

# Introduction to Regression Models for Panel Data Analysis in Stata

Indiana University  
Workshop in Methods  
January 24, 2025

Patricia A. McManus

**Abstract:** This workshop provides an introduction to the analysis of panel data (sometimes called cross-section time-series data) using Stata statistical software. The focus is on the *linear error components model*. We cover *what* differentiates panel data from other longitudinal data, *why* use panel analysis techniques and *how* to use Stata's "xt" suite of commands to facilitate data exploration and analysis. The workshop introduces linear fixed effects models (in three flavors), random effects models, and Allison's (2009) hybrid model. Participants may have the opportunity to follow along using a small example dataset.

## What are Panel Data?

Panel data are a type of *longitudinal data*, or data collected at different points in time. Three main types of *longitudinal data*:

- Time series data. Many observations (large  $t$ ) on as few as one unit (small  $N$ ). Examples: stock price trends, aggregate national statistics.
- Pooled cross sections. Two or more *independent* samples of many units (large  $N$ ) drawn from the same population at different time periods:
  - General Social Surveys
  - US Decennial Census extracts
  - Current Population Surveys\*
- Panel data. Two or more observations (small  $t$ ) on many units (large  $N$ ).
  - Panel surveys of households and individuals (PSID, NLSY, ANES)
  - Data on organizations and firms at different time points
  - Aggregated regional data over time
- This workshop is a basic introduction to the analysis of *panel data*. In particular, I will cover the *linear error components model*.

## Why Analyze Panel Data?

- We are interested in *describing* change over time
  - social change, e.g. changing attitudes, behaviors, social relationships
  - individual growth or development, e.g. life-course studies, child development, career trajectories, school achievement
  - occurrence (or non-occurrence) of events
- We want *superior estimates* trends in social phenomena
  - Panel models can be used to inform policy – e.g. health, obesity
  - Multiple observations on each unit can provide superior estimates as compared to cross-sectional models of association
- We want to estimate *causal models*
  - Policy evaluation
  - Estimation of treatment effects

## **A few examples of questions we can address with panel:**

- What is the wage penalty for motherhood?
- What is Men's wage premium for heterosexual marriage?
- What is the effect of regulation of nursing pay on hospital quality?
- What is the effect of incarceration on wages and income inequality?
- What is the effect of parental divorce on mental health over the life-course?
- What factors are associated with the Death Penalty in US states?
- What is the association between strength of Democracy and Economic Growth?

## What kind of data are required for panel analysis?

- Basic panel methods require at least two “waves” of measurement.

Consider student GPAs and job hours during two semesters of college.

- One way to organize the panel data is to create a single record for each combination of unit and time period. Stata prefers this “long” format:

<i>StudentID</i>	<i>Semester</i>	<i>Female</i>	<i>HSGPA</i>	<i>GPA</i>	<i>JobHrs</i>
17	5	0	2.8	3.0	0
17	6	0	2.8	2.1	20
23	5	1	2.5	2.2	10
23	6	1	2.5	2.5	10

- Notice that the data include:
  - A time-invariant (TI) unique identifier for each unit (*StudentID*)
  - An indicator for time (*Semester*).
  - A time-varying (TV) outcome (*GPA*)
  - Three predictor variables (*Female*, *HSGPA*, *JobHrs*) -- TI or TV?
- Panel datasets can include other time-varying or time-invariant variables

- An alternative way to structure the data is to keep all the measures related to each student in a single record. This is sometimes called “wide” format.

<i>StudentID</i>	<i>Female</i>	<i>HSGPA</i>	<i>GPA5</i>	<i>JobHrs5</i>	<i>GPA6</i>	<i>JobHrs6</i>
17	0	2.8	3.0	0	2.1	20
23	1	2.5	2.2	10	2.5	10

(Stata prefers the long format)

- Why are there two variables for *GPA* and *JobHrs* ?
- Why is there only one variable for gender and high school GPA?
- Where is the indicator for time?

We can write a simple panel equation predicting GPA from hours worked:

$$GPA_{it} = \beta_0 + TERM_{it}\beta_T + HSGPA_{it}\beta_H + JOB_{it}\beta_J + c_i + u_{it}$$

$$v_i = c_i + u_{it}$$

We have a persistent student-specific error and an idiosyncratic error!

