Pooled and Panel Data Analysis

Panel Data Econometrics

Prof. Alexandre Gori Maia

State University of Campinas

Topics

Pooled Data

Fixed Effects – Binary Variables

Fixed Effects – Within Transformation

Reference

Baltagi, B. Econometric analysis of panel data. Third Edition. John Wiley & Sons. 2005, Chapters 1-4.

Wooldridge, J. M. 2001. Econometric analysis of cross section and panel data. Cap. 10.

Sample Designs

Cross-Sectional data

$$Y_i$$

$$i = 1,2,\ldots,n$$

$$\begin{array}{c} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{array}$$

Different units in a specific period of time

Time Series

$$Y_t$$

$$t = 1,2,\ldots,T$$

$$\begin{array}{cccc} Y_1 & Y_2 & \cdots & Y_T \end{array}$$

The same unit in different periods of time

Pooled Data

$$Y_{it}$$

$$i = 1,2,\ldots,n_T$$

$$t = 1,2,\ldots,T$$

$$\begin{array}{cccc} Y_{11} & Y_{12} & \cdots & Y_{1T} \\ Y_{21} & Y_{22} & & Y_{2T} \\ \cdots & \cdots & & \cdots \\ Y_{n_1 1} & Y_{n_2 2} & & Y_{n_T T} \end{array}$$

Cross-sectional samples (not necessarily the same) are observed in different periods of time

Panel Data

$$Y_{it}$$

$$i = 1,2,\ldots,n$$

$$t = 1,2,\ldots,T$$

$$\begin{array}{cccc} Y_{11} & Y_{12} & \cdots & Y_{1T} \\ Y_{21} & Y_{22} & \cdots & Y_{2T} \\ \cdots & \cdots & \cdots & \cdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nT} \end{array}$$

The same cross—sectional sample is observed in different periods of time

Panel Data - Examples

Balanced Panel Data

Y_{it}

$$\begin{array}{cccc} Y_{11} & Y_{12} & \cdots & Y_{1T} \\ Y_{21} & Y_{22} & \cdots & Y_{2T} \\ \cdots & \cdots & \cdots & \cdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nT} \end{array}$$

Each cross-sectional units is observed in all periods

Unbalanced Panel Data

Y_{it}

$$\begin{array}{cccc} Y_{11} & Y_{12} & \cdots & \\ Y_{21} & & \cdots & Y_{21} \\ \cdots & \cdots & \cdots & \cdots \\ & Y_{n2} & \cdots & Y_{n3} \end{array}$$

Some cross-sectional units are not observed in some periods

Rotating Panel Data

Y_{it}

$$\begin{array}{cccc} Y_{11} & & & \\ Y_{21} & Y_{22} & & \\ & Y_{32} & & \\ & & \cdots & Y_{n-1T} \\ & & & Y_{nT} \end{array}$$

Groups of cross-sectional units (*rotation groups*) are brought in and out of the sample in some periods.

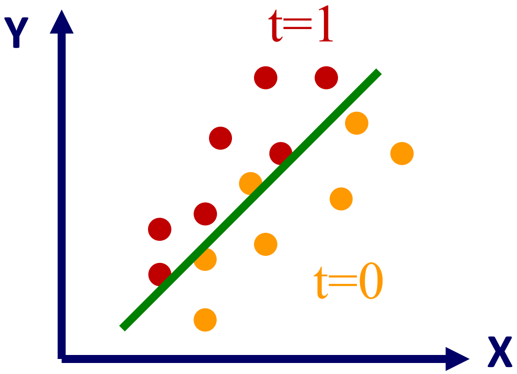
Split Panel

Y_{it}

$$\begin{array}{cccc} Y_{11} & Y_{12} & \cdots & Y_{1T} \\ Y_{21} & Y_{22} & \cdots & Y_{2T} \\ \cdots & \cdots & & \cdots \\ Y_{n_1 1} & & & \\ & Y_{n_2 2} & \cdots & Y_{n_T T} \end{array}$$

Combines cross-sectional and panel samples at each period.

