

# INTRODUCTION TO PANEL DATA

Aleksandra Gregoric

Center for Corporate Governance, Copenhagen Business School

15 April 2010

# Overview of the lecture

- (1) Pooling independent cross-section data across time.
- (2) Simple Panel Data methods: two periods, first differenced estimator.
- (3) Advanced Panel Data Methods:
  - ▶ Fixed-effects estimator and random-effects estimator.
  - ▶ Estimations with Stata.

# Overview of the lecture

## Suggested readings:

Wooldridge, J. M. (2003), *Introductory Econometrics. A Modern Approach*, Thomson South-Western.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MIT Press.

Wintoki et al., (2008), *Endogeneity and the Dynamics of Corporate Governance*, available at [www.ssrn.com](http://www.ssrn.com).

# Features of the data sets

(1) **Cross-section**: a sample of individuals (units), taken at a given point of time ( $i = 1, 2, \dots, N$ ).

(2) **Time series**: observations on a variable or several variables over time ( $t = 1, 2, \dots, T$ ).

(3) **Pooled independent cross-section**: cross-sections of data at two or more points of time (the units in the cross section are mostly not the same).

# Features of the data sets

4. **Panel or longitudinal data**: a time series of the same cross section of individual units:

- ▶ Two sample dimensions: cross-sectional ( $j = 1, \dots, N$ ), time-series ( $t = 1, \dots, T$ ).
- ▶ In this lecture, we focus on **short panels** ( large  $N$  and small  $T$ ) and **static** models (all explanatory variables are dated contemporaneously with the dependent variable).

# Pooling independent cross sections across time

- ▶ Two dimensions of the data: cross-sectional and time-series.
- ▶ Obtained by SAMPLING RANDOMLY from the populations at DIFFERENT POINTS in time (i.e. households surveys).
- ▶ Given the independent sampling, we can rule out the correlation in the error terms across different observations.

## **Why using independently pooled cross-section?**

- ▶ To increase the sample size, obtain more precise estimators and test statistics with more power.
- ▶ Minor statistical complications (see next slide).

# Pooling independent cross sections across time

## Note:

(1) Normally, year dummy variables should be included in the model to account for the different distribution of the population at different time periods.

(2) We generally assume that the relation between the explanatory variables and the dependent variable remains constant over time. However, we can release this assumption by interacting a specific explanatory variable with the year dummies.

(3) Heteroskedasticity-robust standard errors should be used to account for the fact that error variance may change over time, and with the values of the explanatory variables.

# Pooling independent cross sections across time

## Example: Wage determinants (Wooldridge, 2003)

Sample: We have information for a set of 550 people in the 1978 sample and a different set of 534 people in the 1985 sample. We want to estimate:

$$\log(\text{wage}) = \beta_0 + \delta_0 y85 + \beta_1 \text{educ} + \delta_1 y85 \cdot \text{educ} + \beta_2 \text{exp er} + \beta_3 \text{exp er}^3 + \beta_4 \text{union} + \beta_5 \text{female} + \delta_5 y85 \cdot \text{female} + u$$

In this case  $\delta_1$  measures how the return to education has changed over the 1978-1985 period.

Note that there is no need for deflating since the wage is measured in logarithms, and the year dummy is included in the regression.

## Wage determinants: Results

Dependent variable: log (wage)		
Independent variables	Coefficients	Standard Errors
female	-0.317	0.037
y85*female	0.085	0.051
union	0.202	0.030
exper	0.0296	0.0036
exper <sup>2</sup>	-0.0004	0.00008
y85*educ	0.0185	0.0094
educ	0.0747	0.0067
y85	0.118	0.124
Constant	0.459	0.093
n=1084 R <sup>2</sup> =0.426 R <sup>2</sup> (Adj)=0.422		

Questions:

Do we observe a significant reduction in the gender gap across the years?

How would you test whether the regression function differs across the two years?

## Policy analysis with pooled cross sections

Assume that we want to analyze the impact of a certain policy or event (taking place in year  $t$ ) on the relationship between  $y_{it}$  and  $x_{it}$ . For example, we want to analyze whether the construction of a new road influenced the real price of houses ( $P$ ) close to the road.

Rather than running a simple OLS at period  $t$ , we need to look at the difference in the effect of being close to the road on the house before and after the road was build (given that we dispose with two periods of data, we can actually do this).

Namely, we use the **Difference- in-difference estimator**, which (for the simple model) gives:

$$\begin{aligned} & \textit{estimated impact of the road construction} = \\ & (\bar{P}_{close(t)} - \bar{P}_{far(t)}) - (\bar{P}_{close(t-1)} - \bar{P}_{far(t-1)}) \end{aligned}$$

## Pooling independent cross sections across time

Which actually means that we estimate (by pooling the data over the two years) the following:

$$P = \beta_0 + \delta_0 \cdot (t) + \beta_1 \mathit{Close} + \delta_1 \cdot (t) \cdot \mathit{Close} + u$$

- ▶ Note that  $\beta_1$  captures the location effect that is not due to the road construction.
- ▶  $(t)$  is the dummy indicating time  $t$ .
- ▶ Treatment/control groups.
- ▶ Better estimates when the **same cross-sectional units appear in each time period (PANEL)**, which allows us to address the "endogeneity" of the treatment effect.

