Panel Data Analysis
Fixed and Random Effects using Stata

(v. 6.0)

Oscar Torres-Reyna

otorres@princeton.edu

December 2007

http://www.princeton.edu/~otorres/

What panel data looks like…

Panel data (also known as longitudinal or cross-sectional time-series data) is a dataset in which the behavior of entities (*i*) are observed across time (*t*).

$(X_{it}, Y_{it}), i=1,\dots n; t=1,\dots T$

These entities could be states, companies, families, individuals, countries, etc.

Entity	Year	Y	X1	X2	X3	…..
1	1	#	#	#	#	…..
1	2	#	#	#	#	…..
1	3	#	#	#	#	…..
:	:	:	:	:	:	:
2	1	#	#	#	#	…..
2	2	#	#	#	#	…..
2	3	#	#	#	#	…..
:	:	:	:	:	:	:
3	1	#	#	#	#	…..
3	2	#	#	#	#	…..
3	3	#	#	#	#	…..

See Stock and Watson, *Introduction to Econometrics*, chapter 10 “Regression with Panel Data”.

Preparing Data into Panel Data format

The data: the long form

To analyze panel data:

- Variables should be in columns.
- Entity and time in rows.

This format is known as long form.

Entity	Year	Y	X1	X2	X3	.....
1	1	#	#	#	#	.....
1	2	#	#	#	#	.....
1	3	#	#	#	#	.....
:	:	:	:	:	:	:
2	1	#	#	#	#	.....
2	2	#	#	#	#	.....
2	3	#	#	#	#	.....
:	:	:	:	:	:	:
3	1	#	#	#	#	.....
3	2	#	#	#	#	.....
3	3	#	#	#	#	.....

Wide form data (time in columns)

If your dataset is in wide format, either entity or time are in columns, you need to reshape it to long format (you can do this in Stata).

Beware that Stata does not like numbers as column names. You need to add a letter to the numbers before importing into Stata. If you have something like the following:

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Country**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**
2	A			8000.01	8212.90	7847.36	7702.89	7288.48	6430.98	6932.45	7486.24	8094.17
3	B	18268.01	18738.99	19360.46	20151.42	20715.54	20866.90	21364.02	21801.41	22404.59	22676.26	23039.43
4	C	21088.14	21608.14	21988.64	22739.28	23436.61	24194.85	24300.57	24411.48	24650.02	25076.01	25346.01
5	D	313.74	321.36	331.76	342.12	351.70	365.33	377.15	386.26	398.86	415.96	432.63
6	E	21123.66	21659.55	22299.13	22972.31	23613.87	24150.86	24788.69	25368.87	25885.48	26582.19	26890.73
7	F	29941.64	30703.73	31716.04	32671.27	33748.21	34599.47	34483.98	34669.47	35312.75	36450.55	37267.33
8	G	4891.60	5063.81	5328.88	5512.59	5647.06	5934.98	5864.12	5852.99	5872.29	6055.92	6162.84

Wide form data (time in columns)

Add a letter to the numeric column names, for example, an ‘x’ before the year:

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Country**	**x1995**	**x1996**	**x1997**	**x1998**	**x1999**	**x2000**	**x2001**	**x2002**	**x2003**	**x2004**	**x2005**
2	A			8000.01	8212.90	7847.36	7702.89	7288.48	6430.98	6932.45	7486.24	8094.17
3	B	18268.01	18738.99	19360.46	20151.42	20715.54	20866.90	21364.02	21801.41	22404.59	22676.26	23039.43
4	C	21088.14	21608.14	21988.64	22739.28	23436.61	24194.85	24300.57	24411.48	24650.02	25076.01	25346.01
5	D	313.74	321.36	331.76	342.12	351.70	365.33	377.15	386.26	398.86	415.96	432.63
6	E	21123.66	21659.55	22299.13	22972.31	23613.87	24150.86	24788.69	25368.87	25885.48	26582.19	26890.73
7	F	29941.64	30703.73	31716.04	32671.27	33748.21	34599.47	34483.98	34669.47	35312.75	36450.55	37267.33
8	G	4891.60	5063.81	5328.88	5512.59	5647.06	5934.98	5864.12	5852.99	5872.29	6055.92	6162.84

Import into Stata

	Country	x1995	x1996	x1997	x1998	x1999	x2000	x2001	x2002	x2003	x2004	x2005
1	A	.	.	8000.01	8212.90	7847.36	7702.89	7288.48	6430.98	6932.45	7486.24	8094.17
2	B	18268.01	18738.99	19360.46	20151.42	20715.54	20866.90	21364.02	21801.41	22404.59	22676.26	23039.43
3	C	21088.14	21608.14	21988.64	22739.28	23436.61	24194.85	24300.57	24411.48	24650.02	25076.01	25346.01
4	D	313.74	321.36	331.76	342.12	351.70	365.33	377.15	386.26	398.86	415.96	432.63
5	E	21123.66	21659.55	22299.13	22972.31	23613.87	24150.86	24788.69	25368.87	25885.48	26582.19	26890.73
6	F	29941.64	30703.73	31716.04	32671.27	33748.21	34599.47	34483.98	34669.47	35312.75	36450.55	37267.33
7	G	4891.60	5063.81	5328.88	5512.59	5647.06	5934.98	5864.12	5852.99	5872.29	6055.92	6162.84

Reshaping from wide to long

Once in Stata, you can reshape it using the command `reshape`:

```
gen id = _n

order id

reshape long x , i(id) j(year)

rename x gdp
```

Type `help reshape` for more details

	id	year	Country	gdp
1	1	1995	A	.
2	1	1996	A	.
3	1	1997	A	8000.01
4	1	1998	A	8212.90
5	1	1999	A	7847.36
6	1	2000	A	7702.89
7	1	2001	A	7288.48
8	1	2002	A	6430.98
9	1	2003	A	6932.45
10	1	2004	A	7486.24
11	1	2005	A	8094.17
12	2	1995	B	18268.01
13	2	1996	B	18738.99
14	2	1997	B	19360.46
15	2	1998	B	20151.42
16	2	1999	B	20715.54
17	2	2000	B	20866.90
18	2	2001	B	21364.02
19	2	2002	B	21801.41
20	2	2003	B	22404.59
21	2	2004	B	22676.26
22	2	2005	B	23039.43
23	3	1995	C	21088.14
24	3	1996	C	21608.14
25	3	1997	C	21988.64
26	3	1998	C	22739.28
27	3	1999	C	23436.61
28	3	2000	C	24194.85
29	3	2001	C	24300.57
30	3	2002	C	24411.48
31	3	2003	C	24650.02
32	3	2004	C	25076.01
33	3	2005	C	25346.01

Wide form data (entity in columns)

If the wide format data has the entities in column and time in rows, like this example:

	A	B	C	D	E	F	G	H
1	**Year**	**A**	**B**	**C**	**D**	**E**	**F**	**G**
2	**1995**		18268.01	21088.14	313.74	21123.66	29941.64	4891.60
3	**1996**		18738.99	21608.14	321.36	21659.55	30703.73	5063.81
4	**1997**	8000.01	19360.46	21988.64	331.76	22299.13	31716.04	5328.88
5	**1998**	8212.90	20151.42	22739.28	342.12	22972.31	32671.27	5512.59
6	**1999**	7847.36	20715.54	23436.61	351.70	23613.87	33748.21	5647.06
7	**2000**	7702.89	20866.90	24194.85	365.33	24150.86	34599.47	5934.98
8	**2001**	7288.48	21364.02	24300.57	377.15	24788.69	34483.98	5864.12
9	**2002**	6430.98	21801.41	24411.48	386.26	25368.87	34669.47	5852.99
10	**2003**	6932.45	22404.59	24650.02	398.86	25885.48	35312.75	5872.29
11	**2004**	7486.24	22676.26	25076.01	415.96	26582.19	36450.55	6055.92
12	**2005**	8094.17	23039.43	25346.01	432.63	26890.73	37267.33	6162.84

Wide form data (entity in columns)

Import it into Stata:

	Year	A	B	C	D	E	F	G
1	1995	.	18268.01	21088.14	313.74	21123.66	29941.64	4891.60
2	1996	.	18738.99	21608.14	321.36	21659.55	30703.73	5063.81
3	1997	8000.01	19360.46	21988.64	331.76	22299.13	31716.04	5328.88
4	1998	8212.90	20151.42	22739.28	342.12	22972.31	32671.27	5512.59
5	1999	7847.36	20715.54	23436.61	351.70	23613.87	33748.21	5647.06
6	2000	7702.89	20866.90	24194.85	365.33	24150.86	34599.47	5934.98
7	2001	7288.48	21364.02	24300.57	377.15	24788.69	34483.98	5864.12
8	2002	6430.98	21801.41	24411.48	386.26	25368.87	34669.47	5852.99
9	2003	6932.45	22404.59	24650.02	398.86	25885.48	35312.75	5872.29
10	2004	7486.24	22676.26	25076.01	415.96	26582.19	36450.55	6055.92
11	2005	8094.17	23039.43	25346.01	432.63	26890.73	37267.33	6162.84

Reshape wide to long format

Once in Stata, you can reshape it using the command `reshape`:

```
* Adding the prefix 'gdp' to column names. Command 'renvars' is user-written, you need to install it, see note below

renvars A-G, pref(gdp)

gen id = _n
order id
reshape long gdp , i(id) j(country) str
```

Type `help reshape` for more details.
You need to install `renvars`, type:
`search renvars`
Click on the link for `dm88_*` then install.

	id	country	Year	gdp
1	1	A	1995	.
2	1	B	1995	18268.01
3	1	C	1995	21088.14
4	1	D	1995	313.74
5	1	E	1995	21123.66
6	1	F	1995	29941.64
7	1	G	1995	4891.60
8	2	A	1996	.
9	2	B	1996	18738.99
10	2	C	1996	21608.14
11	2	D	1996	321.36
12	2	E	1996	21659.55
13	2	F	1996	30703.73
14	2	G	1996	5063.81
15	3	A	1997	8000.01
16	3	B	1997	19360.46
17	3	C	1997	21988.64
18	3	D	1997	331.76
19	3	E	1997	22299.13
20	3	F	1997	31716.04
21	3	G	1997	5328.88
22	4	A	1998	8212.90
23	4	B	1998	20151.42
24	4	C	1998	22739.28
25	4	D	1998	342.12
26	4	E	1998	22972.31
27	4	F	1998	32671.27
28	4	G	1998	5512.59
29	5	A	1999	7847.36
30	5	B	1999	20715.54
31	5	C	1999	23436.61
32	5	D	1999	351.70
33	5	E	1999	23613.87

More than one variable in same column

To reshape data from wide to long where more than one variable is in the same column like the example below, see slides 29 to 32 in this document:

https://www.princeton.edu/~otorres/DataPrep101.pdf#page=29

Entity	Year	Variable	Value
A	1	Var1	###
A	2	Var1	###
A	3	Var1	###
A	1	Var2	###
A	2	Var2	###
A	3	Var2	###

If you are downloading data from the World Development Indicators, see slide 21 in the link below to get it in the proper panel data form without the need to reshape: https://www.princeton.edu/~otorres/FindingData101.pdf#page=21

Assign numbers to strings

The `encode` command assigns a number to the string variable in alphabetical order.

