NOVA IMS INFORMATION MANAGEMENT SCHOOL

Pooled OLS.
Fixed and Random Effects.

Statistics for Data Science

Bruno Damásio
bdamasio@novaims.unl.pt
@bmpdamasio

2025/2026

Table of contents

Motivation

What is panel data?

- So far, we have covered cross-sectional and time series data.
- Panel data have a cross-sectional and a time series dimension, where the cross-sectional units are followed over time.
- Panel data can be used to account for time-invariant unobservables.
- For a linear regression model with panel data, we have:

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it} \tag{1}$$

  where y_{it} and x_{it} are observed for $i = 1, \ldots, N$ individuals (firms, countries, stocks, ...) over $t = 1, \ldots, T$ time periods (days, months, quarters, years, ...)
- Panel data helps to address issues that cannot be tackled with pure cross-sectional or time-series data.

Motivation

Why not consider the univariate cross-sectional linear regression model?

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{2}$$

The main challenge in identifying the effect of x on y from a univariate cross-sectional model is that individuals in the sample not only differ in terms of x, but also differ along other dimensions.

The textbook example: Crime and Unemployment

Cross-Sectional Regression is biased

- Data for 46 cities in 1982 and 1987

```
##
## Call:
## lm(formula = crmrte ~ unem, data = crime2)
##
## Coefficients:
## (Intercept)         unem
##    103.2434      -0.3077
```

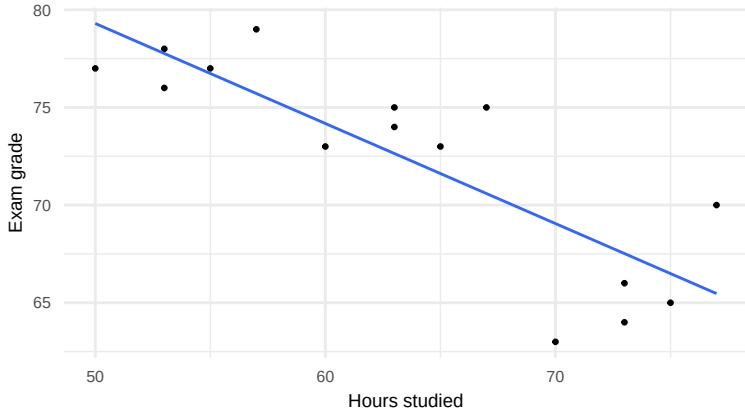
- Higher unemployment decreases crime?

The textbook example: Crime and Unemployment

Omitted Variable Problem

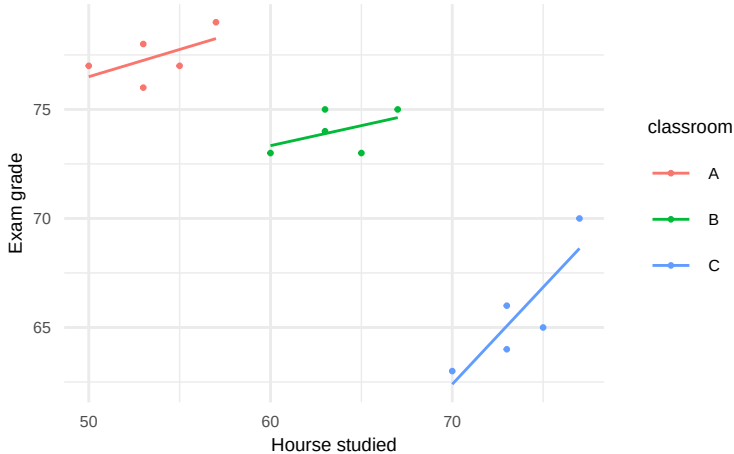
- Problem: omitted variables, such as size, age distribution, education, law enforcement, geography, poverty rates, social security, etc,
- For some complex combination of many different observable and unobservable factors, some cities have much higher average crime rates than other cities.

Another example: Grades and Hours studied



Studying more decreases grades?