## Basic Considerations: Linear panel data models

(1) Panel data involve repeated measurements on different units, normally at regular intervals (i.e. end of each year). If all individual units are observed in each all time periods, we have a BALANCED panel, otherwise the panel is UNBALANCED.

(2) The choice between the main estimators is based on the assumptions with regards to the time-invariant component of the error term ( $\alpha_j$ ).

(3) The identification of the regression coefficients (for some estimators) may depend on the type of the regressor (i.e. time-invariant).

(4) Since the same units are observed across time, the error terms are likely to be correlated across units: correction of the default OLS standard errors is necessary and efficiency can be gained by using generalized least squares models.

## Panel data have several attractive features (Kennedy,2003)

- (1) Panel data can be used to **deal with the unobserved heterogeneity in the micro-units** (i.e. to account for unmeasured explanatory variables that affect the behavior of the micro units ( $a_i$ ) and time-series variables that influence the behavior of the units uniformly but differently in each time period).
- (2) Panel data create more variability, combining variation across micro units with variation across time (alleviating multicollinearity problems).
- (3) Panel data allow us to examine issues that cannot be studied using time-series or cross-section alone (i.e. separating economies of scale effect from the effect of a technological change), dynamic adjustments, lagged effects of the regressors, etc.

# Linear Panel data models

Before going to linear panel data models, recall the assumptions of the **classical linear model**:

(1) **Linear in parameters**: the population model can be written as  
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots \beta_k x_{ik} + u_i.$$

(2) **Random sampling**: we have a random sample of N observations from the population model.

(3) **Zero conditional mean**:  $E(u_i | x_1, x_2, \dots, x_k) = 0$ .

- ▶ Violations of Assumption 3: functional misspecification, **omitted variables problem (unobserved heterogeneity)**, measurement errors;

## Classical linear model

(4) **No perfect collinearity**: none of the explanatory variables in the sample is constant and there are no exact linear relationships among the independent variables.

(5) **Homoskedasticity** ( $\text{Var}(u_i | x_1, x_2, \dots, x_k) = \sigma^2$ ): the variance of the error term, conditional on the explanatory variables, is the same for all combinations of the outcomes of the explanatory variables (i.e. does not depend on the levels of  $x$ ). Assumption (5) is necessary in order to obtain unbiased estimation of the error variance  $\sigma^2$  and the s.e. of the coefficients.

(6) **Normality**:  $\sigma^2 : u \sim \text{Normal}(0, \sigma^2)$ .

## Classical linear model

If 1-4 holds:  $E(\hat{\beta}_j) = \beta_j, j = 0, 1 \dots k$  (OLS is unbiased estimator).

Under assumptions 1-5 (Gauss-Markow assumptions), the OLS is the best linear unbiased estimator (BLUE).

Under the assumptions 1-6:  $\hat{\beta}_j \sim \text{Normal}(\beta_j, \text{Var}(\hat{\beta}_j))$

To apply the usual OLS statistics in panel data (pooled OLS regression across  $i$  and  $t$ ), we need to add the assumption on no serial correlation in the error term:  $\text{Cov}(u_{it}, u_{is} | x_1, x_2, \dots, x_k) = 0$

## Linear Panel Data models: How does it work?

**Example:** We want to estimate whether large firms have on average less concentrated ownership. We dispose with ownership and other firm-level data over the 2000-2004 period (information refers to the end of each year).

In testing this, we could simply pool the data together and run an OLS (i.e. **pooled OLS estimator**):

$$\text{Logit}(C_1)_{it} = \beta_0 + \delta_0 d_t + \beta_1 \ln(\text{size})_{it} + \beta_2 \text{Roa}_{it} + \beta_3 \text{risk}_{it} + \dots + v_{it};$$
$$i = 1..N, t = 1, ..T$$

- ▶ Since the ownership share ( $C_1$ ) is bounded between 0 and 1, we apply the transformation in the log-odds ratio ( $\text{Logit}(C_1)_{it} = \ln(\frac{C_1}{1-C_1})$ ).
- ▶ We use logarithms of firm size (i.e. proxied by firm sales) to reduce the influence of the outliers.
- ▶  $d_t$ - a set of time dummies that capture the aggregate time effects (two-way-effects model).

## Why is OLS not appropriate?

(1) The OLS estimation in our case is very likely to suffer from the "**omitted variable problem**" due to factors that are not included in the regression (i.e. unobserved heterogeneity captured by the error term) and that have an impact on both the regressor ( $x_{it}$ ) and the dependent variable ( $y_{it}$ ).

- ▶ **If the error term is correlated with  $x_{it}$ , the OLS coefficients  $\beta_j$  will be BIASED and INCONSISTENT** (Violation of the Assumption 3).