The new variable is a labeled variable where the labels are the original strings assigned to specific number.

Notice that string variables have the color red, while labeled variables have color blue.

Type `help encode` for more info.

	id	year	Country	gdp	country1
1	1	1995	A	.	A
2	1	1996	A	.	A
3	1	1997	A	8000.01	A
4	1	1998	A	8212.90	A
5	1	1999	A	7847.36	A
6	1	2000	A	7702.89	A
7	1	2001	A	7288.48	A
8	1	2002	A	6430.98	A
9	1	2003	A	6932.45	A
10	1	2004	A	7486.24	A
11	1	2005	A	8094.17	A
12	2	1995	B	18268.01	B
13	2	1996	B	18738.99	B
14	2	1997	B	19360.46	B
15	2	1998	B	20151.42	B
16	2	1999	B	20715.54	B
17	2	2000	B	20866.90	B
18	2	2001	B	21364.02	B
19	2	2002	B	21801.41	B
20	2	2003	B	22404.59	B
21	2	2004	B	22676.26	B
22	2	2005	B	23039.43	B
23	3	1995	C	21088.14	C
24	3	1996	C	21608.14	C
25	3	1997	C	21988.64	C
26	3	1998	C	22739.28	C
27	3	1999	C	23436.61	C
28	3	2000	C	24194.85	C
29	3	2001	C	24300.57	C
30	3	2002	C	24411.48	C
31	3	2003	C	24650.02	C
32	3	2004	C	25076.01	C
33	3	2005	C	25346.01	C

Setting data as panel

Once the data is in long form, we need to set it as panel so we can use Stata's panel data **xt** commands and the time series operators. Using the example from the previous page type:

```
xtset country year
    string variables not allowed in varlist;
    Country is a string variable
```

Given the error, we need to have 'country' in numeric format. Type

```
encode country, gen(country1)
```

Then using 'country1' type

```
xtset country1 year
    Panel variable: country1 (strongly balanced)
     Time variable: year, 2000 to 2021
             Delta: 1 unit
```

Balanced panel: all entities are observed across all times. **Unbalanced** panel: some entities are not observed in some years. Stata algorithms automatically account for this.

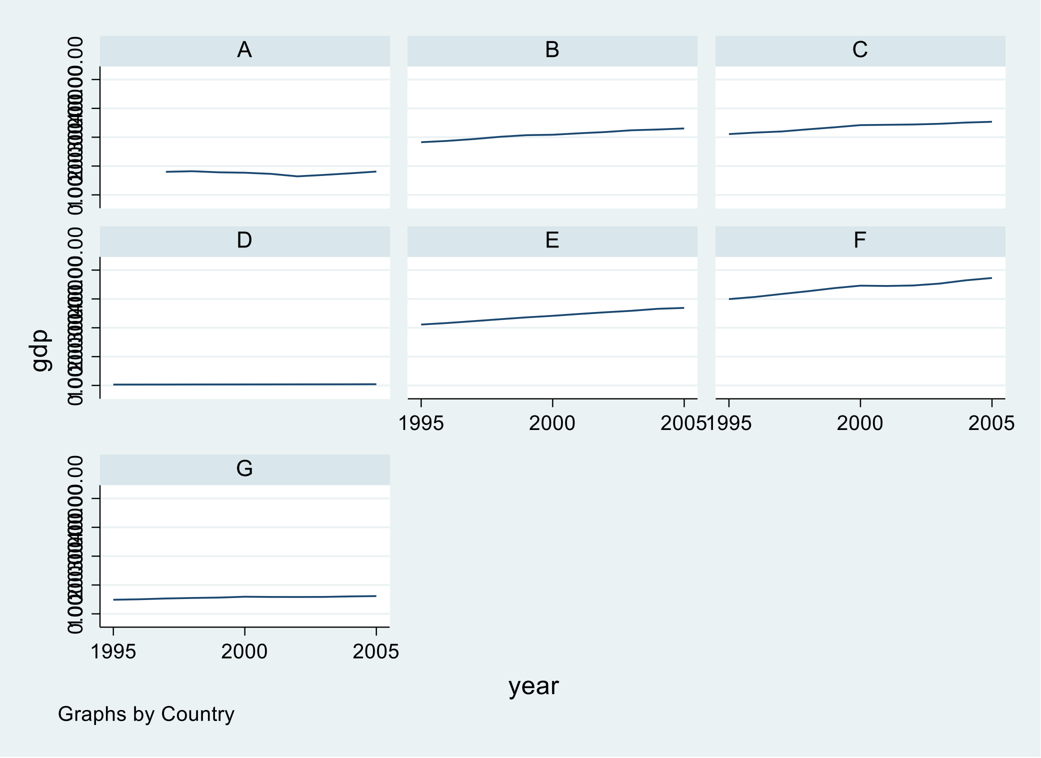
Visualizing panel data

Once the data is set as panel, you can use a series of *xt* commands to analyze it. For more information type:

```
help xt
```

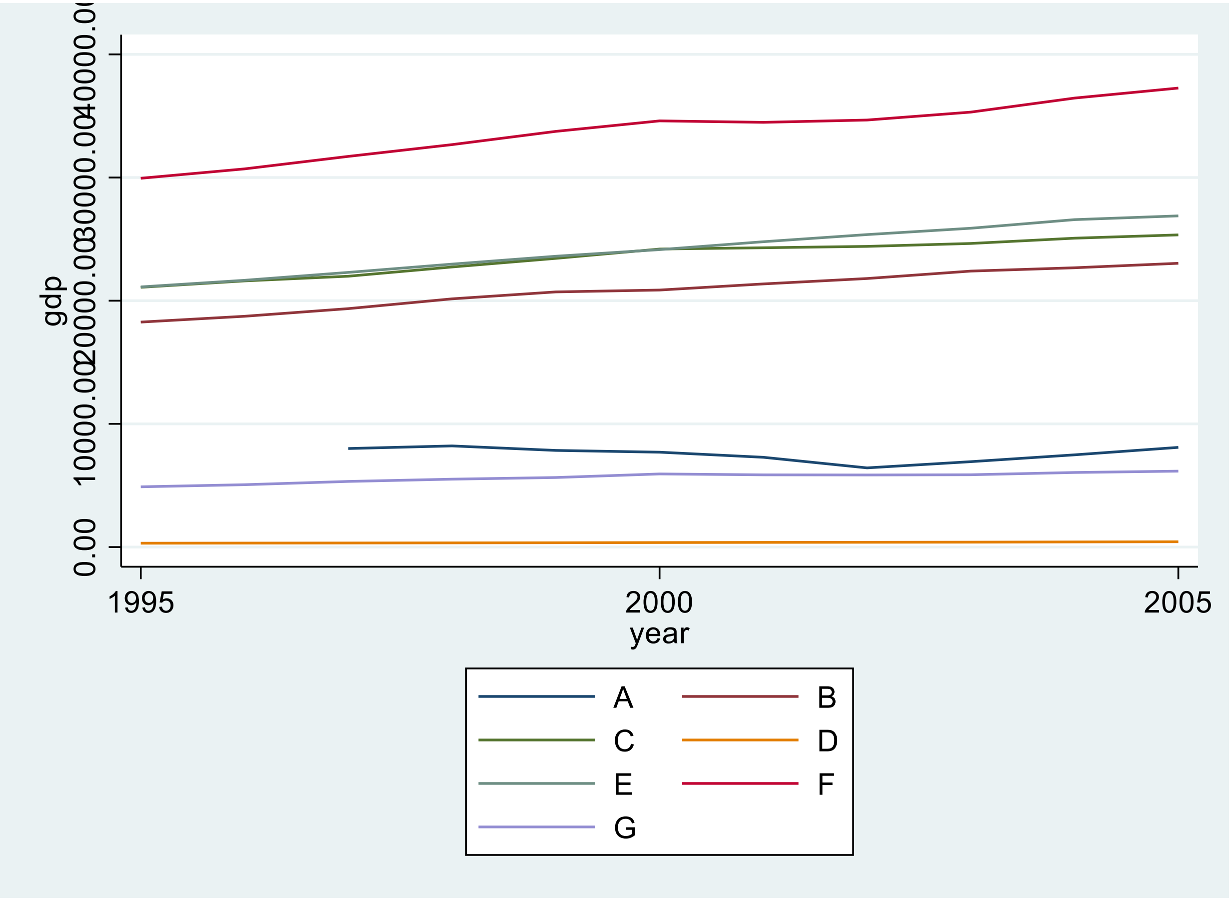
A useful visualization command is `xtline`, type:

```
xtline gdp
```



Visualizing panel data

```
* All in one, type:
xtline gdp, overlay
```



Usage

Panel data deals with omitted variable bias due to heterogeneity in the data. It does this by controlling for variables that we cannot observe, are not available, and/or can not be measured but are correlated with the predictors. Two types:

1. Variables that do not change over time but vary across entities (cultural factors, difference in business practices across companies, etc.) → Entity fixed effects.
2. Variables that change over time but not across entities (i.e. national policies, federal regulations, international agreements, etc.) → Time fixed effects.