## Essential Linear Unobserved Effects Panel Models: FE, RE and Hybrid Estimation

- Motivation: Exploit unit-specific unobserved heterogeneity
  - There are two basic approaches: Fixed effects and Random Effect
  - FE eliminates some sources of bias to produce **consistent** estimates
  - If no evidence of bias, RE produces more **efficient** estimates than POLS
  - Correlated Random Effects (CRE) is a **hybrid** approach that combines FE,RE
- Challenges
  - An array of estimation choices, some appropriate, others not
  - The assumption for consistency is **strict exogeneity** more stringent than OLS
  - Strict exogeneity** means that there cannot be any feedback loop. The errors from any given period must be uncorrelated with past and future covariates.
- Opportunities: many specification tests to guide appropriate analysis

## Basic Questions for the Panel Analyst

### What's the story you want to tell?

- Is this a descriptive analysis? Less worry, fewer controls are usually better.
- Is this an attempt at causal analysis using observational data? Careful specification *AND* theory are essential.

### How does time matter?

- Some analysts may be interested in growth trajectories.
- Some analyses, e.g. difference-in-difference analysis associates time with an event (before and after)
- Time may be irrelevant. Panel models are a powerful tool for estimating relationships even if there is no interest in time. For example, studies using cross-sectional data collected at fixed periods of time often use dummy variables in a two-way specification with fixed-effects for time.

### Are the data up to the demands of the analysis?

- Panel analysis is data-intensive. Two waves are a bare minimum
- Can you perform the necessary specification tests?
- How will you address panel attrition?

But we're getting ahead of ourselves.

## Brief Review of the Classical Linear Regression Model

,  $i=1,2,3,\dots,N$

Where we assume that the linear model is correct and:

- Covariates are Exogenous:  $E(u_i | x_{1i}, x_{2i}, \dots, x_{ki}) = 0$
- Uncorrelated errors:  $Cov(u_i, u_j) = 0$
- Homoskedastic errors:  $Var(u_i) = Var(y_i | x_{1i}, x_{2i}, \dots, x_{ki}) = \sigma^2$

If assumptions do not hold, OLS estimates are BIASED and/or INEFFICIENT

- Biased - Expected value of parameter estimate is different from true.
  - Consistency. If an estimator is unbiased, or if the bias shrinks as the sample size increases, it is *CONSISTENT*
- Inefficient - An estimator is inefficient if an alternative estimator converges more rapidly on the true coefficients as sample size increases.
  - Estimators that exploit all available information are more efficient