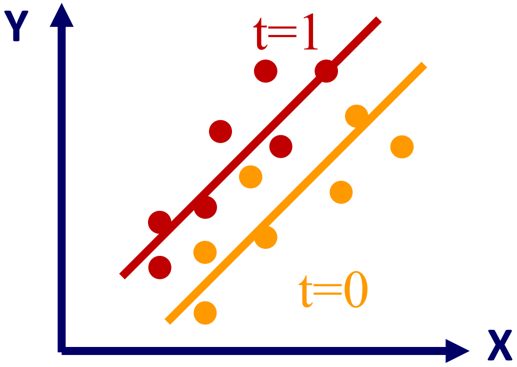
Regression with Pooled Data



Constant intercept and slope coefficients

$$Y = \alpha + \beta X + e$$

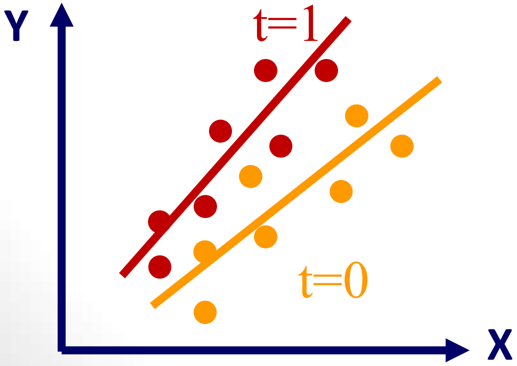
Assumes that the relation between *Y* and *X* is the same in both periods *t*=0 and 1.



Different intercepts and constant slope coefficients

$$Y = \alpha + \beta X + \delta t + e$$

Assume that *Y* varies in time but the relation between *Y* and *X* remains constant.



Different intercepts and slope coefficients

$$Y = \alpha + \beta X + \delta t + \theta(t \times X) + e$$

Both the intercept and the marginal impact of *X* on *Y* change over time.

Pooled Data - Definition

- Pooled data presents some main advantages when comparted to cross-sectional data: i) larger sample size; ii) allows us to identify changes in the relation over time;
- If we assume that the relation is the same over time:

$$Y = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + e_i$$

- If we assume that the expected value of *Y* varies over time and the relation between *Y* and *X* remains constant:

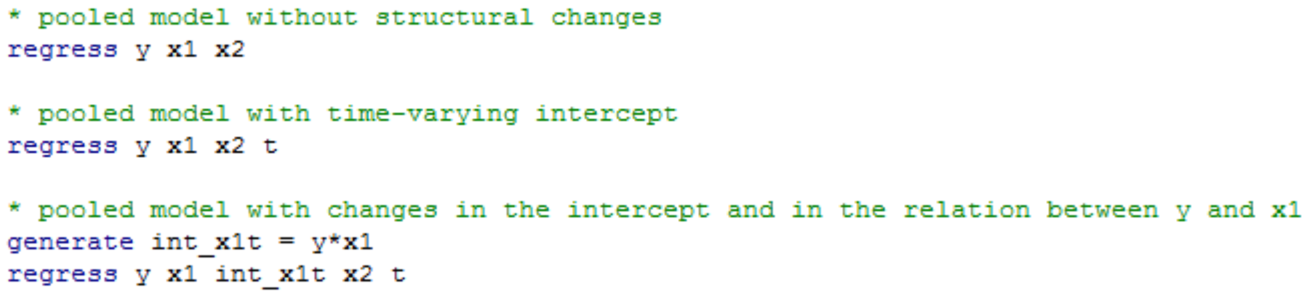
$$Y = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + \delta t + e_i$$

- If we assume changes in both the expected value of *Y* and in the relation between *Y* and *X* over time:

$$Y = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + \delta t + \sum_{j=1}^{k} \theta_j X_j \times t + e_i$$

Example – Stata & R

- Suppose we have a pooled data with information for the regressand *y* and two exogenous variables (*x*1 and *x*2) across two periods (t=0 and 1):