Some drawbacks when working with panel data are data collection issues (i.e. sampling design, coverage), non-response in the case of micro panels or cross-country dependency in the case of macro panels (i.e. correlation between countries).

For a comprehensive list of advantages and disadvantages of panel data see Baltagi, *Econometric Analysis of Panel Data* (chapter 1).

FIXED-EFFECTS MODEL

(Covariance Model, Within Estimator, Individual Dummy Variable Model, Least Squares Dummy Variable Model)

The fixed effects idea

Entities have individual 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. Business practices, cultural, or political variables are, most of the time unavailable or hard to measure.

The fixed effects idea

Since individual characteristics are not random and may impact the predictor or outcome variables, we need to control for them. In this way, the effect of the predictors will not be influenced by those fixed characteristics.*

In entity's fixed effects it is assumed a correlation between the entity's error term and predictor variables. However, an entity's fixed effects cannot be correlated with another entity's.

* See Stock and Watson, 2003, p.289-290

The model (1)

The entity fixed effects regression model is

$$Y_{it} = \alpha_i + \beta X_{it} + u_i + e_{it}$$

i = 1…n ; t = 1….T

Where:

Y_{it} outcome variable (for entity *i* at time *t*).
α_i is the unknown intercept for each entity (*n* entity-specific intercepts).
X_{it} is a vector of predictors (for entity *i* at time *t*) .
u_i within-entity error term ; e_{it} overall error term.

Interpretation of the β coefficient: for a given entity, when a predictor changes one unit over time, the outcome will increase/decrease by β units (assuming no transformation is applied).* Here, β represents a common effect across entities controlling for individual heterogeneity.

* See Bartels, Brandom, "Beyond "Fixed Versus Random Effects": A framework for improving substantive and statistical analysis of panel, time-series cross-sectional, and multilevel data", Stony Brook University, working paper, 2008

The model (2)

The entity and time fixed effects regression model is

$$Y_{it} = \alpha_i + \beta X_{it} + \delta_t + u_i + e_{it}$$

i = 1…n ; t = 1….T

Where:

Y_{it} outcome variable (for entity *i* at time *t*).

α_i is the unknown intercept for each entity (*n* entity-specific intercepts).

X_{it} is a vector of predictors (for entity *i* at time *t*) .

δ_t is the unknow coefficient for the time regressors (*t*)

u_i within-entity error term ; e_{it} overall error term.

Interpretation of a β coefficient: for a given entity, when a predictor changes one unit over time, the outcome will increase/decrease by β units (assuming no transformation is applied).* Here, β represents a common effect across entities controlling for individual and time heterogeneity.

* See Bartels, Brandom, "Beyond "Fixed Versus Random Effects": A framework for improving substantive and statistical analysis of panel, time-series cross-sectional, and multilevel data", Stony Brook University, working paper, 2008

Data example

The data used in the following slides was extracted from the World Development Indicators database: https://databank.worldbank.org/source/world-development-indicators

Selected variables since 2000, all countries only:

- GDP per capita (constant 2015 US$)
- Exports of goods and services (constant 2015 US$)
- Imports of goods and services (constant 2015 US$)
- Labor force, total

Data was further cleaned to remove regions, subregions, and missing values across years and variables resulting in 126 countries.

Variable ‘`trade`’ was added by adding imports + exports.

Setting data as panel

Once the data is in long form, we need to set it as panel so we can use Stata's panel data **`xt`** commands and the time series operators. Using the example from the previous page type:

```
xtset country year
    string variables not allowed in varlist;
    Country is a string variable
```

Given the error, we need to have '`country`' in numeric format. Type

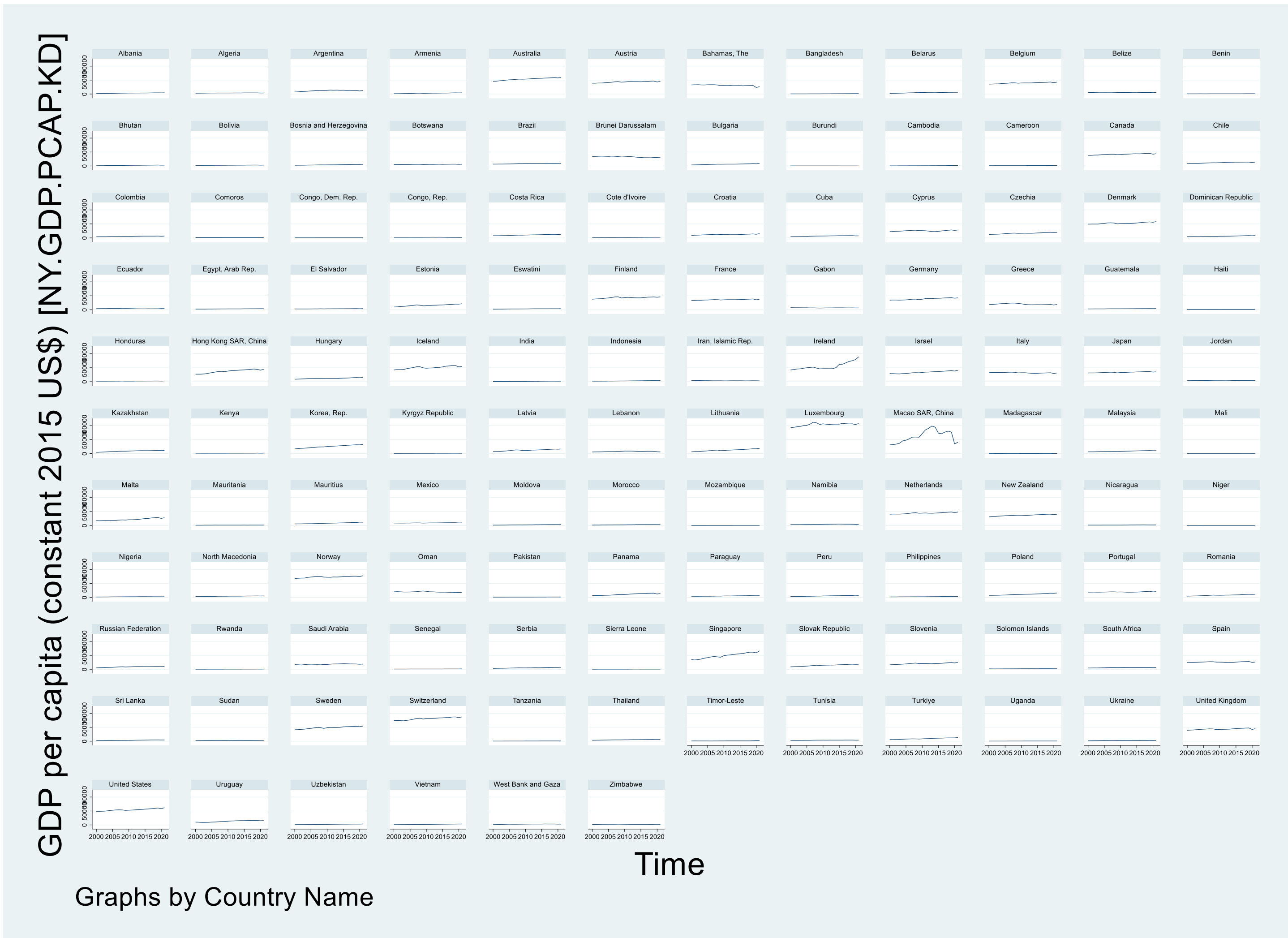
```
encode country, gen(country1)
```

Then using '`country1`' type

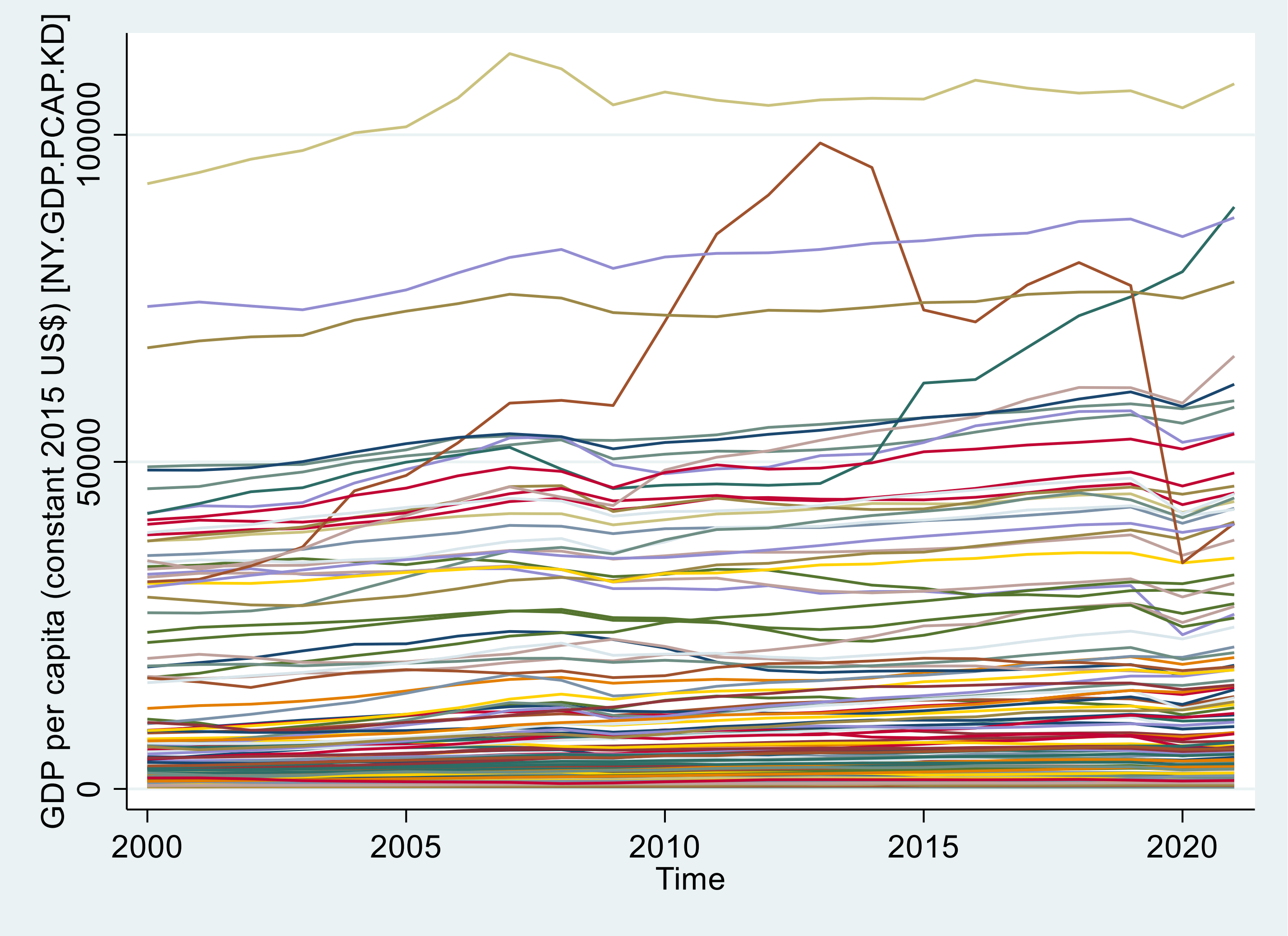
```
xtset country1 year
    Panel variable: country1 (strongly balanced)
     Time variable: year, 2000 to 2021
             Delta: 1 unit
```