## OLS Inefficiency due to Correlated Errors

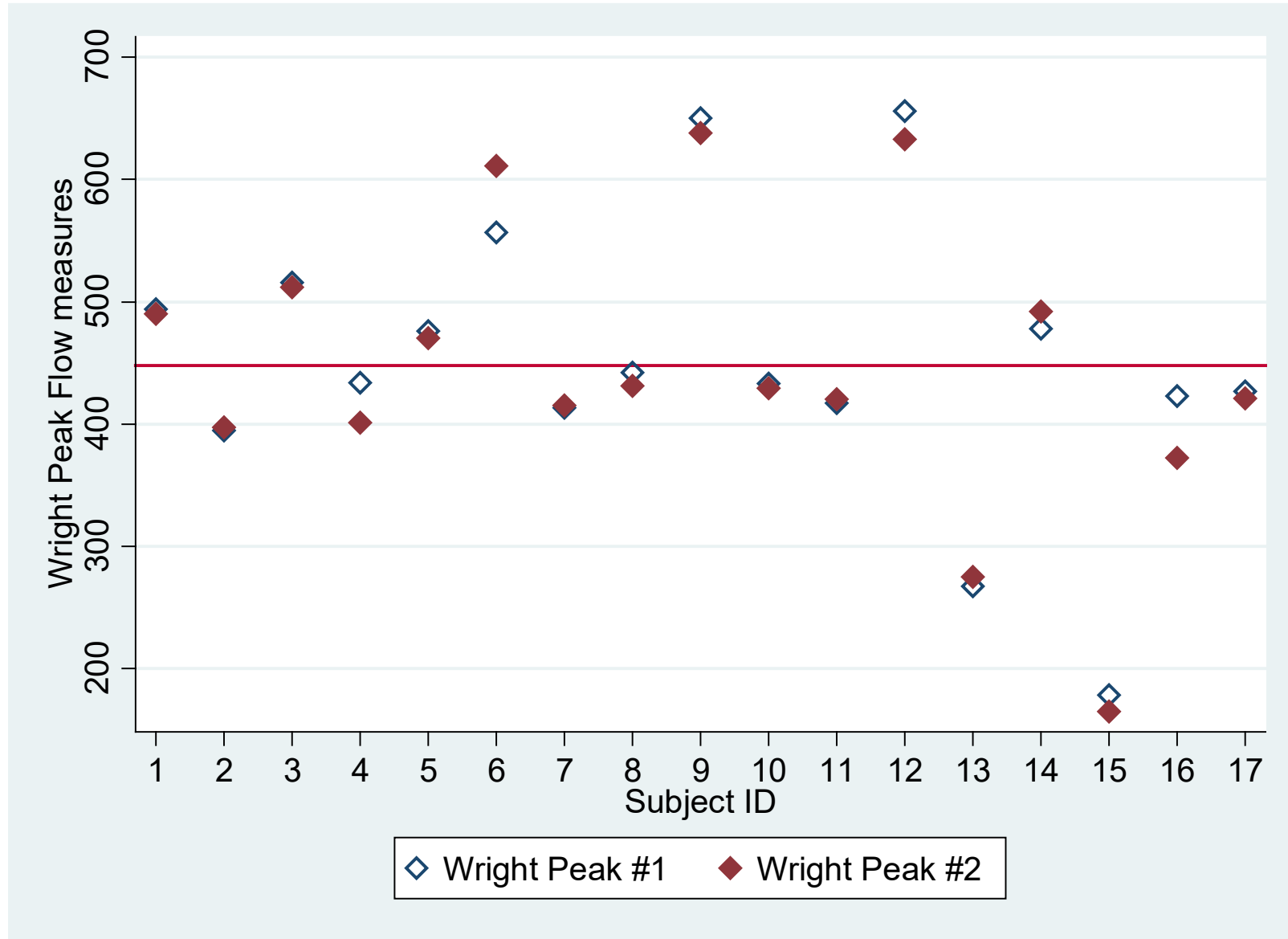
Many data structures are susceptible to error correlation:

- Hierarchical data sample multiple individuals from each unit, e.g. household members, employees in firms, multiple pupils from each school.
- Multistage probability samples often incorporate cluster-based sampling designs with errors that may be correlated within clusters.
- Repeated observations data often show within-unit error correlation.
- Time series data often have errors that are serially correlated, that is, correlated over time.
- Panel data have errors that can be correlated within unit (e.g. individuals), within period.

## Conventional regression-based strategies to address correlated errors

- Cluster-consistent covariance matrix estimator to adjust standard errors.
- Generalized Least Squares instead of OLS to exploit correlation structure.
- Generalized Estimation Equations (GEE)
- Mixed Effects Estimators for multilevel models

# Illustration of Within-unit correlation. Peak-flow Measurements



## Linear Panel Data Model (LPM)

Suppose the data are on each cross-section unit over  $T$  time periods:

$$\begin{aligned} y_{i,t1} &= \mathbf{x}'_{i,t1} \boldsymbol{\beta}_{t1} + u_{i,t1} \\ y_{i,t2} &= \mathbf{x}'_{i,t2} \boldsymbol{\beta}_{t2} + u_{i,t2}, & t=1,2,\dots,T \\ &\vdots \\ y_{i,T} &= \mathbf{x}'_{i,T} \boldsymbol{\beta}_T + u_{i,T} \end{aligned}$$

We can estimate this using wave-by-wave analysis or pooled OLS (POLS)

Example: Begin with two conventional OLS linear regression models using the GPA data, one for each period. Then estimate a POLS specification.

Note that the variables `female` `highgpa` (HS GPA) are time-invariant.