**The main advantage in using the PANEL data is exactly the fact that it allows for the correlation between  $a_i$  and  $x_{it}$ .**

This is important since, even if we assume that  $u_{it}$  is not correlated with  $x_{it}$ ,  $\beta_1$  estimated with pooled OLS will still be biased and inconsistent if the  $a_i$  is correlated with  $x_{it}$ .

## Why is OLS not appropriate?

(2) Due to unobserved time-invariant heterogeneity ( $a_i$ ), the **standard errors of pooled OLS are incorrect** because of the serial correlation in the error term.

The OLS estimator is not as efficient as the estimators that account for serial correlation in errors (efficiency gains possible).

# The composite error

Generally, we can classify the unobserved factors ( $v_{it}$ ) as time-variant and time-invariant factors as follows (for a simple model):

$$y_{it} = \beta_0 + \delta_0 d_t + \beta_1 x_{it} + v_{it}; \quad i = 1, \dots, N, \quad t = 1, 2, \dots, T$$

$$y_{it} = \beta_0 + \delta_0 d_t + \beta_1 x_{it} + a_i + u_{it}, \quad i = 1, \dots, N, \quad t = 1, 2, \dots, T$$

$v_{it}$  is the composite error, which consists of:

- ▶  $a_i$  - the unobserved time-invariant effect (i.e. fixed effect, such as industry, type of privatization, etc.).
- ▶  $u_{it}$  - the idiosyncratic (time-varying) error.

# Cross-sectional time-invariant heterogeneity

Examples of  $a_i$ :

- ▶ Unobserved workers' ability in a regression model estimating the return on education.
- ▶ Geographical features, historical factors, demographic features of the populations etc. in a regression model estimating the impact of the unemployment on the level of city crime.
- ▶ Unobserved managerial skills in a regression estimating the impact of ownership structure on firm performance.

## Simple solutions (two periods)

We can "get rid" of the time-invariant effect by **DIFFERENCING THE DATA** over time, to obtain:

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

(**first-differenced equation/estimator**)

and apply OLS under the assumption that  $\Delta u_i$  is not correlated with  $\Delta x_i$ ,  
that  $x_i$  changes over time, and has some variation across  $i$ .

- ▶ **Note** that differencing reduces the variation in  $x_i$ , which can lead to large S.E. Thus, (in order to increase the variation in  $\Delta x_i$ ), use longer differences over time.
- ▶  $\delta_0$  is actually the change in the intercept between the two periods.

## Differencing with more than two periods

The same applies as for the two-period, assuming that the strict exogeneity assumption holds.

Take care not to create "bogus" observations by creating differences between observation  $i$  and  $j$ .

Account for the possibility that the first differences of the original errors ( $\Delta u_i$ ) are serially correlated:

- ▶ Use feasible GLS to correct for serial correlation in the differences in errors (i.e. when  $u_i$  follows a stable AR(1) model).
- ▶ Correct standard errors (i.e. clustered standard errors or robust standard errors for heteroskedasticity);

# Assumptions for Pooled OLS using First Differences

- (1) For each  $i$ , the model is  $y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \dots a_i + u_{it}$ ,  $t = 1, \dots T$ .
- (2) We have a random sample from the cross section.
- (3) **Strictly exogenous explanatory variables** conditional on the unobserved effect:  $E(u_{it} | X_i, a_i) = 0$ , which implies  $E(\Delta u_{it} | X_i, a_i) = 0$ ,  $t = 2, \dots T$ .
- (4) Each explanatory variable changes over time (for at least some  $i$ ), and no perfect linearity relationships exist among the explanatory variables.
- (5) Homoskedasticity:  $\text{Var}(\Delta u_{it} | X_i) = \sigma^2$ ,  $t = 2, \dots T$ .
- (6) No serial correlation in the differenced errors:  
 $\text{Cov}(\Delta u_{it}, \Delta u_{is}, | X_i) = 0$ ,  $t \neq s$ .
- (7) Conditional on  $x$ ,  $\Delta u_{it}$  are independent and identically distributed normal random variables.

# First-differenced estimators

## **Example: The factors influencing the extent of ownership concentration in a privatized firm**

We dispose with data on ownership, performance, size, percentage of tangible assets in total assets, leverage, industry and listing for a sample of 484 privatized firms in the period 2000 to 2004.

First assume that we only have two periods (2000 and 2004). Let's compare:

- ▶ OLS estimates separately for each period.
- ▶ Pooled OLS with clustered-robust standard errors.
- ▶ First-differenced estimates that single out any time-invariant firm-specific effects ( $\alpha_i$ ).