Balanced panel: all entities are observed across all times. **Unbalanced** panel: some entities are not observed in some years. Stata algorithms automatically account for this.

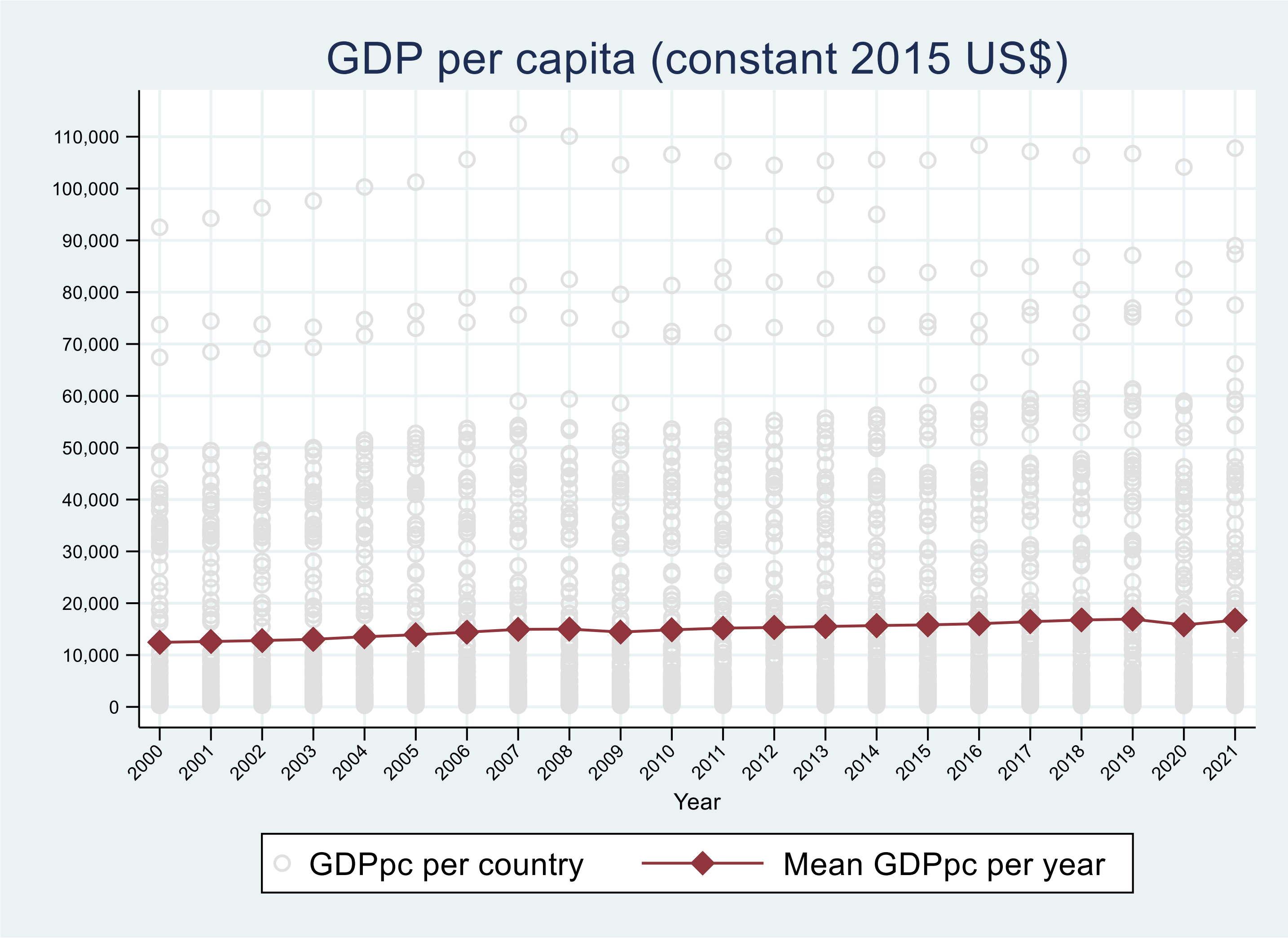
```
* Not ideal with many panels
xtline gdppc
```



```
* Getting the big picture
xtline gdppc, overlay legend(off)
```



```
* Heterogeneity across years
bysort year: egen mean_gdppc = mean(gdppc)
twoway scatter gdppc year, msymbol(circle_hollow) color(gs14) || ///
       connected mean_gdppc year, msymbol(diamond) sort ///
             ylabel(0(10000)115000, angle(0) labsize(2) format(%7.0fc)) ///
             xlabel(2000(1)2021, angle(45) labsize(2) grid) ///
             title(GDP per capita (constant 2015 US$)) xtitle(Year, size(2.5)) ///
             legend(label(1 "GDPpc per country") label(2 "Mean GDPpc per year"))
```



Data example – transformations

To log-transformed a variable use the function `ln()` :

```
gen ln_gdppc = ln(gdppc)
gen ln_labor = ln(labor)
gen ln_trade = ln(trade)
```

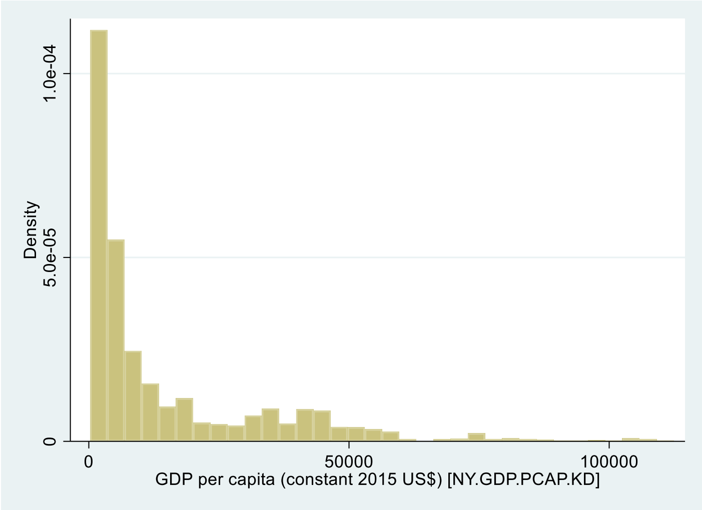
If the variable has negative values, you need to add a value high enough so the minimum value is over zero (preferable 1). For example, if the lowest value in ‘varX’ is -1, then type:

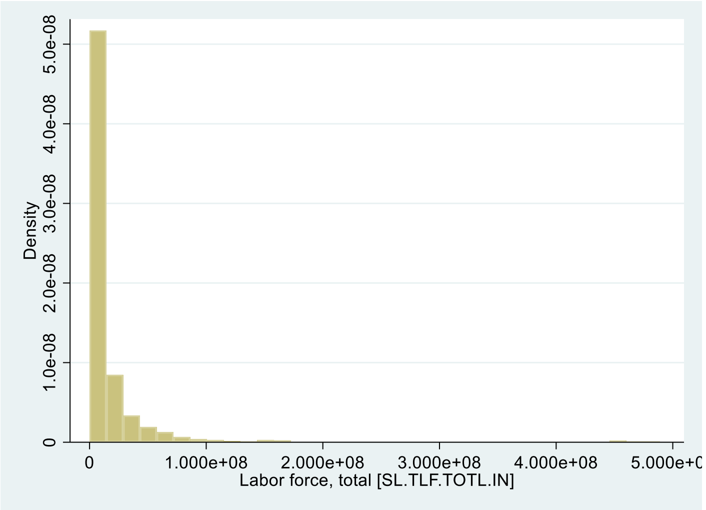
```
gen ln_varX = ln(varX + 2)
```

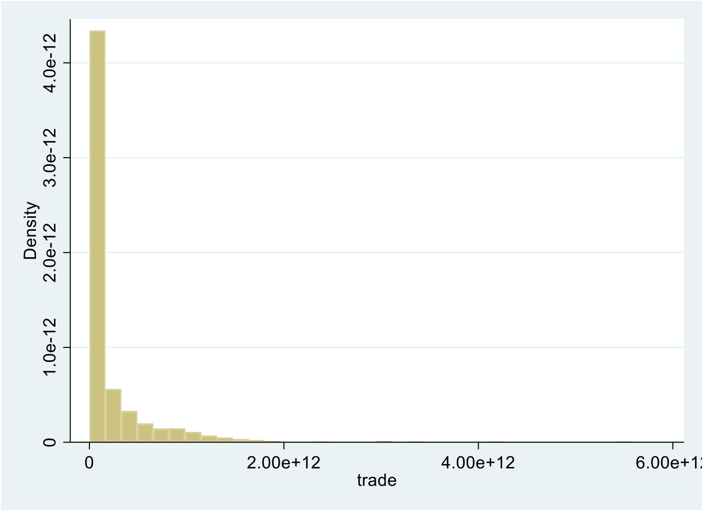
The natural log of 1 is zero.