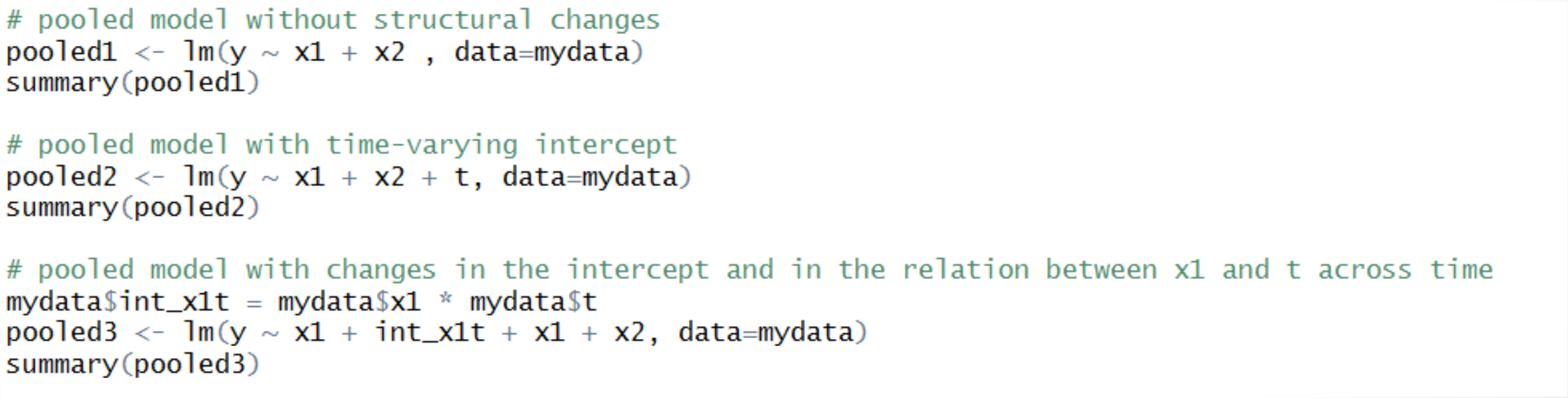


```
* pooled model without structural changes
regress y x1 x2

* pooled model with time-varying intercept
regress y x1 x2 t

* pooled model with changes in the intercept and in the relation between y and x1
generate int_x1t = y*x1
regress y x1 int_x1t x2 t
```

- The equivalent in R:



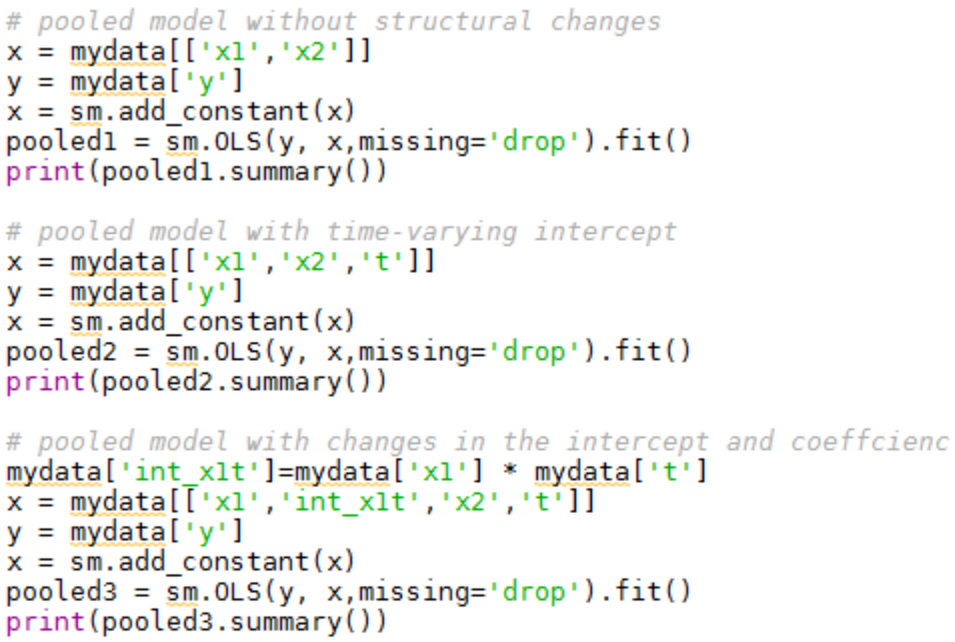
```
# pooled model without structural changes
pooled1 <- lm(y ~ x1 + x2 , data=mydata)
summary(pooled1)

# pooled model with time-varying intercept
pooled2 <- lm(y ~ x1 + x2 + t, data=mydata)
summary(pooled2)

# pooled model with changes in the intercept and in the relation between x1 and t across time
mydata$int_x1t = mydata$x1 * mydata$t
pooled3 <- lm(y ~ x1 + int_x1t + x1 + x2, data=mydata)
summary(pooled3)
```