## OLS Results for each term:

	Term 5 GPA			Term 6 GPA		
	Estimate	SE	t-stat	Estimate	SE	t-stat
Intercept	3.02	0.17	17.8	3.02	0.17	18.3
jobhrs	-0.182	0.05	-4.0	-0.174	0.05	-3.6
female	0.108	0.04	2.5	0.145	0.05	3.2
highgpa	-0.004	0.04	-0.1	0.003	0.04	0.1

## Pooled OLS Results for both terms:

	Term 5&6 GPA			Term 5&6 GPA (Clustered SE)		
	Estimate	SE	t-stat	Estimate	SE	t-stat
Intercept	2.97	0.17	25.1	2.97	0.17	17.2
jobhrs	-0.178	0.05	-5.4	-0.178	0.05	-5.8
female	0.125	0.04	4.1	0.125	0.04	3.0
highgpa	-0.0001	0.03	-0.01	0.0001	0.03	-0.0004
term6	0.095	0.016	6.1	0.095	0.016	6.1

## Linear Unobserved Effects Panel Data Model

- Motivation: Unobserved heterogeneity

Suppose we have a model with an unobserved, time-constant variable  $c$ :

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \dots + x_k\beta_k + c + u$$

Where  $u$  is uncorrelated with all explanatory variables in  $\mathbf{x}$ .

Because  $c$  is unobserved it is absorbed into the error term, so we can write the model as follows:

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \dots + x_k\beta_k + v$$
$$v = c + u$$

The error term  $v$  consists of two components, an “idiosyncratic” component  $u$  and an “unobserved heterogeneity” component  $c$ .

## POLS Estimation of the Error Components Model

- If the unobserved heterogeneity  $c$  is uncorrelated with the explanatory variables in  $\mathbf{x}_i$ , OLS is unbiased whether a single cross-section or pooled.
- But...If we have more than one observation on any unit, the errors will be correlated and OLS estimates will be inefficient

$$y_{i,1} = \beta_0 + x_{1_{i1}}\beta_1 + x_{2_{i1}}\beta_2 + \dots + x_{k_{i1}}\beta_k + v_{i,1}$$

$$y_{i,2} = \beta_0 + x_{1_{i2}}\beta_1 + x_{2_{i2}}\beta_2 + \dots + x_{k_{i2}}\beta_k + v_{i,2}$$

$$v_{i,1} = c_i + u_{i,1}$$

$$v_{i,2} = c_i + u_{i,2}$$

$$\text{cov}(v_{i,1}, v_{i,2}) \neq 0$$

- One strategy is to combine pooled OLS with cluster-consistent standard errors.
- Panel methods over OLS to exploit *OR* remove unobserved heterogeneity.

Even if estimation is consistent, pooled OLS may not be efficient.

In the next sections, we consider the dominant approaches to estimation of the error components panel model: **fixed effects** and **random effects**.

## Fixed Effects Methods for Panel Data

Suppose the unobserved effect  $c_i$  is correlated with the covariates.

Example: Motherhood wage penalty

- We observe that mothers earn less than other women, *cet par.*

$\hat{\beta}_{KIDS_{OLS}} = -0.08$  in a log wage model suggests that each additional child reduces mothers' hourly wages by about 8%

But if women who are less oriented towards work are also more likely to have more children, omitting “work orientations” from the model will bias the coefficient on children.

- Fixed-effects methods transform the model to remove  $c_i$

$\hat{\beta}_{KIDS_{FE}} = -0.03$  FE estimates a persistent but much smaller penalty.

## Fixed Effects Transformation - the “Within” Estimator

Suppose we have the UEM model:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + c_i + u_{it}, \quad t=1,2,\dots,T$$

For each unit, average this equation over all time periods  $t$ :

$$\bar{y}_i = \bar{\mathbf{x}}'_i\boldsymbol{\beta} + \bar{c}_i + \bar{u}_i$$

Subtract the within-unit average from each observation on that unit:

$$y_{it} - \bar{y}_i = (\mathbf{x}'_{it} - \bar{\mathbf{x}}'_i)\boldsymbol{\beta} + (c_i - \bar{c}_i) + (u_{it} - \bar{u}_i), \quad t=1,2,\dots,T$$

This is the **fixed effects transformation**. We can write it as:

$$\ddot{y}_{it} = \ddot{\mathbf{x}}'_{it}\boldsymbol{\beta} + \ddot{u}_{it},$$

where  $c_i - \bar{c}_i = 0$  and  $\ddot{y}_{it} = y_{it} - \bar{y}_i$ ,  $\ddot{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$ ,  $\ddot{u}_{it} = u_{it} - \bar{u}_i$

and  $\ddot{\mathbf{x}}_{it}$  does not contain an intercept term.