Data example – histograms

```
hist gdppc
hist labor
hist trade
```

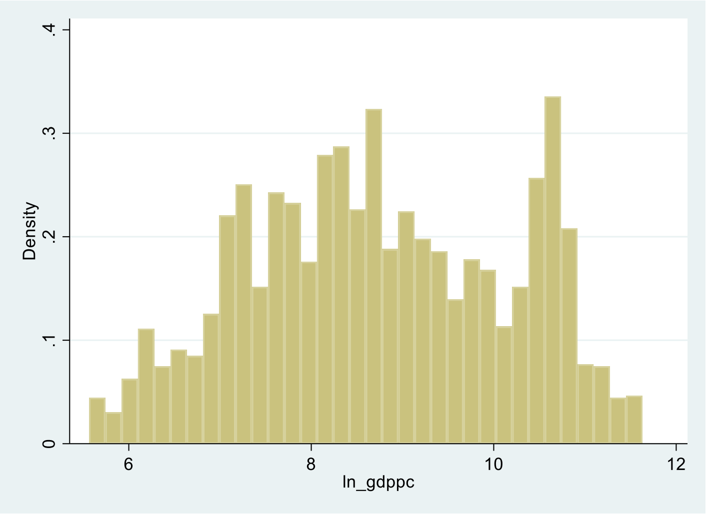


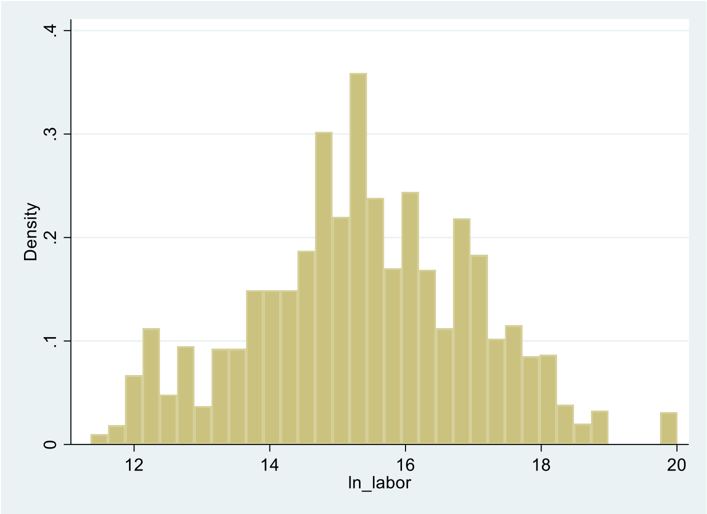


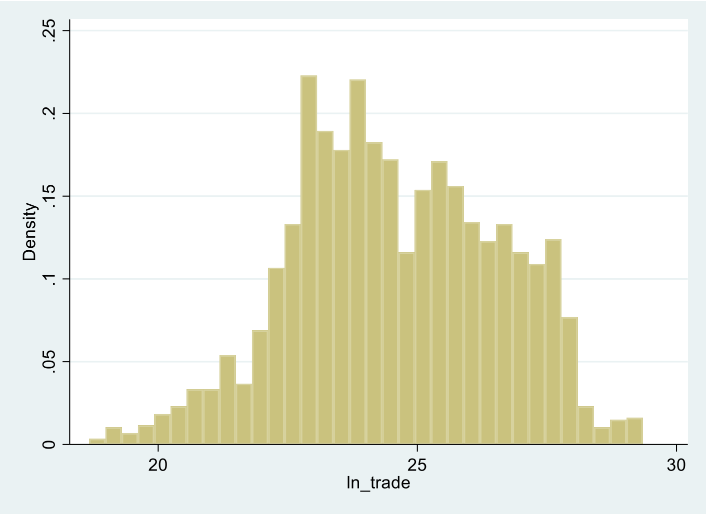


Data example – histograms

```
hist ln_gdppc
hist ln_labor
hist ln_trade
```







Descriptive statistics

```
. sum gdppc trade labor // Pooled data

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       gdppc |      2,772    14925.78       19561   261.0194   112417.9
       trade |      2,772    2.39e+11    5.33e+11   1.28e+08   5.58e+12
       labor |      2,772    1.70e+07    4.54e+07      85987   4.89e+08

. xtsum gdppc trade labor // Heterogeneity by panel and time

Variable         |      Mean   Std. dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
gdppc    overall |  14925.78       19561   261.0194   112417.9 |     N =    2772
         between |              19404.61   293.4895   104003.7 |     n =     126
         within  |              2991.204  -14918.74   52165.38 |     T =      22
                 |                                            |
trade    overall |  2.39e+11    5.33e+11   1.28e+08   5.58e+12 |     N =    2772
         between |              5.20e+11   3.14e+08   4.33e+12 |     n =     126
         within  |              1.27e+11  -1.14e+12   1.49e+12 |     T =      22
                 |                                            |
labor    overall |  1.70e+07    4.54e+07      85987   4.89e+08 |     N =    2772
         between |              4.54e+07     132657   4.53e+08 |     n =     126
         within  |               3154440  -4.24e+07   5.27e+07 |     T =      22
```


See https://www.stata.com/manuals/xtxtsum.pdf

Fixed effects regression using xtreg, fe

$$Y_{it} = \alpha_i + \beta X_{it} + u_i + e_{it}$$

Outcome

Predictor(s)

Fixed effects option

Controlling for heteroskedasticity

Total number of cases (rows)

Total number of entities (*i*)

```
. xtreg ln_gdppc ln_trade ln_labor, fe robust

Fixed-effects (within) regression               Number of obs     =      2,772
Group variable: country1                        Number of groups  =        126

R-squared:                                      Obs per group:
     Within  = 0.6267                                         min =         22
     Between = 0.3872                                         avg =       22.0
     Overall = 0.3906                                         max =         22

                                                F(2,125)          =      87.57
corr(u_i, Xb) = 0.1067                          Prob > F          =     0.0000

                            (Std. err. adjusted for 126 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
    ln_gdppc | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ln_trade |   .3603947   .0737076     4.89   0.000     .2145182    .5062712
    ln_labor |    .053167   .1608747     0.33   0.742     -.265224     .371558
       _cons |  -.9384681   1.075791    -0.87   0.385    -3.067592    1.190656
-------------+----------------------------------------------------------------
     sigma_u |  1.1155513
     sigma_e |  .10989953
         rho |  .99038791   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The within entity errors u_i are correlated with the regressors in the fixed effects model.

If this number is < 0.05 then your model is ok. This is an *F-test* to see whether all the coefficients in the model are jointly different than zero.

Beta coefficients indicate the change in the output (y) when the predictors change one unit over time. In this example, all the variables are log-transformed, the interpretation is: when the predictor increases 1% over time, the output (y) changes β% (elasticity).

Two-tail p-values test the hypothesis that each coefficient is different from 0 (according to its *t*-value). A value lower than 0.05 will reject the null and conclude that the predictor has a significant effect on the outcome (95% significance).

Intraclass correlation (rho), shows how much of the variance in the output is explained by the difference across entities. In this example is 99%.

$$rho = \frac{(sigma\_u)^2}{(sigma\_u)^2 + (sigma\_e)^2}$$

sigma_u = sd of residuals within groups u_i
sigma_e = sd of residuals (overall error term) e_{it}

Entity and time fixed effects regression using xtreg, fe

$$Y_{it} = \alpha_i + \beta X_{it} + \delta_t + u_i + e_{it}$$

Outcome

Predictor(s)

Time fixed effects

Fixed effects option

Controlling for heteroskedasticity

Total number of cases (rows)

Total number of entities (i)

```
. xtreg ln_gdppc ln_trade ln_labor i.year, fe robust

Fixed-effects (within) regression               Number of obs     =      2,772
Group variable: country1                        Number of groups  =        126

R-squared:                                      Obs per group:
     Within  = 0.7083                                         min =         22
     Between = 0.7977                                         avg =       22.0
     Overall = 0.7581                                         max =         22

                                                F(23,125)         =      34.28
corr(u_i, Xb) = 0.7525                          Prob > F          =     0.0000

                             (Std. err. adjusted for 126 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
    ln_gdppc | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ln_trade |   .2401329   .0695213     3.45   0.001     .1025416    .3777242
    ln_labor |  -.2958837    .081081    -3.65   0.000     -.456353   -.1354145
             |
        year |
       2001  |   .0119809   .0042779     2.80   0.006     .0035144    .0204475
      ...          ...         ...        ...   ...          ...          ...
      ...          ...         ...        ...   ...          ...          ...
       2021  |   .2878247   .0705454     4.08   0.000     .1482065    .4274428
             |
       _cons |   7.213881   1.961627     3.68   0.000     3.331578    11.09619
-------------+----------------------------------------------------------------
     sigma_u |  1.0561892
     sigma_e |  .09753735
         rho |  .99154389   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

If this number is < 0.05 then your model is ok. This is an *F-test* to see whether all the coefficients in the model are jointly different than zero.

The within entity errors u_i are correlated with the regressors in the fixed effects model.

Beta coefficients indicate the change in the output (y) when the predictors change one unit over time. In this example, all the variables are log-transformed, the interpretation is: when the predictor increases 1% over time, the output (y) changes β% (elasticity).

Two-tail p-values test the hypothesis that each coefficient is different from 0 (according to its *t*-value). A value lower than 0.05 will reject the null and conclude that the predictor has a significant effect on the outcome (95% significance).