## Descriptive statistics

	Mean(Sd)	Mean(Sd)
VARIABLES	2000	2004
C1 (largest owner's share) in %	36.78 (20.58)	49.88 (27.46)
Firm sales ( $10^9$ SIT)	4952667 (15000000)	7015585 (2280000)
Tangible assets in total assets (share)	0.52 (0.22)	0.46 (0.23)
Risk (ratio)	4.04 (15.16)	3.71 (10.12)
Leverage (%)	37.30 (20.68)	42.51 (21.67)
Roa (in %)	1.33 (6.64)	1.07 (8.45)
Listing dummy	0.16 (0.37)	0.16 (0.366)

# Results

	Dependent variable: logitc1		Dependent variable: logitc1	Dependent variable: logitc1
	2000	2004	Pooled OLS	First-differences
ln (sales)	-0.007 (0.070)	0.027 (0.060)	0.023 (0.048)	-0.282* (0.070)
Tangible assets in total assets	0.151 (0.289)	0.033 (0.365)	-0.111 (0.241)	-1.42** (0.563)
Risk	-0.002 (0.002)	-0.009* (0.005)	-0.006** (0.002)	-0.001 (0.003)
Leverage (%)	-0.001 (0.002)	-0.002 (0.003)	0.001 (0.002)	-0.0001 (0.006)
Roa (in %)	-0.013* (0.007)	-0.003 (0.012)	-0.009 (0.008)	0.004 (0.009)
Listing dummy	-0.812*** (0.115)	-0.959*** (0.191)	-0.868*** (0.116)	
Constant	-0.704 (1.018)	0.152 (0.926)	-0.418 (0.728)	0.755*** (0.074)
Industry dummies	Yes	Yes	Yes	
Observations	484	484	967	484
R-Squared	0.10	0.08	0.07	0.05

## Note on clustered standard errors

Note that the error term ( $u_{it}$ ) in panel data is likely to be correlated over time for a given individual:

- ▶ Should use clustered standard errors that cluster on individual since the default standard errors for OLS assume i.i.d. errors (and are thus misleadingly small).
- ▶ The difference between default and clustered standard errors for pooled OLS can be very large and increases with increasing  $T$ , increasing autocorrelation in model errors, and increasing autocorrelation of the regressor of interest.
- ▶ Clustered errors require that  $N \rightarrow \infty$  and that errors are independent over  $i$  (individual unit, i.e. individual), or a more aggregate level (i.e. country/household).

## Panel data methods

Before looking at more advanced solutions, let us first prepare our data:

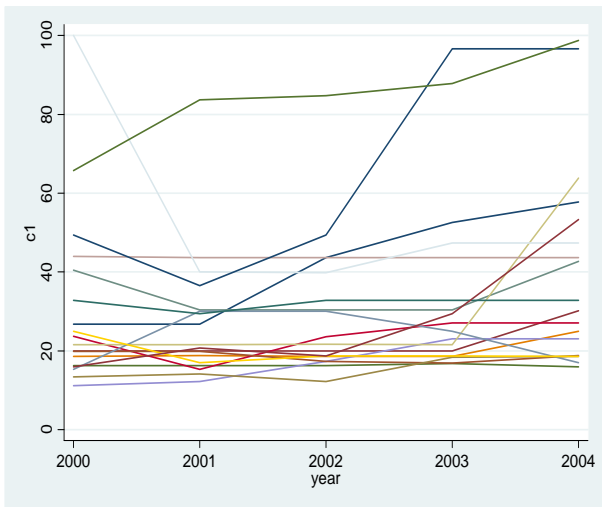
1. Arrange the data in the correct form: several records per unit, one record per year.
2. Declare panel and time identifiers (*xtset panelid timeid*).
3. Check the variation in the dependent and independent variables over time and over individuals (*xtsum*).
  - ▶ Plot separate time-series plots for some of the individual units:
  - ▶ (*quietly xtline c1 if idcode<=5037816&idcode>5034523, overlay legend(off)*)

# Panel data methods

```
. xtset idcode year
      panel variable:  idcode (strongly balanced)
      time variable:  year, 2000 to 2004
      delta: 1 year
```

```
list idcode year c1 sales tang_a
```

	idcode	year	c1	sales	tang_a
1.	1214381	2000	80.647	2705263	.7742845
2.	1214381	2001	39.876	3181732	.7969728
3.	1214381	2002	19.028	3865465	.8484114
4.	1214381	2003	25.662	3706275	.8115399
5.	1214381	2004	27.498	3807850	.7230686
6.	1318829	2000	52	3205191	.0758154
7.	1318829	2001	52	3670199	.0713515
8.	1318829	2002	52	3972128	.0531684
9.	1318829	2003	52	4417878	.0391962
10.	1318829	2004	52	5010137	.0386127



# Within and between variation (variance decomposition)

```
xtsum idcode year c1 sales tang_a leverage list
```

Variable		Mean	Std. Dev.	Min	Max	Observations	
idcode	overall	5125325	313127.3	1214381	5971101	N =	2420
	between		313386.5	1214381	5971101	n =	484
	within		0	5125325	5125325	T =	5
year	overall	2002	1.414506	2000	2004	N =	2420
	between		0	2002	2002	n =	484
	within		1.414506	2000	2004	T =	5
c1	overall	43.24797	24.35096	4.428	99.998	N =	2420
	between		21.33252	6.422379	99.1978	n =	484
	within		11.77476	-13.71143	108.5254	T =	5
sales	overall	5987713	1.86e+07	338	3.21e+08	N =	2420
	between		1.83e+07	2568.2	2.76e+08	n =	484
	within		3381635	-4.55e+07	8.65e+07	T =	5
tang_a	overall	.4863972	.2228702	0	.9859735	N =	2420
	between		.2129004	.0025596	.9689188	n =	484
	within		.0664795	-.0134246	.9352757	T =	5
list	overall	.1590909	.3658364	0	1	N =	2420
	between		.3661393	0	1	n =	484
	within		0	.1590909	.1590909	T =	5