The **fixed-effects estimator**, also called the **within estimator**, applies pooled OLS to the transformed equation:

$$\hat{\beta}_{FE} = \left( \sum_{i=1}^N \ddot{\mathbf{X}}_i' \ddot{\mathbf{X}}_i \right)^{-1} \left( \sum_{i=1}^N \ddot{\mathbf{X}}_i' \ddot{\mathbf{y}}_i \right) = \left( \sum_{i=1}^N \sum_{t=1}^T \ddot{\mathbf{x}}_{it}' \ddot{\mathbf{x}}_{it} \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T \ddot{\mathbf{x}}_{it}' \ddot{y}_{it} \right)$$

Recall the student GPA Data:

<i>StudentID</i>	<i>Semester</i>	<i>Female</i>	<i>HSGPA</i>	<i>GPA</i>	<i>JobHrs</i>
17	5	0	2.8	3.0	0
17	6	0	2.8	2.1	20
23	5	1	2.5	2.2	10
23	6	1	2.5	2.5	10

After applying the fixed-effects transform, the demeaned (mean-centered) data:

<i>StudentID</i>	<i>Semester</i>	<i>CFemale</i>	<i>CHSGPA</i>	<i>CGPA</i>	<i>CJobHrs</i>
17	-.5	0	0	.45	-10
17	.5	0	0	-.45	10
23	-.5	0	0	-.15	0
23	.5	0	0	.15	0

# STATA Example. FE "Within" Estimates for Student GPA & Pooled OLS (Terms 5 & 6)

```

/* 1. FE "Within" Estimates for Term 5 and Term 6 */
. xtreg gpa job female highgpa,fe
note: female omitted because of collinearity
note: highgpa omitted because of collinearity

Fixed-effects (within) regression           Number of obs   =       400
Group variable: student                    Number of groups =       200
::::
corr(u_i, Xb) = 0.3989                     F(1,199)        =       7.59
                                           Prob > F         =       0.0064
-----+-----
      gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
      job |   -.0730158   .0265     -2.76   0.006   -0.1252727   -0.0207589
  female |           0   (omitted)
 highgpa |           0   (omitted)
   _cons |    3.22673   .0550286   58.64   0.000    3.118216    3.335244
-----+-----
  sigma_u |   .34081956
  sigma e |   .14873104
      rho |   .84002677   (fraction of variance due to u_i)
-----+-----
F test that all u_i=0: F(199, 199) = 8.83          Prob > F = 0.0000.

```

```

/* 2. Pooled OLS Estimates for Term 5 and Term 6 */
. reg gpa job female highgpa

      Source |      SS          df           MS       Number of obs   =       400
-----+-----+-----+-----+-----+-----
      F(3, 396)          =       53.6
::::
-----+-----
      gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
      job |   -.3312173   .0316199   -10.47   0.000   -0.3933813   -0.2690534
  female |   .1921272   .0316134    6.08   0.000    0.1299762    0.2542783
 highgpa |   .0675687   .0265919    2.54   0.011    0.0152898    0.1198476
   _cons |   3.455251   .1100984   31.38   0.000    3.238801    3.671702

```

## Fixed Effects Dummy Variables Regression

Up to now, we've treated the unobservables  $c_i$  as random variables:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + c_i + u_{it}$$

An alternative approach is to treat  $c_i$  as a fixed parameter for each unit. In this case, we can use dummy variables regression to estimate  $c_i$ .

Step one: Create a dummy variable for each of sample unit  $i$

Step two: Substitute the vector of  $N-1$  dummies for  $c_i$ :

$$y_{it} = \gamma_1 + \mathbf{x}'_{it}\boldsymbol{\beta} + d2\gamma_2 + d3\gamma_3 + \dots + dN\gamma_N + u_{it},$$

(where the intercept  $\gamma_1$  estimates the effect when  $d1=1$ )

Step three: Estimate the equation using pooled OLS.

- The fixed effects dummy variables (FEDV) estimator produces precisely the same coefficient vector and standard errors as the FE estimator.

## First Differences Regression

An alternative approach is to remove  $c_i$  using first differences. In this case, we first transform the variables and then run OLS.