Another example: Grades and Hours studied



Omitted Variable Problem

- Thus, when using panel data, the main motivation is to solve the omitted variables problem.
- Let y and $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ be observable random variables, and let c be an **unobservable random variable** or **unobserved heterogeneity**.
- As is often the case, we are interested in the partial effects of the observable explanatory variable x_j in the population regression function:

$$E[y \mid x_1, x_2, \ldots, x_k, c]$$

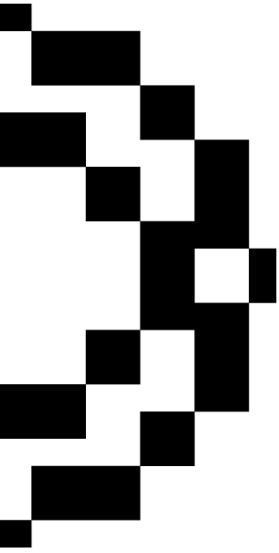
- In other words, we would like to hold c constant when obtaining partial effects of the observable explanatory variables.

Omitted Variable Problem

- Assuming a linear model, with c entering additively along with the x_j, we have:

$$\mathrm{E}[y \mid \mathbf{x}, c] = \beta_0 + \mathbf{x}\boldsymbol{\beta} + c$$

- If c is uncorrelated with each x_j, then c is just another unobserved factor affecting y that is not systematically related to the observable explanatory variables whose effects are of interest.
- On the other hand, if $Cov(x_j, c) \neq 0$ for some j, putting c into the error term can cause some problems.



Estimators

Pooled OLS Estimation

- A pooled OLS simply applies the OLS for a sample of $N \times T$ observations, ignoring the panel structure of the data.
- Under certain assumptions, the pooled OLS estimator can be used to obtain a consistent estimator of $\boldsymbol{\beta}$ in the model below:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}, t = 1, 2, \ldots, T, i = 1, 2, \ldots, N \tag{3}$$

  where $v_{it} = c_i + u_{it}$ are the **composite errors**
- For each t, v_{it} is the sum of the unobserved effect and an idiosyncratic error (u_{it}).

Pooled OLS estimation is appropriate to use when the regressors are not correlated with c_i.

Pooled OLS Assumptions

POLS.1: Linearity

The data sequence $\{y_{it}, x_{1it}, x_{2it}, \ldots, x_{kit}\}$ follows the linear model:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}, t = 1, 2, \ldots, T, i = 1, 2, \ldots, N \tag{4}$$

where $\{v_{it}\}$ is the sequence of composite errors.

POLS.2: Random Sampling

The individuals are select by random sampling. This implies that $\{y_i, x_i\}$ are i.i.d., i.e., the observations are independent across individuals but not necessarily across time.

Pooled OLS Assumptions

POLS.3: Contemporaneous exogeneity

$E[v_{it} \mid \mathbf{x}_{it}] = 0$, which implies that
$E[\mathbf{x}_{it} c_i] = 0$ and $E[\mathbf{x}_{it} u_{it}] \Rightarrow E[\mathbf{x}_{it} v_{it}] = 0$

POLS.4: No Perfect Collinearity

In the sample, no independent variable is constant nor a perfect linear combination of the others.

Pooled OLS Properties

Consistency

Knowing that we have a sample of N independent cross sections observed during T periods of time and that c_i is independent of u_{it}, under POLS1-POLS4, the pooled OLS is consistent, however if c is correlated with any element of $\mathbf{x}_t$, then pooled OLS is biased and inconsistent.

Pooled OLS: Assumptions

POLS.5: Homoskedasticity and no serial correlation

$\text{Var}(u_i|\mathbf{x}_i, c_i) = \sigma_u^2 \mathbf{I}_T$

- However, in this setting, the composite errors v_{it} will be serially correlated due to the presence of c_i in each time period.
- Therefore, inference using pooled OLS requires the **robust variance matrix estimator**.

Example

Example

Suppose we want to study the factors that influence wage for working adults. Our regression is:

$$log(wage_{it}) = \beta_0 + \beta_1 educ_{it} + \beta_2 exper_{it} + \beta_3 married_{it} + \beta_4 union_{it} + c_i + u_{it} \quad (5)$$

where *educ* is the number of years of schooling, *exper* the number of years working, *married* is a dummy variable that assumes the value 1 if the individual is married and *union* is a dummy variable that assumes the value 1 if the individual is unionized.