## Panel data methods

Lets first start with pooled OLS regression with cluster-robust standard errors:

- ▶ regress  $y$   $x_1$   $x_2 \dots x_n$  ,  $T$  (*time dummies*), vce (cluster idcode)

Alternatively, we could apply the population-average estimators (pooled FGLS estimators), which still assume that regressor are exogenous but allow different assumptions about the corellation structure of the error term ( $u_{it}$ ), and are thus more efficient than OLS:

- ▶ xtreg, pa corr(independent) vce(robust), which equals pooled OLS.
- ▶ xtreg, pa corr(exchangeable) vce(robust), which is asytmotically equivalent to RE.
- ▶ xtreg, pa corr(ar k) vce(robust), which assumes an autoregressive process of order k for  $u_{it}$ .

## Panel data methods

However, the OLS or PA estimator may be BIASED when the composite error is correlated with the  $x_{it}$ .

Apart from first-differencing, two ways of improving the estimation have been suggested:

1. Fixed effects estimator (FE estimator) or within-estimator;
2. Random effects estimator (RE estimator);

(1) The **FIXED EFFECTS estimator (FE)** uses a transformation to remove the unobserved effect  $a_i$  prior to estimation. This estimator relies on the **time-variation in dependent and explanatory variables within each cross-sectional observation (!)**.

# Fixed and Random Effects Estimators

(2) **RANDOM EFFECTS estimator (RE)** is based on the assumption that **unobserved effects  $a_i$  are not correlated with  $x_{it}$** .

- ▶ With RE estimator it is possible to estimate the coefficients for the time-invariant variables, which is not the case when applying the FE estimator.
- ▶ However, note that with FE you can still estimate the change in the impact of a time-invariant variable by interacting this variable with time dummies.

# Fixed and Random Effects Estimators

## Additional notes:

- ▶ Conditional on the assumption that effects  $a_i$  are not correlated with  $X_{it}$ , RE is more efficient than FE estimator.
- ▶ RE estimator permits the serial correlation in the error term (i.e. serial correlation restricted to be the same at all lags: exchangeable errors).
- ▶ For both methods to work,  $N$  must be sufficiently large relative to  $T$ .

## Fixed effect estimator

For simplicity, consider a model with a single explanatory variable:

$$y_{it} = \beta_1 x_{it} + a_i + u_{it}, \quad t = 1, 2, \dots, T$$

for each  $i$ , average this equation over time (i.e.  $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}$ )

$$\bar{y}_{it} = \beta_1 \bar{x}_i + a_i + \bar{u}_i, \quad t = 1, 2, \dots, T$$

finally, subtract the second equation from the first one to get:

$$y_{it} - \bar{y}_{it} = \beta_1 (x_{it} - \bar{x}_i) + (a_i - a_i) + (u_{it} - \bar{u}_i)$$

which is then estimated by OLS (correct df to NT-N-k).

- ▶ Under the assumption that the idiosyncratic error is uncorrelated with each explanatory variables across all time periods (**strict exogeneity**), the FE estimator is **unbiased** and **consistent**. For the OLS analysis to be valid, we also need the assumption that idiosyncratic errors are homoskedastic and serially uncorrelated.

## Fixed -effects estimator

The fixed effect estimator can be obtained by the DUMMY variable regression, namely by including a dummy variable for each cross-sectional unit. This is however not very practical when  $N$  is large, and it still requires panel data.

Note that if the statistical package (i.e. Stata) reports the intercept with FE estimates, then this is the average across  $i$  for  $a_i$ .

### When to use FE and when to use first-differencing?

- ▶ for  $T = 2$  FE and first differencing estimates are identical.
- ▶ for  $T = 3$  both are unbiased and consistent but FE is efficient when  $u_i$  are serially uncorrelated and homoscedastic.
- ▶ for  $T > 3$ , serially positively correlated idiosyncratic errors and uncorrelated differences ( $\Delta u_i$ ), first -differencing is more efficient.
- ▶ for large  $T$ , the FE is less sensitive to strict exogeneity assumption.

# Assumptions for FE

- (1) For each  $i$ , the model is  $y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \dots a_i + u_{it}$ ,  $t = 1, \dots, T$ .
- (2) We have a random sample in cross-sectional dimensions.
- (3) **Strictly exogenous explanatory variables** conditional on the unobserved effect:  $E(u_{it} | X_i, a_i) = 0$ .
- (4) Each explanatory variable changes over time (for at least some  $i$ ), and no perfect linearity relationships exist among the explanatory variables.
- (5) Homoskedasticity:  $Var(u_{it} | X_i) = \sigma^2$ ,  $t = 2, \dots, T$ .
- (6) No serial correlation in the errors:  $Cov(u_{it}, u_{is}, | X_i) = 0$ ,  $t \neq s$ .
- (7) Conditional on  $x$ ,  $a_i$  and  $u_{it}$  are independent and identically distributed as normal.