Start with the error components model:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + c_i + u_{it}, \quad t=1,2,\dots,T$$

For each observation, subtract the previous within-unit observation:

$$y_{it} - y_{i,t-1} = (\mathbf{x}'_{it} - \mathbf{x}'_{i,t-1})\boldsymbol{\beta} + (c_i - c_i) + (u_{it} - u_{i,t-1})$$

This is the **first-difference transformation**. We can write it as:

$$\begin{aligned} \Delta y_{it} &= \Delta \mathbf{x}'_{it}\boldsymbol{\beta} + \Delta c_i + \Delta u_{it} \\ &= \Delta \mathbf{x}'_{it}\boldsymbol{\beta} + \Delta u_{it} \end{aligned}$$

where  $\Delta c_i = 0$  and  $\Delta \mathbf{x}'_{it}$  does not contain an intercept term.

## First Differences Regression

Recall the student GPA Data:

<i>StudentID</i>	<i>Semester</i>	<i>Female</i>	<i>HSGPA</i>	<i>GPA</i>	<i>JobHrs</i>
17	1	0	2.8	3.0	0
17	2	0	2.8	2.1	10
23	1	1	2.5	2.2	10
23	2	1	2.5	2.5	10

After the first-difference transform, the first observation on each unit is lost:

<i>StudentID</i>	<i>DSemester</i>	<i>DFemale</i>	<i>DHSGPA</i>	<i>DGPA</i>	<i>DJobHrs</i>
17	.	.	.	.	.
17	1	0	0	-0.9	10
23	.	.	.	.	.
23	1	0	0	0.3	0

- With two waves, the first difference (FD) estimator produces precisely the same coefficient vector and standard errors as the FE estimator.
- With more than two waves the estimates will be different.

## Caution: Fixed effects has some disadvantages

- ⇒ FE is not a panacea for all sources of endogeneity bias.
  - time-varying* unobserved effects
  - time-varying* measurement error
  - simultaneity* or feedback loops
  
- ⇒ All time-constant effects are removed.
  - No estimation of effects of race, gender, birth order, etc.
  - Poor estimates if little variation (e.g. education in adulthood)
  
- ⇒ FE trades consistency for efficiency.
  - FE uses *only* within-unit change, ignores between-unit variation.
  - Parameter estimates may be imprecise, standard errors large.
  
- Despite limitations, FE is an indispensable tool in the panel analyst's toolbox.

## Random Effects Methods

If we can assume that the unobserved heterogeneity will not bias the estimates:

- Fixed effects methods are *inefficient*. They throw away information.
- Pooled OLS is *inefficient* because it does not exploit the autocorrelation in the composite error term.
- Random effects methods use feasible GLS estimation (RE FGLS) to exploit within-cluster correlation
- Random effects estimation is more *efficient* than FE or OLS
- The “random effects assumption” of no bias due to  $c_i$  is more stringent

$$E(c_i \mid \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}) = E(c_i) = 0$$

## A Conventional FGLS Random Effects Estimator

Assume the errors are correlated within each unit

Assume the errors are uncorrelated across units

Assume the variance in the composite errors is equal to the sum of the variances in the unobserved effect  $c_i$  and the idiosyncratic error  $u_i$ :

$$\sigma_v^2 = \sigma_u^2 + \sigma_c^2$$

RE strategy: If  $\sigma_v^2 = \sigma_u^2 + \sigma_c^2$ , find estimators such that  $\hat{\sigma}_v^2 = \hat{\sigma}_u^2 + \hat{\sigma}_c^2$

## Practical Feature of Random Effects Estimation

- Recall that the fixed effects “within” estimator essentially transforms the data by centering each variable on the unit-specific mean.

OLS is then performed on the “fully demeaned” transformed data.

- The random effects estimator essentially transforms the data by “partially demeaning” each variable. Instead of subtracting the entire unit-specific mean, only part of the mean is subtracted.

The demeaning factor  $\lambda$  is between 0 and 1, with the specific value based on the variance components estimation.