Pooled OLS - Example in R

```
#Pooled ols
library(plm)
pool <- plm(lwage ~ educ + exper
   + married + union, data=wagepan, model='pooling')
summary(pool)


## Pooling Model
##
## Call:
## plm(formula = lwage ~ educ + exper + married + union, data = wagepan,
##     model = "pooling")
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -5.232401 -0.253150  0.033334  0.301096  2.605001
##
## Coefficients:
##              Estimate Std. Error t-value  Pr(>|t|)
## (Intercept) 0.0189577  0.0618246  0.3066    0.7591
## educ        0.1032319  0.0044971 22.9553 < 2.2e-16 ***
## exper       0.0487251  0.0029001 16.8013 < 2.2e-16 ***
## married     0.1277006  0.0155485  8.2131 2.813e-16 ***
## union       0.1720027  0.0170683 10.0773 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1236.5
## Residual Sum of Squares: 1018.7
## R-Squared:      0.1762
```

Pooled OLS - Example in R

To obtain the robust standard errors:

```
library(lmtest)
coeftest(pool,vcov=vcovHC(pool,type="HC0",cluster="group"))


##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.0189577  0.1117643  0.1696    0.8653
## educ        0.1032319  0.0086614 11.9186 < 2.2e-16 ***
## exper       0.0487251  0.0039445 12.3526 < 2.2e-16 ***
## married     0.1277006  0.0254884  5.0101 5.655e-07 ***
## union       0.1720027  0.0278805  6.1693 7.484e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Random Effects Estimation

- As with pooled OLS, a **random effects analysis** puts c_i into the error term.
- The rationale behind random effects model is that the variation across individuals is assumed to be random and uncorrelated with the predictor or independent variables included in the model.

Random effects estimation is appropriate when c_i is random and influences the dependent variable but is not correlated with the independent variables.

Random Effects Properties

RE.1: Linearity

The data sequence $\{y_{it}, x_{1it}, x_{2it}, \ldots, x_{kit}\}$ follows the linear model:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}, t = 1, 2, \ldots, T, i = 1, 2, \ldots, N \tag{6}$$

where $\{v_{it}\}$ is the sequence of composite errors.

RE.2: Random Sampling

The individuals are select by random sampling. This implies that $\{y_i, x_i\}$ are i.i.d., i.e., the observations are independent across individuals but not necessarily across time.

Random Effects Properties

Random effects estimation imposes more assumptions than those needed for pooled OLS: strict exogeneity in addition to orthogonality between c_i and $\mathbf{x}_{it}$.

RE.3: Strict exogeneity

(a) $E[u_{it} \mid \mathbf{x}_i, c_i] = 0, t = 1, \ldots, T$

(b) $E[c_i \mid \mathbf{x}_i] = E[c_i] = 0$

where $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT})$

Random Effects

- The random effects approach exploits the serial correlation in the composite error, $v_{it} = c_i + u_{it}$, in a generalized least squares (GLS) framework.
- It can be shown that

$$E\left[v_{it}^2\right] = E\left[c_i^2\right] + 2E\left[c_i u_{it}\right] + E\left[u_{it}^2\right] = \sigma_c^2 + \sigma_u^2 \tag{7}$$

  and, for $t \neq s$ we have

$$E\left[v_{it}, v_{is}\right] = E\left[(c_i + u_{it})(c_i + u_{is})\right] = E\left[c_i^2\right] = \sigma_c^2 \tag{8}$$

- Hence, $\mathbf{\Omega}$ is given by

$$\begin{pmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \dots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \dots & \sigma_c^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_c^2 & \sigma_c^2 & \dots & \sigma_c^2 + \sigma_u^2 \end{pmatrix}$$

Random Effects

For consistency of GLS, we need the usual rank condition for GLS:

RE.4: No perfect colinearity

$\text{rank}(E\left[\mathbf{X}_i\boldsymbol{\Omega}^{-1}\mathbf{X}_i\right]) = K$

RE.5: Homoskedasticity

(a) $E\left[\mathbf{u}_i\mathbf{u}_i' \mid \mathbf{x}_i, c_i\right] = \sigma_u^2\mathbf{I}_T$

(b) $E\left[c_i^2 \mid \mathbf{x}_i\right] = \sigma_c^2$

Random Effects Properties

Consistency and Efficiency

Let $\hat{\beta}_{RE}$ be the **random effects estimator**:
Under assumptions RE.1 to RE.4 $\Rightarrow \hat{\beta}_{RE}$ is consistent.
Under assumptions RE.1 to RE.5 $\Rightarrow \hat{\beta}_{RE}$ is asymptotically efficient.

Random Effects - in R

