Econometric Softwares: Stata

Jérémy Do Nascimento Miguel

Class 7 Spring 2024

Session 7: Outline

- Preparing panel data
- Visualizing panel data
- Panel data regression methods

How it should look like

Entity	Year	X	Z
1	2010	10	200
1	2011	3	300
1	2012	12	400
2	2010	9	100
2	2011	2	200
2	2012	3	300
3	2010	6	400
3	2011	8	500
3	2012	14	200

- Panel data (also known as longitudinal data) is a dataset in which cross-sectional units/entities (i) are observed across time (t):
  $\rightarrow$ $(X_{it}, Y_{it}])$ with $i = 1, ...n; t = 1, ...T$
- Units can be countries, districts, firms, households, etc.

Preparing Panel data

To analyze panel data:

- Variables should be in columns
- Entity (i) and time (t) should be in rows

This format is also known as long format

Entity	Year	X	Z
1	2010	10	200
1	2011	3	300
1	2012	12	400
2	2010	9	100
2	2011	2	200
2	2012	3	300
3	2010	6	400
3	2011	8	500
3	2012	14	200

Preparing Panel data: wide form data

Sometimes data may come in wide format: time or entities are in columns, not in rows: need to **reshape** the data to long format

- Case 1: Time in columns

Entity	2010	2011	2012
A	10	3	12
B	9	2	3
C	6	8	14

- Case 2: Units in columns

Year	A	B	C
2010	10	9	6
2011	3	2	8
2012	12	3	14

Preparing Panel data: wide form data

- Case 1: Time in columns

Entity	2010	2011	2012
A	10	3	12
B	9	2	3
C	6	8	14

(!) Beware that Stata does not accept numbers as column names: Add a letter to the numbers before importing (Excel file) into Stata

Preparing Panel data: wide form data

Case 1: Time (t) in columns:

1. Import into Stata (using import excel or import delimited for Excel or CSV files): import excel using panel.xlsx, sheet(wide1) clear first
2. Use reshape longcommand to transform data from wide to long format: Beware that if entity identifiers are string characters, you need to generate numeric ID variable that uniquely identifies each entity before using reshape command.

   ```
   bys Entity: gen ID = _n
   reshape long X_, i(ID) j(year)
   rename X_ X
   ```

Preparing Panel data: wide form data

- Final result should look like

ID	year	Entity	X
1	2010	A	10
1	2011	A	3
1	2012	A	12
2	2010	B	9
2	2011	B	2
2	2012	B	3
3	2010	C	6
3	2011	C	8
3	2012	C	14

Preparing Panel data: wide form data

- Case 2: Units (i) in columns

Year	A	B	C
2010	10	9	6
2011	3	2	8
2012	12	3	14

Preparing Panel data: wide form data

Case 2: (t) in columns:

1. Import into Stata (using import excel or import delimited for Excel or CSV files): import excel using panel.xlsx, sheet(wide2) clear first
2. Use reshape longcommand to transform data from wide to long format: Beware that if column names are solely entity identifiers/names, you need to add prefix (which can be name of variable, e.g. X) to each column name, before using reshape command.

```
rename A X_A

rename B X_B

rename C X_C

reshape long X_, i(Year) j(Entity) string
rename X_ X
```

Option string added here in reshape to specify that entity identifier is string (i.e. A, B, C)

Preparing Panel data: wide form data

- Final result should look like

Year	Entity	X
2010	A	10
2010	B	9
2010	C	6
2011	A	3
2011	B	2
2011	C	8
2012	A	12
2012	B	3

Preparing panel data: encoding

In Case 2 above, we see that in the final result, entity identifier is a string variable (variable Entity)

- It is always better to have numeric identifiers: it increases speed of running commands (and some specific commands may require identifiers to be numeric)
- Two ways of assigning numeric values to string identifiers (i.e. A=1, B=2, C=3):
  1. Generate a new variable that gives unique numeric value to each unique value of string variable, using egen ... = group() command: egen ID = group(Entity)
  2. Generate a new variable that assigns a numeric value to each unique value of string variable AND is labeled with those string values: encode Entity, gen(ID)

Preparing Panel data: encoding

1. Using `egen`:

Year	Entity	X	ID
2010	A	10	1
2010	B	9	2
2010	C	6	3
2011	A	3	1
2011	B	2	2
2011	C	8	3
2012	A	12	1
2012	B	3	2
2012	C	14	3

2. Using `encode`:

Year	Entity	X	ID
2010	A	10	A
2010	B	9	B
2010	C	6	C
2011	A	3	A
2011	B	2	B
2011	C	8	C
2012	A	12	A
2012	B	3	B
2012	C	14	C

Declaring panel structure

- After preparing panel data (i.e. entity and time in rows, variables in columns), we can declare to Stata that it is panel data: xtset ID year
- After panel data is set, we can use a set of xt commands to analyze the data. See the full list of commands at help xt