## STATA Example Random Effects Estimates for Student GPA over Six Terms

```
. xtreg gpa job female highgpa i.term, re theta
```

```
Random-effects GLS regression      Number of obs      =      1,200
Group variable: student            Number of groups   =      200
```

```
R-sq:                               Obs per group:
  within = 0.4298                      min =      6
  between = 0.2802                     avg  =     6.0
  overall = 0.3480                      max  =      6
```

```
Wald chi2(8) = 795.01
Prob > chi2 = 0.0000
```

```
corr(u_i, X) = 0 (assumed)
theta = .52872425
```

gpa	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
job	-.1723449	.018694	-9.22	0.000	-.2089844	-.1357054
female	.1479121	.0298693	4.95	0.000	.0893693	.2064548
highgpa	.0854096	.0250853	3.40	0.001	.0362434	.1345758
term						
2	.1271703	.024284	5.24	0.000	.0795747	.174766
3	.2182234	.0242782	8.99	0.000	.170639	.2658078
4	.3210531	.0242804	13.22	0.000	.2734644	.3686417
5	.416021	.0242992	17.12	0.000	.3683954	.4636467
6	.5267124	.0243235	21.65	0.000	.4790392	.5743856
_cons	2.626918	.091072	28.84	0.000	2.44842	2.805416
sigma_u	.18035047					
sigma_e	.23605128					
rho	.36858452	(fraction of variance due to u_i)				

## Random Effects or Fixed Effects - How to decide?

Hausman test for the Exogeneity of the Unobserved Error Component

If the unobserved effects are exogenous, the FE and RE are asymptotically equivalent. This suggests the null hypothesis for the Hausman test:

$$H_0 : \hat{\beta}_{RE} = \hat{\beta}_{FE} ,$$

where  $\hat{\beta}_{RE}$  and  $\hat{\beta}_{FE}$  are coefficient vectors for the time-varying explanatory variables, excluding the time variables.

If the null hypothesis is rejected, we conclude that RE is inconsistent, and the FE model is preferred.

If the null hypothesis cannot be rejected, random effects is preferred because it is a more efficient estimator.

## Conventional Hausman Test in Stata:

```
. xtreg gpa job sex highgpa,fe  
. estimates store fe  
. xtreg gpa job sex highgpa,re  
. estimates store re
```

```
. hausman fe re
```

```
----- Coefficients -----  
      |      (b)      (B)      (b-B)      sqrt(diag(V_b-V_B))  
      |      fe      re      Difference      S.E.  
-----+-----  
job |  -.0748115  -.1232374      .048426      .0088051  
-----
```

b = consistent under  $H_0$  and  $H_a$ ; obtained from xtreg

B = inconsistent under  $H_a$ , efficient under  $H_0$ ; obtained from xtreg

Test:  $H_0$ : difference in coefficients not systematic

```
chi2(1) = (b-B)' [(V_b-V_B)^(-1)] (b-B)  
        =      30.25  
Prob>chi2 =      0.0000
```

- We reject the null and conclude the fixed effects estimator is appropriate.

## An “Alternative” Hausman Test and FE/RE “Hybrid” Methods

### The Correlated Random Effects Model (Allison 2005,2009; Wooldridge 2005)

Suppose we have both time-varying and time-invariant covariates:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\delta} + c_i + u_{it}, \quad t=1,2,\dots,T$$

Add a vector of within-unit means for the time-varying covariates:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\delta} + \bar{\mathbf{x}}'_i\boldsymbol{\xi} + c_i + u_{it}$$

Estimate using the random effects estimator and the result is a “hybrid”

- Alternative Hausman: If a Wald test for the joint statistical significance of the coefficient estimates in  $\boldsymbol{\xi}$  rejects the null, FE is preferred
- The coefficient vector  $\boldsymbol{\beta}$  yields the fixed effects estimates
- The coefficient vector  $\boldsymbol{\xi}$  produces the between estimates
- The coefficient vector  $\boldsymbol{\delta}$  produces estimates for the time-invariant covariates. Interpret with caution: these might still be correlated with the unobserved error.

## The “Hybrid Model” and Correlated Random Effects Methods for Panel Data

Insights from the regression-based test of the RE assumption motivate a model that combines the advantages of the FE and RE model.

Sociologists more often use the virtually identical “Hybrid Model” following Allison (2005, 2009). A related model is Wooldridge (2005) “Correlated Random Effects Model (CRE).