```
library(plm)
library(wooldridge)
re <- plm(lwage ~ educ + exper
    + married + union, model='random', data=wagepan)
summary(re)


## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = lwage ~ educ + exper + married + union, data = wagepan,
##     model = "random")
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Effects:
##                   var std.dev share
## idiosyncratic 0.1250  0.3535 0.539
## individual    0.1070  0.3272 0.461
## theta: 0.6431
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -4.546962 -0.144396  0.025829  0.191661  1.555052
##
## Coefficients:
##               Estimate Std. Error z-value  Pr(>|z|)
## (Intercept) -0.0613771  0.1070096 -0.5736    0.5663
## educ         0.1082765  0.0087264 12.4079 < 2.2e-16 ***
## exper        0.0576255  0.0025010 23.0409 < 2.2e-16 ***
## married      0.0796721  0.0167171  4.7659 1.880e-06 ***
```

Fixed Effects Estimation

- Individuals have characteristics that may or may not influence the outcome and/or predictor variables.
- For example, the business practices of a company may influence its stock price or level of spending; attitudes or policies towards guns in a particular state may affect its levels of gun violence.

When the c_i may be correlated with the explanatory variables, we use the fixed effects estimation.

Fixed Effects Assumptions

Let's consider again the linear unobserved effects model for T time periods:

FE.1: Linearity

The data sequence $\{y_{it}, x_{1it}, x_{2it}, \ldots, x_{kit}\}$ follows the linear model:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, t = 1, 2, \ldots, T, i = 1, 2, \ldots, N \quad (9)$$

FE.2: Random Sampling

The individuals are select by random sampling. This implies that $\{y_i, x_i\}$ are i.i.d., i.e., the observations are independent across individuals but not necessarily across time.

Fixed Effects Assumptions

But relax the RE3.b assumptions:

FE.3: Strict exogeneity

$E[u_{it} \mid \mathbf{x}_i, c_i] = 0, t = 1, 2, \ldots, T$

Fixed Effects Assumptions

- In fixed effects analysis, $E[c_i \mid x_i]$ is allowed to be any function of $\mathbf{x}_i$.
- Essentially, we relax the assumption RE.3b, which allows to consistently estimate partial effects in the presence of time-constant omitted variables that can be arbitrarily related to the observable $\mathbf{x}_{it}$.
- Therefore, fixed effects analysis is more robust than random effects analysis.

Fixed Effects Estimation

The idea for estimating β under the Assumption FE.3 is to transform the equations to eliminate the unobserved effect c_i. This transformation is denoted the **fixed effects transformation** or **within transformation**.

Fixed Effects Estimation

Considering the mean values **within** each individual:

$$\bar{y}_i = T^{-1} \sum_{t=1}^{T} y_{it} \quad \bar{\mathbf{x}}_i = T^{-1} \sum_{t=1}^{T} \mathbf{x}_{it} \quad \bar{u}_i = T^{-1} \sum_{t=1}^{T} u_{it}$$

Subtracting these terms for each t gives the FE transformed equation:

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{x}_i)\boldsymbol{\beta} + u_{it} - \bar{u}_i \tag{10}$$

or

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it}, t = 1, 2, \ldots, T \tag{11}$$

Downside: cannot be used to investigate time-invariate causes of y

Fixed Effects Estimation

- Since the assumption $E\left[\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{u}}_{it}\right] = 0$ holds under Assumption FE.3, we can apply pooled OLS.
- The **fixed effects (FE) estimator**, denoted by $\hat{\beta}_{FE}$ is the pooled OLS estimator from the regression $\ddot{y}_{it}$ on $\ddot{\mathbf{x}}_{it}$.

Fixed Effects Assumptions

FE.4: No perfect collinearity

$\text{rank}(\sum_{t=1}^{T} E\left[\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\right]) = k$

FE.5: Homoskedasticity

$E\left[\mathbf{u}_i\mathbf{u}_i' \mid \mathbf{x}_i, c_i\right] = \sigma_u^2\mathbf{I}_T$

It can be shown that the errors ($\ddot{\mathbf{u}}_{it}$) are negatively serially correlated, however as T gets large, the correlation tends to zero.

Fixed Effects Properties

Consistency

Under FE.1 to FE.4, the fixed effects estimator is consistent.

Fixed Effects Estimation - in R