Intraclass correlation (rho), shows how much of the variance in the output is explained by the difference across entities. In this example is 99%.

sigma_u = sd of residuals within groups u_i
sigma_e = sd of residuals (overall error term) e_{it}

$$rho = \frac{(sigma\_u)^2}{(sigma\_u)^2 + (sigma\_e)^2}$$

Fixed effects regression using xtreg, fe (with lags on predictors)

$$Y_{it} = \alpha_i + \beta X_{it-1} + u_i + e_{it}$$

Outcome

Predictor(s)

Fixed effects option

Controlling for heteroskedasticity

Total number of cases (rows)

Total number of entities (*i*)

```
. xtreg ln_gdppc L1.ln_trade L1.ln_labor, fe robust

Fixed-effects (within) regression               Number of obs     =      2,646
Group variable: country1                        Number of groups  =        126

R-squared:                                      Obs per group:
     Within  = 0.6054                                         min =         21
     Between = 0.3771                                         avg =       21.0
     Overall = 0.3799                                         max =         21

                                                F(2,125)          =      81.17
corr(u_i, Xb) = 0.1265                          Prob > F          =     0.0000

                             (Std. err. adjusted for 126 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
    ln_gdppc | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ln_trade |
         L1. |   .3385586   .0703993     4.81   0.000     .1992297    .4778875
             |
    ln_labor |
         L1. |   .0581167   .1566956     0.37   0.711    -.2520033    .3682367
             |
       _cons |  -.4600892   1.082489    -0.43   0.672     -2.60247    1.682291
-------------+----------------------------------------------------------------
     sigma_u |  1.1260807
     sigma_e |  .10685653
         rho |  .99107579   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The within entity errors u_i are correlated with the regressors in the fixed effects model.

If this number is < 0.05 then your model is ok. This is an *F-test* to see whether all the coefficients in the model are jointly different than zero.

Beta coefficients indicate the change in the output (y) when the predictors one unit over time (a year before –"L1."). In this example, all the variables are log-transformed, the interpretation is: when the predictor increases 1% over time (a year before –"L1."), the output (y) changes β% (elasticity).

Two-tail p-values test the hypothesis that each coefficient is different from 0 (according to its *t*-value). A value lower than 0.05 will reject the null and conclude that the predictor has a significant effect on the outcome (95% significance).

Intraclass correlation (rho), shows how much of the variance in the output is explained by the difference across entities. In this example is about 98%.

$$rho = \frac{(sigma\_u)^2}{(sigma\_u)^2 + (sigma\_e)^2}$$

sigma_u = sd of residuals within groups u_i
sigma_e = sd of residuals (overall error term) e_{it}

Entity fixed effects regression using reghdfe

$$Y_{it} = \alpha_i + \beta X_{it} + u_i + e_{it}$$

Outcome

Predictor(s)

Fixed effects option

Controlling for correlation within panels

Total number of cases (rows)

```
. reghdfe ln_gdppc ln_trade ln_labor , absorb(country1) vce(cluster country1)
(MWFE estimator converged in 1 iterations)

HDFE Linear regression                            Number of obs   =      2,772
Absorbing 1 HDFE group                            F(   2,     125) =      87.57
Statistics robust to heteroskedasticity           Prob > F        =     0.0000
                                                  R-squared       =     0.9943
                                                  Adj R-squared   =     0.9940
                                                  Within R-sq.    =     0.6267
Number of clusters (country1) =        126        Root MSE        =     0.1099

                            (Std. err. adjusted for 126 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
    ln_gdppc | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ln_trade |   .3603947   .0737076     4.89   0.000     .2145182    .5062712
    ln_labor |    .053167   .1608747     0.33   0.742    -.265224     .371558
       _cons |  -.9384681   1.075791    -0.87   0.385    -3.067592    1.190656
------------------------------------------------------------------------------

Absorbed degrees of freedom:
-----------------------------------------------------+
 Absorbed FE | Categories  - Redundant  = Num. Coefs |
-------------+---------------------------------------|
    country1 |       126         126           0    *|
-----------------------------------------------------+
* = FE nested within cluster; treated as redundant for DoF computation
```

Total number of entities (*i*)

If this number is < 0.05 then your model is ok. This is an *F-test* to see whether all the coefficients in the model are jointly different than zero.

`R-squared` shows the percent of the variance in the outcome explained by the model. The `Adj R-squared`, accounts for the number of variables and their significant contribution to explaining the variation in the output variable.

Beta coefficients indicate the change in the output (y) when the predictors change one unit over time. In this example, all the variables are log-transformed, the interpretation is: when the predictor increases 1% over time, the output (y) changes β% (elasticity).

Two-tail p-values test the hypothesis that each coefficient is different from 0 (according to its *t*-value).
A value lower than 0.05 will reject the null and conclude that the predictor has a significant effect on the outcome (95% significance).

NOTE: Use `reghdfe` when controlling for multiple fixed effects or when `xtreg, fe` cannot run due to the number of panels.

Entity and time fixed effects regression using reghdfe

$$Y_{it} = \alpha_i + \beta X_{it} + u_i + e_{it}$$

Outcome

Predictor(s)

Fixed effects option

Controlling for correlation within panels

Total number of cases (rows)

```
. reghdfe ln_gdppc ln_trade ln_labor , absorb(country1 year) vce(cluster country1)
(MWFE estimator converged in 2 iterations)

HDFE Linear regression                            Number of obs   =      2,772
Absorbing 2 HDFE groups                           F(   2,    125) =      11.37
Statistics robust to heteroskedasticity           Prob > F        =     0.0000
                                                  R-squared       =     0.9955
                                                  Adj R-squared   =     0.9953
                                                  Within R-sq.    =     0.3050
Number of clusters (country1) =        126        Root MSE        =     0.0976

                           (Std. err. adjusted for 126 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
    ln_gdppc | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ln_trade |   .2401329   .0695213     3.45   0.001     .1025416    .3777242
    ln_labor |  -.2958837    .081081    -3.65   0.000     -.456353   -.1354145
       _cons |   7.381277   1.999695     3.69   0.000     3.423632    11.33892
------------------------------------------------------------------------------

Absorbed degrees of freedom:
-----------------------------------------------------+
 Absorbed FE | Categories  - Redundant  = Num. Coefs |
-------------+---------------------------------------|
    country1 |       126         126           0    *|
        year |        22           0          22     |
-----------------------------------------------------+
* = FE nested within cluster; treated as redundant for DoF computation
```

If this number is < 0.05 then your model is ok. This is an *F-test* to see whether all the coefficients in the model are jointly different than zero.

Total number of entities (*i*)

`R-squared` shows the percent of the variance in the outcome explained by the model. The `Adj R-squared`, accounts for the number of variables and their significant contribution to explaining the variation in the output variable.

Beta coefficients indicate the change in the output (y) when the predictors change one unit over time. In this example, all the variables are log-transformed, the interpretation is: when the predictor increases 1% over time, the output (y) changes β% (elasticity).

Two-tail p-values test the hypothesis that each coefficient is different from 0 (according to its *t*-value).
A value lower than 0.05 will reject the null and conclude that the predictor has a significant effect on the outcome (95% significance).

Entity fixed effects regression with lags using reghdfe

$$Y_{it} = \alpha_i + \beta X_{it} + u_i + e_{it}$$

Outcome

Predictor(s)

Fixed effects option

Controlling for correlation within panels

Total number of cases (rows)

```
reghdfe ln_gdppc L1.ln_trade L1.ln_labor , absorb(country1 ) vce(cluster country1)
(MWFE estimator converged in 1 iterations)

HDFE Linear regression                            Number of obs   =      2,646
Absorbing 1 HDFE group                            F(   2,     125) =      81.17
Statistics robust to heteroskedasticity           Prob > F        =     0.0000
                                                  R-squared       =     0.9946
                                                  Adj R-squared   =     0.9943
                                                  Within R-sq.    =     0.6054
Number of clusters (country1) =        126        Root MSE        =     0.1069

                          (Std. err. adjusted for 126 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
    ln_gdppc | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ln_trade |
         L1. |   .3385586   .0703993     4.81   0.000     .1992297    .4778875
             |
    ln_labor |
         L1. |   .0581167   .1566956     0.37   0.711    -.2520033    .3682367
             |
       _cons |  -.4600892   1.082489    -0.43   0.672     -2.60247    1.682291
------------------------------------------------------------------------------

Absorbed degrees of freedom:
-----------------------------------------------------+
 Absorbed FE | Categories  - Redundant  = Num. Coefs |
-------------+---------------------------------------|
    country1 |       126         126           0    *|
-----------------------------------------------------+
* = FE nested within cluster; treated as redundant for DoF computation
```

If this number is < 0.05 then your model is ok. This is an *F-test* to see whether all the coefficients in the model are jointly different than zero.

Total number of entities (*i*)

`R-squared` shows the percent of the variance in the outcome explained by the model. The `Adj R-squared`, accounts for the number of variables and their significant contribution to explaining the variation in the output variable.

Beta coefficients indicate the change in the output (y) when the predictors change one unit over time. In this example, all the variables are log-transformed, the interpretation is: when the predictor increases 1% over time, the output (y) changes β% (elasticity).

Two-tail p-values test the hypothesis that each coefficient is different from 0 (according to its *t*-value). A value lower than 0.05 will reject the null and conclude that the predictor has a significant effect on the outcome (95% significance).

NOTE: must type `xtset country1 year`, before using lags in `reghdfe`

A note on fixed effects

"...The fixed-effects model controls for all time-invariant differences between the individuals, so the estimated coefficients of the fixed-effects models cannot be biased because of omitted time-invariant characteristics...[like culture, religion, gender, race, etc].

One side effect of the features of fixed-effects models is that they cannot be used to investigate time-invariant causes of the dependent variables. Technically, time-invariant characteristics of the individuals are perfectly collinear with the person [or entity] dummies. Substantively, fixed-effects models are designed to study the causes of changes within a person [or entity]. A time-invariant characteristic cannot cause such a change, because it is constant for each person." [(Underline is mine) Kohler, Ulrich, Frauke Kreuter, *Data Analysis Using Stata*, 2nd ed., p.245]

RANDOM-EFFECTS MODEL

(Random Intercept, Partial Pooling Model)

The random effects idea

The rationale behind random effects model is that, unlike the fixed effects model, the variation across entities is assumed to be random and uncorrelated with the predictor or independent variables included in the model:

> "...the crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are stochastic or not" [Green, 2008, p.183]

If you have reason to believe that differences across entities have some influence on your dependent variable but are not correlated with the predictors then you should use random effects. An advantage of random effects is that you can include time invariant variables (i.e. gender). In the fixed effects model these variables are absorbed by the intercept.

Thanks to Mateus Dias for useful feedback.

The random effects idea

Random effects assume that the entity's error term is not correlated with the predictors which allows for time-invariant variables to play a role as explanatory variables.

In random-effects you need to specify those individual characteristics that may or may not influence the predictor variables. The problem with this is that some variables may not be available therefore leading to omitted variable bias in the model.

RE allows to generalize the inferences beyond the sample used in the model.

Random effects regression using xtreg, re

$$Y_{it} = \alpha_i + \beta X_{it} + \gamma Z_i + e_{it}$$

Outcome

Predictor(s)

Random effects option

Controlling for heteroskedasticity

Total number of cases (rows)

Total number of entities (i)