# Example in Stata (FE)

```
xtreg logitc1 lnsales tang_a leverage risk roa list 102 103 104, fe vce (cluster idcode)
```

```
Fixed-effects (within) regression      Number of obs      =      2417  
Group variable: idcode                 Number of groups   =      484
```

```
R-sq:  within = 0.1359                Obs per group: min =      4  
        between = 0.0015                avg =      5.0  
        overall = 0.0113                max =      5
```

```
corr(u_i, Xb) = -0.2258                F(8,483)           =      16.22  
                                                Prob > F            =      0.0000
```

(Std. Err. adjusted for 484 clusters in idcode)

logitc1	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
lnsales	-.1501619	.1182895	-1.27	0.205	-.3825874	.0822636
tang_a	-1.063162	.394739	-2.69	0.007	-1.83878	-.2875443
leverage	-.0003393	.0036698	-0.09	0.926	-.0075499	.0068714
risk	.0005677	.00185	0.31	0.759	-.0030673	.0042028
roa	.0001584	.0029305	0.05	0.957	-.0055997	.0059165
list	(dropped)					
102	.2138944	.0358379	5.97	0.000	.1434769	.2843118
103	.3706559	.0449681	8.24	0.000	.2822986	.4590131
104	.6642239	.0645636	10.29	0.000	.5373637	.7910842
<b>cons</b>	<b>2.1921</b>	<b>1.724856</b>	<b>1.27</b>	<b>0.204</b>	<b>-1.197047</b>	<b>5.581248</b>
<b>sigma_u</b>	<b>1.2005239</b>					
<b>sigma_e</b>	<b>.76965136</b>					
<b>rho</b>	<b>.70871502</b>	<b>(fraction of variance due to u_i)</b>				

# Example in Stata (Dummy variable regression)

```
areg logitc1 lnsales tang_a leverage roa risk list 102 103 104, absorb(idcode)  
vce(cluster idcode)
```

```
Linear regression, absorbing indicators                                Number of obs =      2417  
                                                                F(   8,   483) =    12.97  
                                                                Prob > F       =    0.0000  
                                                                R-squared     =    0.7460  
                                                                Adj R-squared =    0.6812  
                                                                Root MSE     =    .76965
```

(Std. Err. adjusted for 484 clusters in idcode)

logitc1	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
lnsales	-.1501619	.1322997	-1.14	0.257	-.410116	.1097922
tang_a	-1.063162	.4414921	-2.41	0.016	-1.930644	-.1956798
leverage	-.0003393	.0041044	-0.08	0.934	-.008404	.0077254
roa	.0001584	.0032776	0.05	0.961	-.0062817	.0065984
risk	.0005677	.0020691	0.27	0.784	-.0034979	.0046334
list	(dropped)					
102	.2138944	.0400825	5.34	0.000	.1351367	.2926521
103	.3706559	.0502941	7.37	0.000	.2718336	.4694781
104	.6642239	.0722106	9.20	0.000	.5223383	.8061096
_cons	2.1921	1.929148	1.14	0.256	-1.598459	5.98266

idcode | absorbed (484 categories)

## Between estimator

Performing an OLS on the time-averages of  $y_i$  and the explanatory variables  $x_i$ , gives us the BETWEEN ESTIMATOR, namely:

$$\bar{y}_{it} = \beta_0 + \beta_1 \bar{x}_i + a_i + \bar{u}_i, \quad t = 1, 2, \dots, T$$

- ▶ Note that with the between estimator the  $a_i$  term is still there  $\rightarrow$  still **INCONSISTENT** when  $a_i$  is correlated with  $\bar{x}_i$ .
- ▶ Note that since only cross-section variation in the data is used, the coefficients of any individual-invariant regressors (i.e. time dummies) cannot be identified.
- ▶ Assuming that  $a_i$  is not correlated with  $\bar{x}_i$ , one should rather use RE estimator since BE disregards important information on how the variables change over time.

# Random Effects Estimator

Consider a model with a single explanatory variable and assume that  $a_i$  is not correlated with  $x_j$ :

$$y_{it} = \beta_1 x_{it} + a_i + u_{it}, \quad t = 1, 2, \dots, T$$

for each  $i$ , average this equation over time (i.e.  $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}$ )

$$\bar{y}_{it} = \beta_1 \bar{x}_i + a_i + \bar{u}_i, \quad t = 1, 2, \dots, T$$

finally, subtract A FRACTION of the time-average from the first equation to get:

$$y_{it} - \lambda \bar{y}_{it} = \beta_0 (1 - \lambda) + \beta_1 (x_{it} - \lambda \bar{x}_i) + (v_{it} - \lambda \bar{v}_i), \quad t = 1, 2, \dots, T$$

which is then estimated by pooled OLS ( $\rightarrow$  **feasible GLS estimator**).