```
library(plm)
fe <- plm(lwage ~ educ + exper
    + married + union, model='within', data=wagepan)
summary(fe)


## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = lwage ~ educ + exper + married + union, data = wagepan,
##     model = "within")
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -4.14628 -0.12503  0.01232  0.16205  1.48094
##
## Coefficients:
##           Estimate Std. Error t-value  Pr(>|t|)
## exper   0.0598672  0.0025835 23.1726 < 2.2e-16 ***
## married 0.0610384  0.0182929  3.3367 0.0008558 ***
## union   0.0837910  0.0194140  4.3160 1.629e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    572.05
## Residual Sum of Squares: 476.43
## R-Squared:      0.16715
## Adj. R-Squared: 0.047646
## F-statistic: 255.026 on 3 and 3812 DF, p-value: < 2.22e-16
```

In summary

1. Pooled OLS
   - c_i is not correlated with x_{it}
   - Ignore panel data structure and estimate through OLS
   - Under POLS.1 to POLS.4, is consistent
2. Random effects
   - c_i is not correlated with x_{it}
   - Consider correlation structure and estimate through GLS
   - Under RE.1 to RE.5, is consistent and efficient
3. Fixed effects
   - c_i is correlated with x_{it}
   - Demean each variable in the regression and estimate through OLS
   - Under FE.1 to FE.4, is consistent

Hausman Test

Hausman test

- Since the key consideration in choosing between a random effects and fixed effects approach is whether c_i and x_{it} are correlated, it is important to have a method for testing this assumption.
- Hausman (1978) proposed a test based on the difference between the random effects and fixed effects estimates. Since FE is consistent when c_i and x_{it} are correlated, but RE is inconsistent, a statistically significant difference is interpreted as evidence against the random effects assumption RE.3b.

Hausman test

- The hypothesis are $H_0 : E[\mathbf{x}_{it} c_i] = 0$ (both FE and RE are consistent estimators, RE is the most efficient estimator) vs $H_1 : E[\mathbf{x}_{it} c_i] \neq 0$ (only FE is consistent).
- The test applies only to variables that vary on time.
- If the assumption RE.4 fails, one must use the Robust Hausman test.

Hausman test

```
phtest(fe, re)


##
##  Hausman Test
##
## data:  lwage ~ educ + exper + married + union
## chisq = 23.188, df = 3, p-value = 3.689e-05
## alternative hypothesis: one model is inconsistent
```

Hausman test

```
#Test heteroskedasticity (verify assumption RE.4)
library(lmtest)
bptest(lwage ~ educ + exper + expersq
    + married + union + factor(nr), data=wagepan)


##
##  studentized Breusch-Pagan test
##
## data:  lwage ~ educ + exper + expersq + married + union + factor(nr)
## BP = 761.47, df = 548, p-value = 3.716e-09
```

Robust Hausman test

```
phtest(fe, re,  vcov = function(x) vcovHC(x, method="white2", type="HC3"))


##
##  Hausman Test
##
## data:  lwage ~ educ + exper + married + union
## chisq = 23.188, df = 3, p-value = 3.689e-05
## alternative hypothesis: one model is inconsistent
```