```
. xtreg ln_gdppc ln_trade ln_labor, re robust

Random-effects GLS regression                   Number of obs     =      2,772
Group variable: country1                        Number of groups  =        126

R-squared:                                      Obs per group:
     Within  = 0.6110                                         min =         22
     Between = 0.7295                                         avg =       22.0
     Overall = 0.7212                                         max =         22

                                                Wald chi2(2)      =     192.71
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =     0.0000

                           (Std. err. adjusted for 126 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
    ln_gdppc | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ln_trade |   .4175909   .0760404     5.49   0.000     .2685543    .5666274
    ln_labor |  -.1597685   .1312262    -1.22   0.223    -.4169671    .0974302
       _cons |   .9295612   .6361615     1.46   0.144    -.3172923    2.176415
-------------+----------------------------------------------------------------
     sigma_u |  .41594682
     sigma_e |  .10989953
         rho |  .93474564   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The between entity errors u_{it} are uncorrelated with the regressors in the random effects model.

If this number is < 0.05 then your model is ok. This is an *F-test* to see whether all the coefficients in the model are jointly different than zero.

Beta coefficients indicate the change in the output (y) when the predictors change one unit over time and across entities (average effect). In this example, all the variables are log-transformed, the interpretation is: when the predictor increases, on average, 1%, the output (y) changes β% (elasticity).

Two-tail p-values test the hypothesis that each coefficient is different from 0 (according to its *t*-value).
A value lower than 0.05 will reject the null and conclude that the predictor has a significant effect on the outcome (95% significance).

Intraclass correlation (rho), shows how much of the variance in the output is explained by the difference across entities. In this example is 99%.

$$rho = \frac{(sigma\_u)^2}{(sigma\_u)^2 + (sigma\_e)^2}$$

sigma_u = sd of residuals within groups u_i
sigma_e = sd of residuals (overall error term) e_{it}

FIXED OR RANDOM?

Which to choose?

Whenever there is a clear idea that individual characteristics of each entity or group affect the regressors, use fixed effects. For example, macroeconomic data collected for most countries overtime. There might be a good reason to believe that countries’ economic performance may be affected by their own internal characteristics: type of government, political environment, cultural characteristics, type of public policies, etc.

Random effects is used whenever there is reason to believe that individual characteristics have no effect on the regressors (uncorrelated).

Which to choose?

The Hausman-test tests whether the individual characteristics are correlated with the regressors (see Green, 2008, chapter 9). The null hypothesis is that they are not (random effects).

```
xtreg ln_gdppc ln_trade ln_labor, fe
estimates store fixed
xtreg ln_gdppc ln_trade ln_labor, re
estimates store random
hausman fixed random, sigmamore
```

```
. hausman fixed random, sigmamore

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed        random       Difference          Std. err.
-------------+----------------------------------------------------------------
    ln_trade |    .3603947     .4175909       -.0571962         .0026039
    ln_labor |     .053167    -.1597685        .2129354          .012825
------------------------------------------------------------------------------
                          b = Consistent under H0 and Ha; obtained from xtreg.
           B = Inconsistent under Ha, efficient under H0; obtained from xtreg.

Test of H0: Difference in coefficients not systematic

    chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            = 484.43
Prob > chi2 = 0.0000
```

If Prob > chi2 is < 0.05 use fixed effects

TESTS / DIAGNOSTICS

Do we need time fixed effects?

To see if time fixed effects are needed when running a FE model use the command `testparm`. It is a joint *F*-test to if all years jointly equal to 0 (type `help testparm` for more details).

```
xtreg ln_gdppc ln_trade ln_labor i.year, fe robust
testparm i.year
```

```
. testparm i.year

 ( 1)  2001.year = 0
 ( 2)  2002.year = 0
 ( 3)  2003.year = 0
 ( 4)  2004.year = 0
 ( 5)  2005.year = 0
 ( 6)  2006.year = 0
 ( 7)  2007.year = 0
 ( 8)  2008.year = 0
 ( 9)  2009.year = 0
 (10)  2010.year = 0
 (11)  2011.year = 0
 (12)  2012.year = 0
 (13)  2013.year = 0
 (14)  2014.year = 0
 (15)  2015.year = 0
 (16)  2016.year = 0
 (17)  2017.year = 0
 (18)  2018.year = 0
 (19)  2019.year = 0
 (20)  2020.year = 0
 (21)  2021.year = 0

       F( 21,   125) =    4.44
            Prob > F =    0.0000
```

The `Prob > F` is < 0.05, we fail to accept the null that the coefficients for the years are jointly equal to zero. In this case, time fixed effects are needed.

Do we need random effects?

The LM test helps you decide between a random effects regression and a simple OLS regression. The null hypothesis in the LM test is that variances across entities is equal to zero. This is, no significant difference across units (i.e. no panel effect). The command in Stata is `xttset0` type it right after running the random effects model

```
xtreg ln_gdppc ln_trade ln_labor, re robust
xttest0
```

```
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

        ln_gdppc[country1,t] = Xb + u[country1] + e[country1,t]

        Estimated results:
                        |       Var     SD = sqrt(Var)
                ---------+-----------------------------
                ln_gdppc |   2.022383       1.422105
                       e |   .0120779       .1098995
                       u |   .1730118       .4159468

        Test: Var(u) = 0
                             chibar2(01) = 19981.51
                          Prob > chibar2 =   0.0000
```

`Prob > chibar2` < 0.05, we fail to accept the null hypothesis and conclude that random effects are needed.

Are the panels correlated? [B-P/LM test]

According to Baltagi, cross-sectional dependence is a problem in macro panels with long time series (over 20-30 years). The null hypothesis in the B-P/LM test of independence is that residuals across entities are not correlated. The user-defined command to run this test is `xttest2` (run it after `xtreg, fe`):

```
ssc install xttest2

xtreg ln_gdppc ln_trade ln_labor, fe robust

xttest2
```

```
. xttest2

Correlation matrix of residuals:

                                   [OMITTED]

Breusch-Pagan LM test of independence: chi2(7875) = 73886.228, Pr = 0.0000
Based on 22 complete observations over panel units
```

`Pr` < 0.05, we fail to accept the null hypothesis and conclude that panel are correlated (cross-sectional dependence).

Are the panels correlated? [Pasaran CD test]

As mentioned in the previous slide, cross-sectional dependence is more of an issue in macro panels with long time series (over 20-30 years) than in micro panels.
Pasaran CD (cross-sectional dependence) test is used to test whether the residuals are correlated across entities*. Cross-sectional dependence can lead to bias in tests results (also called contemporaneous correlation). The null hypothesis is that residuals are not correlated. The command for the test is xtcsd, you have to install it typing:

```
ssc install xtcsd
xtreg ln_gdppc ln_trade ln_labor, fe robust
xtcsd, pesaran abs
```

`Pr` < 0.05, we fail to accept the null hypothesis and conclude that panel are correlated (cross-sectional dependence).

```
. xtcsd, pesaran abs

Pesaran's test of cross sectional independence =     9.266, Pr = 0.0000

Average absolute value of the off-diagonal elements =     0.588
```

Had cross-sectional dependence be present Hoechle suggests to use Driscoll and Kraay standard errors using the command xtscc (install it by typing `ssc install xtscc`). Type `help xtscc` for more details.

*Source: Hoechle, Daniel, “Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence”, http://fmwww.bc.edu/repec/bocode/x/xtscc_paper.pdf

Testing for heteroskedasticity

A test for heteroskedasticity is available for the fixed- effects model using the command `xttest3`. The null hypothesis is homoskedasticity (or constant variance). This is a user-written program, to install it type:

```
ssc install xttest3

xtreg ln_gdppc ln_trade ln_labor, fe robust

xttest3

. xttest3

Modified Wald test for groupwise heteroskedasticity

in fixed effect regression model


H0: sigma(i)^2 = sigma^2 for all i


chi2 (126)   =   3.3e+05

Prob>chi2 =        0.0000
```

We reject the null and conclude heteroskedasticity.

NOTE: Use the option 'robust' to obtain heteroskedasticity-robust standard errors (also known as Huber/White or sandwich estimators).

Testing for serial correlation

Serial correlation tests apply to macro panels with long time series (over 20-30 years). Not a problem in micro panels (with very few years). Serial correlation causes the standard errors of the coefficients to be smaller than they actually are and higher R-squared. A Lagram-Multiplier test for serial correlation is available using the command xtserial. This is a user-written program, to install it type:

```
ssc install xtserial

xtreg ln_gdppc ln_trade ln_labor, fe robust

xtserial ln_gdppc ln_trade ln_labor
```

```
. xtserial ln_gdppc ln_trade ln_labor

Wooldridge test for autocorrelation in panel data
H0: no first order autocorrelation
    F(  1,     125) =    289.854
           Prob > F =      0.0000
```

We reject the null and conclude serial correlation.

Type `help xtserial` for more details.

Table 1: Selection of Stata commands and options that produce robust standard error estimates for linear panel models.