Example – Python

- The equivalent in Python:



```
# pooled model without structural changes
x = mydata[['x1','x2']]
y = mydata['y']
x = sm.add_constant(x)
pooled1 = sm.OLS(y, x,missing='drop').fit()
print(pooled1.summary())

# pooled model with time-varying intercept
x = mydata[['x1','x2','t']]
y = mydata['y']
x = sm.add_constant(x)
pooled2 = sm.OLS(y, x,missing='drop').fit()
print(pooled2.summary())

# pooled model with changes in the intercept and coeffcienc
mydata['int_x1t']=mydata['x1'] * mydata['t']
x = mydata[['x1','int_x1t','x2','t']]
y = mydata['y']
x = sm.add_constant(x)
pooled3 = sm.OLS(y, x,missing='drop').fit()
print(pooled3.summary())
```

Exercise

1) The dataset *Data_AgricultureClimate.csv* contains information on agricultural production and climate change in São Paulo, Brazil (GORI MAIA, A., MIYAMOTO, B. C, GARCIA, J. R. Climate change and agriculture: Do environmental preservation and ecossystem services matter? Ecoloogical Economics, v. 152 (October 2018), 2018):

   a) Develop a regression model for pooled data to analyze the relation between the (log of) production value, (log of) area, temperature and precipitation;

   b) Consider changes in the relation before and after 2005 (variable *periodo*);

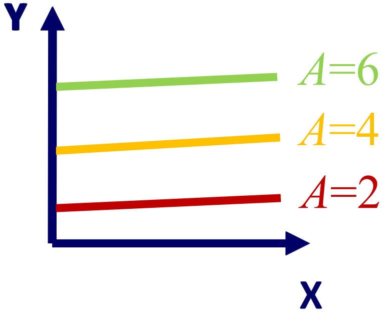
Omitted Variable Bias

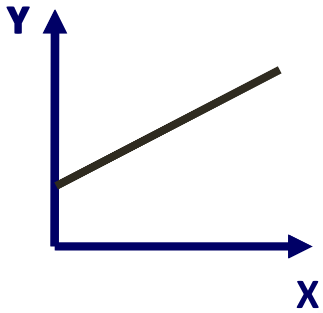
- Suppose that the production *Y* depends on the credit (*X*) and the land size *A*;
- If we can not observe the value of land size *A*, the simple relation between production *Y* and credit *X* tends to be biased;

A=2	*A*=2	*A*=4	*A*=4	*A*=6	*A*=6
Y=2000	*Y*=2200	*Y*=4000	*Y*=4000	*Y*=6200	*Y*=6000
X=2	*X*=4	*X*=6	*X*=8	*X*=10	*X*=12

$$Y = \alpha + \beta_1 X + \beta_2 A + e$$

$$Y = \alpha + \beta X + e$$





Controlling for Unobersvables

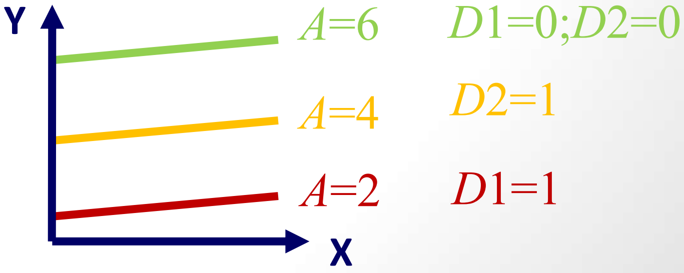
- Suppose that each farm (*i*=1,2,3) is observed in two distinct periods (t=0,1);
- If we assume that the land size *A* is different between the farms but constant over time, we can control the effect of land size on *Y* by using binary variables to identify each farm (for example, *D*1=1 para *i*=1, *D*2=1 para *i*=2, farm 3 is the reference);
- In other words, although land size A is non-observable, we can control its effect on *Y* by including a component *c*, in our model, called *unobserved heterogeneity*.