Due to  $a_i$  being part of the composite error and fixed across the same cross-sectional unit, the composite errors  $v_{it}$  are serially correlated across time, namely:

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

In order to account for this serial correlation, we define  $\lambda$  as:

$$\lambda = 1 - \left( \frac{\sigma_u^2}{(\sigma_u^2 + T\sigma_a^2)} \right)^{1/2}$$

to obtain the transformed equation (see previous slide).

- ▶ Under the random effects assumptions (see next slide), the RE estimator is CONSISTENT (but not unbiased) and FULLY EFFICIENT.

# Comparison between different estimators

Note that:

- ▶  $\lambda = 0 \rightarrow$  pooled OLS estimator.
- ▶  $\lambda = 1 \rightarrow$  fixed effects estimator.
- ▶  $\lambda$  *close to 0* when the unobserved effect  $a_i$  is relatively unimportant  $\rightarrow$  RE and OLS estimates very similar.
- ▶ when T becomes very large,  $\lambda$  goes towards 0  $\rightarrow$  RE and FE estimates very similar.

# Assumptions for RE

(1) For each  $i$ , the model is  $y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \dots a_i + u_{it}$ ,  $t = 1, \dots T$ .

(2) We have a random sample in cross-sectional dimensions.

(3) **Strictly exogenous explanatory variables** conditional on the unobserved effect:  $E(u_{it} | X_i, a_i) = 0$ .

(3a) The expected value of  $a_i$  given all explanatory variables is **constant**:  
 $E(a_i | X_i) = \beta_0$ .

(4) There are no perfect linear relationships among the explanatory variables.

(5) Homoskedasticity:  $\text{Var}(u_{it} | X_i) = \sigma_u^2$ ,  $t = 2, \dots T$ .

(5a)  $\text{Var}(a_i | X_i) = \sigma_a^2$ ,  $t = 2, \dots T$ .

(6) No serial correlation in the errors:  $\text{Cov}(u_{it}, u_{is}, | X_i) = 0$ ,  $t \neq s$ .

# Comparison of results

	Dependent variable: logit C1	Dependent variable: logit C1	Dependent variable: logit C1	Dependent variable: logit C1
	Pooled OLS (Clustered S.E.)	FE (Clustered S.E.)	BE (Clustered S.E.)	RE (Clustered S.E.)
In (sales)	-0.009 (0.043)	-0.150 (0.117)	0.054 (0.05)	-0.072** (0.062)
Tangible assets in total assets	0.140 (0.247)	-1.063*** (0.395)	0.078 (0.261)	-0.397** (0.265)
Risk	-0.005* (0.002)	0.001 (.002)	-0.009*** (0.003)	-0.001 (0.002)
Leverage (%)	-0.001 (0.002)	-0.0003 (0.004)	-0.002 (0.003)	-0.0003 (0.002)
Roa (in %)	-0.013** (0.005)	0.0002 (0.002)	-0.035*** (0.012)	-0.002 (0.003)
Listing dummy	-0.800*** (0.119)	dropped	-0.767*** (0.135)	-0.789*** (0.121)
Constant	-0.780 (0.648)	2.192 (1.725)	-3.037 (2.084)	0.679 (0.938)
Industry dummies	Yes	-	Yes	Yes
Time dummies	Yes	Yes	Yes	Yes
Observations	2417	2417	2417	2417
R-Squared	0.11	0.011	0.01	0.09
R-within		0.14	0.01	0.13
R-between		0.001	0.13	0.08



## Comparison between different estimators

Assuming that that  $a_i$  is not correlated with  $x$ , can coefficient  $\beta$  be consistently estimated with pooled OLS ?

→ Yes, but since OLS ignores the positive serial correlation in the error term (due to  $a_i$ ), it is not efficient unless  $\sigma_a^2 = 0$ .

Why using RE?

- ▶ More information is used.
- ▶ Accounts for the serial correlation in the error term and is thus, more efficient.

# Comparison between different estimators

When to use OLS rather than RE?

- ▶ In the case that the list of included explanatory variables is so exhausting that there are no individual effect ( $a_i$ ), use OLS since more efficient.
- ▶ Use the Breusch-Pagan Lagrange multiplier test to check for the presence of random effects.

# When to use FE and when RE?

Why not just using FE (within estimator) all the time?

- ▶ Because of efficiency:  $\text{Var}(\hat{\beta}_{FE}) \geq \text{Var}(\hat{\beta}_{RE})$ .
- ▶ Because FE does not allow us to estimate the coefficients for time-invariant variables.

When can we use RE?

- ▶ **HAUSMAN TEST**

Under the null hypothesis:  $H_0 : E(a_i | x_{it}) = 0$

Under the alternative hypothesis:  $H_1 : E(a_i | x_{it}) \neq 0$

# Hausman test

If  $H_0$  can not be rejected  $\rightarrow$  RE and FE are both consistent but FE inefficient since it involves estimating  $N$  dummy variables.