Command	Option	SE estimates are robust to disturbances being	Notes
`reg`, `xtreg`	`robust`	heteroscedastic	
`reg`, `xtreg`	`cluster()`	heteroscedastic and autocorrelated	
`xtregar`		autocorrelated with AR(1)[1]	
`newey`		heteroscedastic and autocorrelated of type MA(q)[2]	
`xtgls`	`panels()`, `corr()`	heteroscedastic, contemporaneously cross-sectionally correlated, and autocorrelated of type AR(1)	$N < T$ required for feasibility; tends to produce optimistic SE estimates
`xtpcse`	`correlation()`	heteroscedastic, contemporaneously cross-sectionally correlated, and autocorrelated of type AR(1)	large-scale panel regressions with `xtpcse` take a lot of time
`xtscc`		heteroscedastic, autocorrelated with MA(q), and cross-sectionally dependent	

[1] AR(1) refers to first-order autoregression

[2] MA(q) denotes autocorrelation of the moving average type with lag length q.

Source: Hoechle, Daniel, "Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence", page 4, http://fmwww.bc.edu/repec/bocode/x/xtscc_paper.pdf

Fixed Effects using Least Squares Dummy Variable model (LSDV)

Using reg, xtreg, reghdfe

```
reg ln_gdppc ln_trade ln_labor bn.country1,  vce(cluster country1) hascons
outreg2 using my_reg.doc, replace ctitle(Using -reg-) keep(ln_trade ln_labor) addtext(Country FE, YES)

xtreg ln_gdppc ln_trade ln_labor, fe robust
outreg2 using my_reg.doc, append ctitle(Using -xtreg-) addtext(Country FE, YES)

reghdfe ln_gdppc ln_trade ln_labor , absorb(country1) vce(cluster country1)
outreg2 using my_reg.doc, append ctitle(Using -reghdfe-) addtext(Country FE, YES)
```

	(1)	(2)	(3)
VARIABLES	Using -reg-	Using -xtreg-	Using -reghdfe-
ln_trade	0.360***	0.360***	0.360***
	(0.0754)	(0.0737)	(0.0737)
ln_labor	0.0532	0.0532	0.0532
	(0.165)	(0.161)	(0.161)
Constant		-0.938	-0.938
		(1.076)	(1.076)
Observations	2,772	2,772	2,772
R-squared	0.994	0.627	0.994
Country FE	YES	YES	YES
Number of country1		126	

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Using reg, xtreg, reghdfe

```
reg ln_gdppc ln_trade ln_labor i.year bn.country1,  vce(cluster country1) hascons
outreg2 using my_reg1.doc, replace ctitle(Using -reg-) ///
                           keep(ln_trade ln_labor) ///
                           addtext(Country FE, YES, Year FE, YES)

xtreg ln_gdppc ln_trade ln_labor i.year, fe robust
outreg2 using my_reg1.doc, append ctitle(Using -xtreg-) ///
                           keep(ln_trade ln_labor) ///
                           addtext(Country FE, YES, Year FE, YES)

reghdfe ln_gdppc ln_trade ln_labor , absorb(country1 year) vce(cluster country1)
outreg2 using my_reg1.doc, append ctitle(Using -reghdfe-) ///
                           addtext(Country FE, YES, Year FE, YES)
```

VARIABLES	(1) Using -reg-	(2) Using -xtreg-	(3) Using -reghdfe-
ln_trade	0.240***	0.240***	0.240***
	(0.0712)	(0.0695)	(0.0695)
ln_labor	-0.296***	-0.296***	-0.296***
	(0.0830)	(0.0811)	(0.0811)
Constant		7.214***	7.381***
		(1.962)	(2.000)
Observations	2,772	2,772	2,772
R-squared	0.996	0.708	0.996
Country FE	YES	YES	YES
Year FE	YES	YES	YES
Number of country1		126	

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

OLS No FE / OLS FE

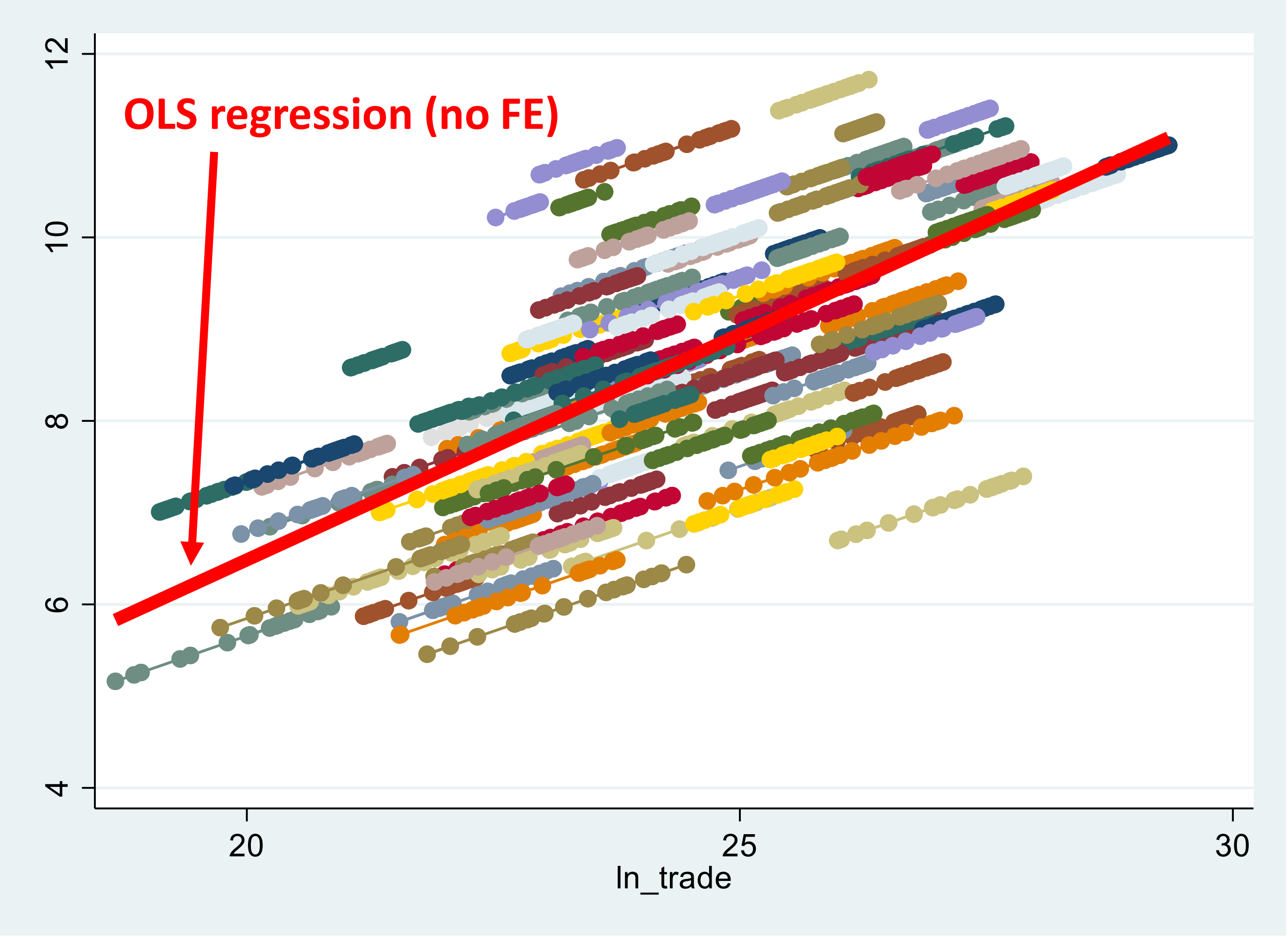
```
reg ln_gdppc ln_trade, robust
outreg2 using my_reg2.doc, replace ctitle(OLS No FE)

reg ln_gdppc ln_trade bn.country1,  vce(cluster country1) hascons
outreg2 using my_reg2.doc, append ctitle(OLS with FE) keep(ln_trade)
```

	(1)	(2)
VARIABLES	OLS No FE	OLS with FE
ln_trade	0.492***	0.371***
	(0.00869)	(0.0479)
Constant	-3.358***	
	(0.215)	
Observations	2,772	2,772
R-squared	0.471	0.994

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

OLS No FE / OLS FE

```
reg ln_gdppc ln_trade bn.country1,  vce(cluster country1) hascons
predict ln_gdppc_hat
separate ln_gdppc_hat, by(country1)
twoway connected ln_gdppc_hat1-ln_gdppc_hat99 ln_trade, legend(off) || ///
       connected ln_gdppc_hat100-ln_gdppc_hat126 ln_trade, legend(off) || ///
       lfit ln_gdppc ln_trade, clwidth(vthick) clcolor(red)
```



Suggested books / references

- *Introduction to econometrics* / James H. Stock, Mark W. Watson. 2nd ed., Boston: Pearson Addison Wesley, 2007.
- *Econometric Analysis of Panel Data*, Badi H. Baltagi, Wiley, 2008.
- *Econometric Analysis* / William H. Greene. 6th ed., Upper Saddle River, N.J. : Prentice Hall, 2008.
- *An Introduction to Modern Econometrics Using Stata*/ Christopher F. Baum, Stata Press, 2006.
- *Data analysis using regression and multilevel/hierarchical models* / Andrew Gelman, Jennifer Hill. Cambridge ; New York : Cambridge University Press, 2007.
- *Data Analysis Using Stata*/ Ulrich Kohler, Frauke Kreuter, 2 nd ed., Stata Press, 2009.
- *Statistics with Stata* / Lawrence Hamilton, Thomson Books/Cole, 2006.
- *Statistical Analysis: an interdisciplinary introduction to univariate & multivariate methods* / Sam Kachigan, New York : Radius Press, c1986
- "Beyond "Fixed Versus Random Effects": A framework for improving substantive and statistical analysis of panel, time-series cross-sectional, and multilevel data" / Brandom Bartels http://polmeth.wustl.edu/retrieve.php?id=838
- "Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence" / Daniel Hoechle, http://fmwww.bc.edu/repec/bocode/x/xtscc_paper.pdf
- *Designing Social Inquiry: Scientific Inference in Qualitative Research* / Gary King, Robert O.Keohane, Sidney Verba, Princeton University Press, 1994.
- *Unifying Political Methodology: The Likelihood Theory of Statistical Inference* / Gary King, Cambridge University Press, 1989.