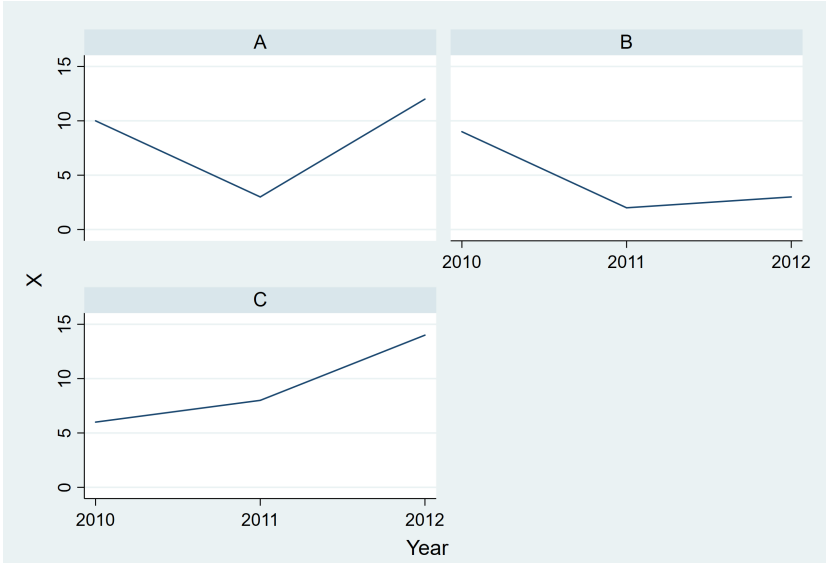
Visualizing Panel data

Note that almost all of commands we used to generate graphs also work with panel data.

- xtline works only with panel data (i.e. after declaring panel structure with xtreg)
- This command generates time-trend line of a variable separately for each entity

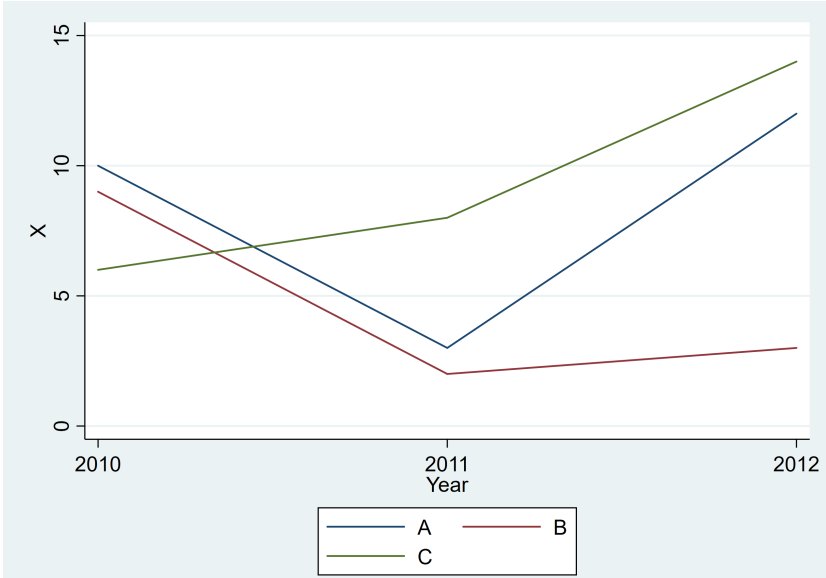
Visualizing Panel data

xtline X, xlab(2010(1)2012)



Visualizing Panel data

xtline X, xlab(2010(1)2012) overlay



Panel data regression methods: Fixed Effects model

In Stata, there are different ways to estimate fixed effects (FE) model:

1. regress with entity/time dummies or with i. operators:
   regress Y X i.ID (with entity FE)
   regress Y X i.ID i.Year (with entity and year FE)
2. xtreg with/without time dummies or with i. operators (works only after setting panel data with xtset):
   xtreg Y X, fe (with entity FE)
   xtreg Y X i.year, fe (with entity and year FE)
3. reghdfe with FE specified in absorb() option of the command:
   reghdfe Y X, absorb(Entity) (with entity FE)
   reghdfe Y X , absorb(Entity Year) (with entity and year FE)

Panel data regression methods: Random Effects model

In Stata, we can estimate random effects model using xtreg, recommand: xtreg y x , re

- In random effects model, we can include time-invariant variables and estimate their effect on dependent variable
- Keep in mind that some of these time-invariant variables may not be available/observed and this may potentially lead to omitted variable bias (if these variables are correlated with other regressors in the model
- In fixed effects model, time-invariant characteristics and all unobserved differences across entities will be captured with entity fixed effects and their effect cannot be estimated

Exercise 7 I

We have data in Excel format **wdi_extract.xlsx** which is an extract from World Bank's World Development Indicators database. The format of the data is wide, i.e. columns are years, and also, variables and countries are in rows.

- Before importing the data, manually add suffix to column names in Excel (e.g. 2010 to y2010)
- Import data into Stata
- Reshape the data so that the final result will be years and countries are in rows, and variables are in columns.
- Make changes to identifiers if necessary and set the panel structure of the data (xtset)
- Use xtline to plot time trends of CO2 emissions in France, Germany, Netherlands, UK, Spain and Italy.

Exercise 7 II

- Estimate fixed effects model by regressing GDP per capita on CO2 emissions and fertility rate. First try with only country fixed effects and then add year fixed effects. Try estimation with xtreg and reghdfe.