If  $H_0$  is false, the RE estimates are subject to unobserved heterogeneity bias, and will thus systematically differ from the FE estimates.

The Hausman test in fact looks at the RE and FE estimates to determine whether the estimates (taken as a group) are significantly different between the FE and RE regression.

- ▶ Note that any time-invariant variables that are dropped from the FE regression have to be excluded from the RE regression when performing the Hausman test.

# Hausman test

```
hausman fe re, sigmamore
```

	---- Coefficients ----			
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	fe	re	Difference	S.E.
lnsales	-.1501619	-.0912047	-.0589572	.0383731
tang_a	-1.063162	-.2666242	-.7965379	.1792689
leverage	-.0003393	.0010383	-.0013775	.0016231
roa	.0001584	-.0015841	.0017425	.000711
risk	.0005677	-.0014091	.0019769	.0011378
102	.2138944	.2219354	-.008041	.0062122
103	.3706559	.3791884	-.0085325	.0092488
104	.6642239	.683561	-.019337	.0125211

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

$\chi^2(8) = (b-B)'[(V_b-V_B)^{-1}](b-B)$   
= 42.84  
Prob>chi2 = 0.0000

## Unbalanced panels

Generally, FE and RE methods apply also in the case of unbalanced panels.

Note that in order for the estimators to be CONSISTENT we need to make sure that the reason for which some time periods are missing for some of the observations (i.e. a firm leaves the sample - attrition) is not be systematically related to the idiosyncratic errors ( $u_{it}$ ). If yes, need to correct for **SAMPLE SELECTION BIAS**.

Note however that FE analysis does allow attrition to be correlated with the unobserved effect  $a_i$ .

Note that there are other solutions to address unobserved heterogeneity in cross-sections:

1. Add more controls, if possible.
2. Find a proxy variable for  $a_i$  (i.e. include  $y_{it-1}$ ).
3. Find instruments for the elements of the  $x_{it}$  that are correlated with  $a_i$ , and use the instrumental variable procedure.
  - ▶ IV estimation can be combined with panel data methods (i.e. first-differencing), to consistently estimate parameters in the presence of unobserved effects and endogeneity in the time-varying explanatory variables. For more, see Wooldridge (2002, 2003).

## Further issues

Note that for the panel data estimators to be CONSISTENT, there should be no correlation between the explanatory variables and the remaining error, once we control for  $a_j$ :

### **STRICT EXOGENEITY ASSUMPTION:**

$Cov(x_{itj}, u_{it}) = 0$ , for all  $t, s$  and  $j$ .

- ▶ This assumption is violated when we omit an important time-varying variable or when future explanatory variables react to current changes in the idiosyncratic errors (i.e. when we include lagged dependent variable as the explanatory variable).
- ▶ This is often the case in corporate governance research.
- ▶ There are tests of strict exogeneity that you can use (see Wooldridge, 2002).

## Further issues: Violation of strict exogeneity

### **Example: Estimating the impact of board structure on firm performance**

$$performance_{it} = \beta_1 GOV_{it} + a_i + u_{it}, \quad t = 1, 2, \dots, T, \quad i = 1, 2, \dots, N.$$

A number of studies estimating this equation assume that past performance has no effect on either on current performance ( $y$ ) or current governance ( $x$ ). However, "*a major determinant of firm's governance and contracting environment is the firm's history and past performance, and failure to account for this may lead to biased inferences*" (Wintoki et al., 2008).

## Further issues: Violation of strict exogeneity

**Sources of endogeneity** in the estimation of the relation between firm governance and performance:

(1) There may be some other unobserved factors that influence both, firm governance and performance, and that can not be proxied by firm past performance (**unobserved heterogeneity**).

(2) Past performance is a proxy for unobservable factors that affect both current governance and performance, i.e. managerial ability, contracting environment, etc. (**dynamic endogeneity**).

(3) It is also possible that within any time period the actors in the firm's nexus of contract chose their governance characteristics (i.e. board structure) in anticipation of their expected performance (**simultaneity**).

# Solution

(1) Rely on a dynamic model that includes lagged performance as an explanatory variable. However:

- ▶ OLS estimation of a dynamic model does not account for unobservable heterogeneity and simultaneity.
- ▶ FE estimators will be inconsistent since the strict exogeneity assumption is violated due to the inclusion of the lagged dependent variable.

## Solution

(2) Estimate the model via GMM using lagged values of the governance variables and performance as INSTRUMENTS (**dynamic panel GMM estimator**).

The estimation involves 2 steps. First, we write the dynamic model in a first-differenced form to eliminate unobserved heterogeneity, and then estimate this model via GMM using lagged values of the explanatory variables as instruments for the current explanatory variables.

The assumption is that a firm's history is a valid instrument for its current governance structure (i.e we use sufficient lags). For more, see Arellano and Bond (1991), Arellano and Bover (1995), Blundell and Bond (1998). See also system GMM estimators, which include both the equations and levels and differences.