D1=1; *D2*=0		*D1*=0; *D2*=1		*D1*=0; *D2*=0	
A=2	*A*=2	*A*=4	*A*=4	*A*=6	*A*=6
i=1	*i*=1	*i*=2	*i*=2	*i*=3	*i*=3
t=0	*t*=1	*t*=0	*t*=1	*t*=0	*t*=1
Y=2000	*Y*=2200	*Y*=4000	*Y*=4000	*Y*=6200	*Y*=6000
X=2	*X*=4	*X*=6	*X*=8	*X*=10	*X*=12

$$Y_{it} = \alpha + \beta_1 X_{it} + \beta_2 A_{it} + e_{it}$$

$$Y_{it} = \alpha + \beta_1 X_{it} + c_1 D1_i + c_2 D2_i + e_{it}$$

$$Y_{it} = \alpha + \beta_1 X_{it} + c_i + e_{it}$$



Unobserved Heterogeneity

- Assume that the relation between *y* and **x** ≡ (X_1, X_2, ..., X_k) is given by:

$$E(y \mid \mathbf{x}, c) = \mathbf{x}\boldsymbol{\beta} + c$$

Where *c* is an unobserved component, also called **unobserved effect** or **unobserved heterogeneity**. One main assumption in the panel data analysis is that the component *c* is constant over time. This means:

$$Y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + e_{it}$$

Where $E(e_{it} | x_{it}, c_i) = 0$

- When *c* isn't correlated to the independent variables – Cov(X_j,*c*)=0 – then the omission of *c* in our model will not generate any kind of bias (omitted variable bias). In this case, we could apply OLS using models for pooled data (*pooled regression*). However, if Cov(X_j,*c*)≠0, the the pooled regression estimates are biased even for large samples.

Fixed Effects – Binary Variables

- Suppose the model with unobserved heterogeneity given by:

$$Y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + e_{it}$$

- The error e_{it} is called **idiosyncratic error**, since it varies randomly for all cross-sectional units and periods.
- A simple solution to control the unobserved heterogeneity c is given by the **fixed effects estimator with binary variables**. This method assumes that c_i represents a parameter that can be estimated using the coefficient associated with the *i*-th binary variable:

$$Y_{it} = \alpha + \sum_{j=1}^{k} \beta_j X_{j_{it}} + c_2 I_{2_i} + \ldots + c_n I_{n_i} + e_{it}$$

Where I_{ji}=1 if $j=i$, I_{ji}=0 if $j \neq i$. The estimators of de c_j are called **binary variables estimators**. The name “fixed effect” come from the idea that *c* is considered to be a parameter (constant value in the population).

Within Transformation

- One main limitation of the fixed effects estimator with binary variable is that the number of binary variables may be quite large. Most estimates tend to be insignificant if the sample is not large enough to compensate the lost degrees of freedoms.
- Alternatively, through an algebraic transformation, we can estimate the same coefficients using the **within estimators.**

Suppose the model with unobserved heterogeneity:

$$Y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + e_{it}$$

This relation is also valid for the average values of each cross-sectional unit:

$$\bar{Y}_i = \bar{\mathbf{x}}_i\boldsymbol{\beta} + c_i + \bar{e}_i$$

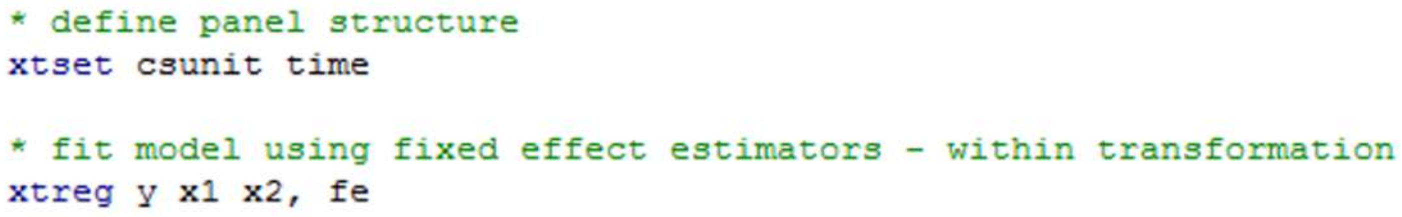
Since c_i is constant over time, its average is the same than c_i.

