the inclusion of $\bar{x}_i$ in the model. The estimator is labelled "c" for CRE

Mundlak's trick

A famous piece of algebra (Mundlak, 1976) shows that:

$$\widehat{\beta}_c = \widehat{\beta}_w \tag{15}$$

Adding the time average $\bar{x}_i$ and using RE is the same as subtracting time averages and using pooled OLS

Why is this useful?

If $\gamma = 0$, then we have RE estimates $\widehat{\beta}_g$. Adding $\bar{x}_i$ delivers $\widehat{\beta}_w$

If $\hat{\beta}_w$ differs (statistically) to $\hat{\beta}_g$

Hence a test of $H_0 : \gamma = 0$ is test of $Cov(a_i, x_{it}) = 0$.

Using the applied example we test test the joint sig. of the 6 $\bar{x}_i$ variables ...

```
. by pid: egen Asoutheast = mean(southeast)

. by pid: egen Aage       = mean(age      )

. by pid: egen Aage2      = mean(age2     )

. by pid: egen Amarried   = mean(married  )

. by pid: egen Amanual    = mean(manual   )

. by pid: egen Acovered   = mean(covered  )

. by pid: egen Awave16       = mean(wave16  )

. by pid: egen Amale16    = mean(male16   )

.
.
. list pid wave age male wave16 male16 Aage Amale Awave16 Amale16 if N==2 in 1/11, sepby(pid)

     +-------------------------------------------------------------------------------+
     |      pid   wave   age   male   wave16   male16   Aage   Amale16   Awave16   Amale16 |
     |-------------------------------------------------------------------------------|
  4. | 10023526      6    43      0        0        0     48         0        .5         0 |
  5. | 10023526     16    53      0        1        0     48         0        .5         0 |
     |-------------------------------------------------------------------------------|
  6. | 10028005      6    35      1        0        0   39.5        .5        .5        .5 |
  7. | 10028005     16    44      1        1        1   39.5        .5        .5        .5 |
     |-------------------------------------------------------------------------------|
 10. | 10055266      6    28      0        0        0   33.5         0        .5         0 |
 11. | 10055266     16    39      0        1        0   33.5         0        .5         0 |
     +-------------------------------------------------------------------------------+
```

```
 xtreg lrhrpay  southeast  age  age2  married  manual  covered  wave16  male16 male    Asoutheas
> t Aage Aage2 Amarried Amanual Acovered Awave16 Amale16 if N==2, robust i(pid) re

Random-effects GLS regression                   Number of obs      =      4290 Group variable: pid
Number of groups   =      2145

R-sq:  within  = 0.4309                         Obs per group: min =         2
       between = 0.3308                                        avg =       2.0
       overall = 0.3590                                        max =         2

                                                Wald chi2(15)      =         .
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =         .

                                (Std. Err. adjusted for 2145 clusters in pid)
------------------------------------------------------------------------------
             |               Robust
     lrhrpay |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   southeast |   .0767908   .0564688     1.36   0.174    -.0338861    .1874677
         age |    .055333   .0331157     1.67   0.095    -.0095725    .1202385
        age2 |  -.0008047   .0000627   -12.84   0.000    -.0009276   -.0006819
     married |    .050731   .0202825     2.50   0.012     .0109781    .0904839
      manual |  -.0258372   .0258614    -1.00   0.318    -.0765246    .0248502
     covered |   .1006201    .022956     4.38   0.000     .0556272    .1456131
      wave16 |   .4661019    .326576     1.43   0.154    -.1739754    1.106179
      male16 |  -.043198   .0195997    -2.20   0.028    -.0816127   -.0047832
        male |   .4521994   .0204948    22.06   0.000     .4120302    .4923685
  Asoutheast |    .001785    .060618     0.03   0.977    -.1170241    .1205942
       Aage |   .0278528   .0333851     0.83   0.404    -.0375808    .0932865
      Aage2 |  -.0001739   .0000967    -1.80   0.072    -.0003634    .0000156
    Amarried |  -.0010897   .0310277    -0.04   0.972     -.061903    .0597235
     Amanual |  -.3914623   .0333805   -11.73   0.000     -.456887   -.3260377
    Acovered |   .0918841   .0306689     3.00   0.003     .0317742    .1519939
     Awave16 |          0  (omitted)
     Amale16 |          0  (omitted)
       _cons |   .0852186   .2036999     0.42   0.676    -.3140258     .484463
-------------+----------------------------------------------------------------
     sigma_u |  .31899204
     sigma_e |  .31928609
         rho |  .49953931   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
. test Asoutheast Aage Aage2 Amarried Amanual Acovered

 ( 1)  Asoutheast = 0
 ( 2)  Aage = 0
 ( 3)  Aage2 = 0
 ( 4)  Amarried = 0
 ( 5)  Amanual = 0
 ( 6)  Acovered = 0

           chi2(  6) =  157.95
         Prob > chi2 =    0.0000
```

Verifies Mundlak's algebra - CRE is equivalent to FE estimates

Notice we have added *male*. This is an RE estimate

The test-statistic is very similar to the one above, but has the correct d.f., i.e. 6 restrictions

If we sum the 6 t-stats, squared, we get close to the test statistic

It is *manual* and *covered* that cause RE to be rejected against FE:

- Ability is a lot lower if manual

Costs of using FE/FD estimators

Although FE/FD estimation can solve the omitted variables bias problem, it comes at a price:

1. One whole cross-section is used up to estimate the fixed effects, so there is *inefficiency* due to data loss
2. The estimator relies on within variation; if this is small, the estimates are imprecisely estimated
   - The estimator relies on *changers*, ie individuals for whom something happens
   - For a lot individuals their personal characteristic/situatiosn do not change much (if at all) from one year to the next; e.g. occupation, industry, location, even wage, unless they change jobs or get promoted...

Costs of using FE/FD estimators

3. If there is no within variation at all, (ie variables are fixed through time), the effects of such variables cannot be estimated and cannot enter the regression
   - Examples are gender, ethnicity, school grades, which precludes investigation of some important issues, eg discrimination
   - For other variables like age, job-tenure, experience, $\Delta x_{it} =$ constant for all t, and so again cannot enter the regression if we have time dummies

1-3 are about information loss.

In 1 and 2, standard errors will be 'high'. These disadvantages are less important the larger T

There is a trade-off between efficiency and consistency, ie it is better to have a wrong estimate with some precision than a correct estimate that can be anywhere...

Finally, measurement error gets worse under differencing/de-meaning, and the bias caused might outweigh the bias from using RE incorrectly

Readings

↪ Wooldridge, Chapter 14