### The Hybrid Model (the Allison Approach):

Substitute a vector of fully demeaned covariates for all time-varying predictors. Augment the error components model with a vector of unit-specific means. Estimate the equation using robust RE.

$$y_{it} = (\mathbf{x} - \bar{\mathbf{x}})'_{it} \boldsymbol{\beta} + \bar{\mathbf{x}}'_i \boldsymbol{\xi} + \mathbf{z}'_i \boldsymbol{\delta}_H + c_i + u_{it}$$

- The coefficient vector  $\boldsymbol{\beta}$  yields the fixed effects estimates
- The coefficient vector  $\boldsymbol{\xi}$  produces exactly the between estimates for time-varying covariates.
- The coefficient vector  $\boldsymbol{\delta}$  produces exactly the between effects for time-invariant covariates.
- The between estimates might still be correlated with the unobserved error, so interpret with caution.

## Correlated Random Effects (CRE) - Wooldridge (2005)

Revisit the Mundlak error components model augmented with covariate means:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\xi + \mathbf{z}'_{it}\boldsymbol{\delta} + c_i + u_{it},$$

The equation decomposes the total “between” unit variance for time-varying covariates into a within and a residual-between-component. Estimate the equation using a robust RE estimator.

- The coefficient vector  $\boldsymbol{\beta}$  yields the fixed effects estimates
- The sum of the coefficient vectors  $\boldsymbol{\beta}$  and  $\xi$  produces the “between” effects estimates
- The coefficient vector  $\boldsymbol{\delta}$  produces exactly the between effects for time-invariant covariates.
- The between estimates might still be correlated with the unobserved error, so interpret with caution.

## Interpretation of Results from the Error Components Model

Since the UEM model is derived as a *levels* model, coefficients can be interpreted much the same as interpretations of a conventional OLS model, but there are nuances:

For example, suppose we estimate the relationship between marriage and men's wages,  $\hat{\beta}_{MARRIED} \simeq 0.05$  in every model.

- **Pooled OLS** cross-section coefficients contain information about average differences between units.

$$E[y_{it} | \mathbf{x}_{it}] = \mathbf{x}_{it}\boldsymbol{\beta} + c_i$$

This is a *population-averaged effect*. On average, married men earn 5% more than men who are not married.

This says nothing about the *causal* effect of marriage on men's earnings.

- **RE/FE/FD** estimate average effects *within* units.

If the unobserved effects are exogenous these are asymptotically equivalent to the population averaged effect.

$$E[y_{it} \mid \mathbf{x}_{it}, c_i] = \mathbf{x}_{it}\boldsymbol{\beta}$$

On average, entering marriage increases men's earnings by 5%.

- **RE** coefficients represent average change *within* units, estimated from all units whether they experience change or not.
- **FE** coefficients represent average changes *within* units, only for units that did experience change

This is akin to a *treatment effect among the treated*.

On average, men who married increased their earnings by 5%.

## Best Practices

### Theorize the model

- What exactly does this unobserved heterogeneity represent?
- Why would you expect it to be correlated / uncorrelated with the regressors?
- Is it likely there is endogeneity due to time-varying unobserved heterogeneity or feedback from the idiosyncratic error to the next wave of covariates?

### Specification Testing for Panel Analysis - Interval/Continuous Outcomes

- Always neglected...but formal test for unobserved effect can be useful.
- Optional: Obtain intraclass correlation coefficient (ICC) as indicator of the extent of within-unit clustering. This is a descriptive statistic, not a test.
- Specification test(s) for strict exogeneity
- Hausman-type specification test for RE vs. FE
- Test for serial correlation in the idiosyncratic errors

## Extensions

### FE Models with Time-Invariant Predictors

- Interactions between time and covariate

### Panel Models for Categorical Outcomes

- Fixed effects logit and random effects logit for binary outcomes
- Fixed and random effects Poisson models can be used for count outcomes.
- Population averaged models can be estimated using General Estimation Equations (GEE).

**Dynamic panel models** i.e. lagged dependent variable as a covariate:

$$GPA_{it} = \beta_0 + GPA_{i,t-1}\beta_{GPA} + TERM_{it}\beta_T + HSGPA_{it}\beta_H + JOB_{it}\beta_J + v_{it}$$

- GLM models for instrumental variables (IV) estimation
- Generalized Method of Moments (GMM) is used for some dynamic panel models because it allows a flexible specification of the instruments