Subtracting the equations, we have:

$$\underbrace{(Y_{it} - \bar{Y}_i)}_{\tilde{Y}_{ij}} = \underbrace{(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)}_{\tilde{\mathbf{x}}_{ij}}\boldsymbol{\beta} + (c_i - c_i) + \underbrace{(e_{it} - \bar{e}_i)}_{\tilde{e}_{ij}} \Longrightarrow \tilde{Y}_{it} = \tilde{\mathbf{x}}_{it}\boldsymbol{\beta} + \tilde{e}_{it}$$

Example – Stata & R

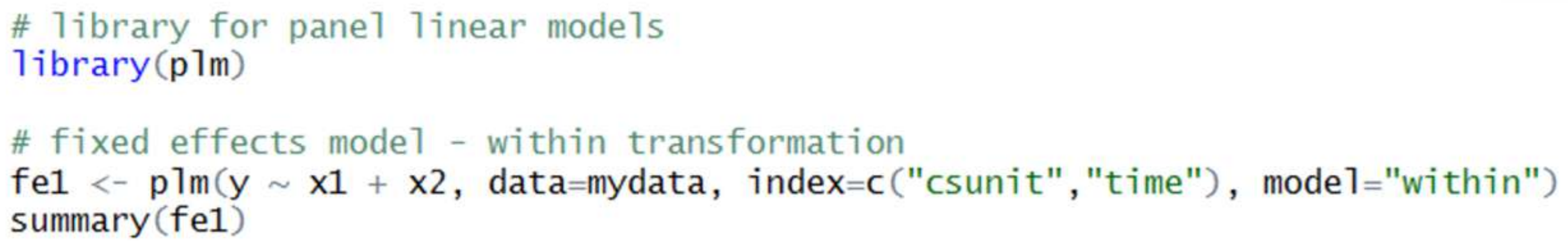
- Suppose we have a panel with information for the regressand *y* and two exogenous variables (*x*1 and *x*2) across *n* cross-sectional units (variable *cs*=1..*n*) and *T* periods (variable *time*=1..*T*). The within estimator is given in Stata by:



```
* define panel structure
xtset csunit time

* fit model using fixed effect estimators - within transformation
xtreg y x1 x2, fe
```

- The equivalent in R:

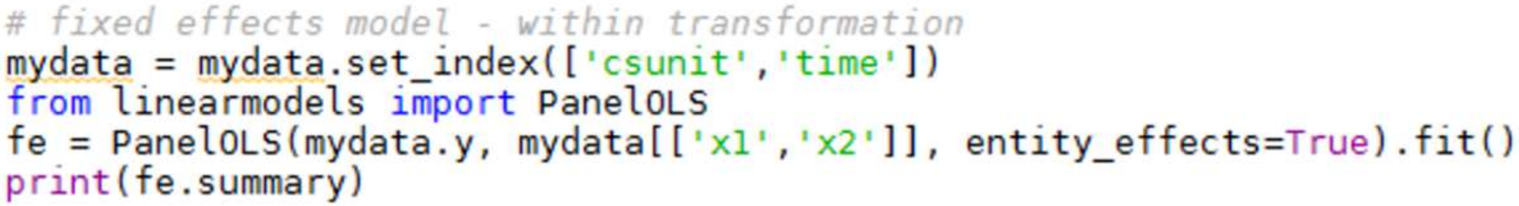


```
# library for panel linear models
library(plm)

# fixed effects model - within transformation
fe1 <- plm(y ~ x1 + x2, data=mydata, index=c("csunit","time"), model="within")
summary(fe1)
```

Example – Stata & R

- The equivalent in Python



```
# fixed effects model - within transformation
mydata = mydata.set_index(['csunit','time'])
from linearmodels import PanelOLS
fe = PanelOLS(mydata.y, mydata[['x1','x2']], entity_effects=True).fit()
print(fe.summary)
```

Two-Way Fixed Effects Estimator

- The model with controls for the heterogeneity across cross-sectional units (c_i) is also called one-way model:

$$Y_{it} = \alpha_t + \sum_{j=1}^{k} \beta_j X_{j_{it}} + c_i + e_{it}$$

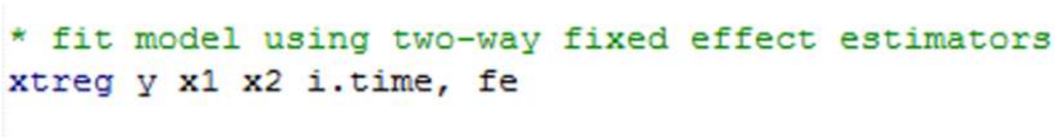
- We can extend this idea, using binary variables to control for the heterogeneity across periods *t*. The two-way model is:

$$Y_{it} = \alpha + \sum_{j=1}^{k} \beta_j X_{j_{it}} + c_i + ct_2 P_{2_t} + \ldots + ct_T P_{T_t} + e_{it}$$

Where P_{ji}=1 if j=t, P_{ji}=0 if $j \neq t$.

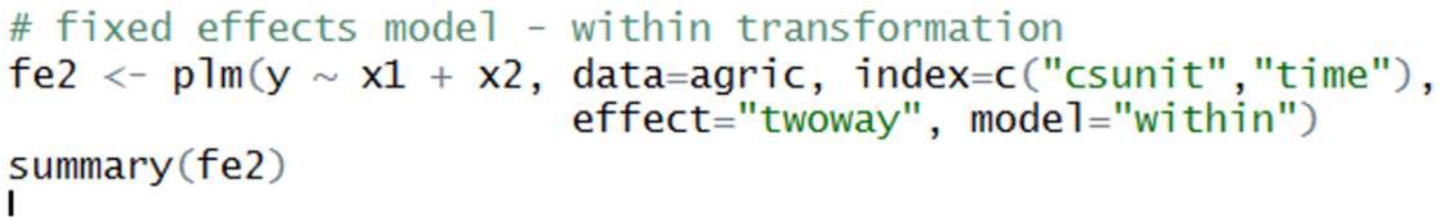
Example – Stata, R & Python

- The two-way estimator in Stata:



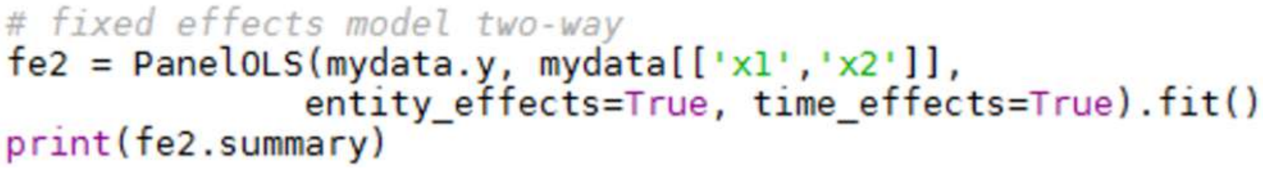
```
* fit model using two-way fixed effect estimators
xtreg y x1 x2 i.time, fe
```

- The equivalent in R:



```
# fixed effects model - within transformation
fe2 <- plm(y ~ x1 + x2, data=agric, index=c("csunit","time"),
                        effect="twoway", model="within")
summary(fe2)
```

- The equivalent in Python:



```
# fixed effects model two-way
fe2 = PanelOLS(mydata.y, mydata[['x1','x2']],
               entity_effects=True, time_effects=True).fit()
print(fe2.summary)
```

Advantages of Panel Data Analysis

1) Differences across individuals and periods: Panel data models allow us to use binaries to control the differences across cross-sectional units (individuals) and periods. Cross-sectional data does not provide enough degrees of freedom for such analysis;

2) Degrees of freedom: the sample size of a panel data is the number of cross-sectional units multiplied by number of periods. In a cross-sectional (time series) data we only have the number of cross-sectional units (periods);

3) Controlling for omitted variable bias: we can control for unobservables that are related to both the regressors and the regressand (omitted variable bias) using binary variables or the within transformation;

Exercise

1) The dataset *Data_AgricultureClimate.csv* contains information on agricultural production and climate variables in the state of São Paulo (GORI MAIA, A., MIYAMOTO, B. C, GARCIA, J. R. Climate change and agriculture: Do environmental preservation and ecossystem services matter? Ecoloogical Economics, v. 152 (October 2018), 2018):
   a) Analyze the relation between the (log) value of agricultural production, (log) area, temperature and precipitation using the one-way fixed-effects estimators;
   b) Now use two-way fixed-effects estimators, identifying the main differences in relation to (a);