PRINCETON UNIVERSITY

Panel Data Analysis
Fixed & Random Effects
(using Stata 10.x)

(ver. 4.1)

Oscar Torres-Reyna
Data Consultant
otorres@princeton.edu

http://dss.princeton.edu/training/

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

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

Panel data looks like this

country	year	Y	X1	X2	X3
1	2000	6.0	7.8	5.8	1.3
1	2001	4.6	0.6	7.9	7.8
1	2002	9.4	2.1	5.4	1.1
2	2000	9.1	1.3	6.7	4.1
2	2001	8.3	0.9	6.6	5.0
2	2002	0.6	9.8	0.4	7.2
3	2000	9.1	0.2	2.6	6.4
3	2001	4.8	5.9	3.2	6.4
3	2002	9.1	5.2	6.9	2.1

Panel data allows you to control for variables you cannot observe or measure like cultural factors or difference in business practices across companies; or variables that change over time but not across entities (i.e. national policies, federal regulations, international agreements, etc.). This is, it accounts for individual heterogeneity.

With panel data you can include variables at different levels of analysis (i.e. students, schools, districts, states) suitable for multilevel or hierarchical modeling.

Some drawbacks 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)

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

In this document we focus on two techniques use to analyze panel data:

- Fixed effects
- Random effects

The Stata command to run fixed/random effecst is `xtreg`.

Before using `xtreg` you need to set Stata to handle panel data by using the command `xtset`. type:

`xtset` *`country year`*

```
. xtset country year
       panel variable:   country (strongly balanced)
        time variable:   year, 1990 to 1999
                delta:   1 unit
```

In this case "`country`" represents the entities or panels (*i*) and "`year`" represents the time variable (*t*).

The note "`(strongly balanced)`" refers to the fact that all countries have data for all years. If, for example, one country does not have data for one year then the data is unbalanced. Ideally you would want to have a balanced dataset but this is not always the case, however you can still run the model.

NOTE: If you get the following error after using `xtset`:

```
varlist:  country:  string variable not allowed
```

You need to convert '`country`' to numeric, type:

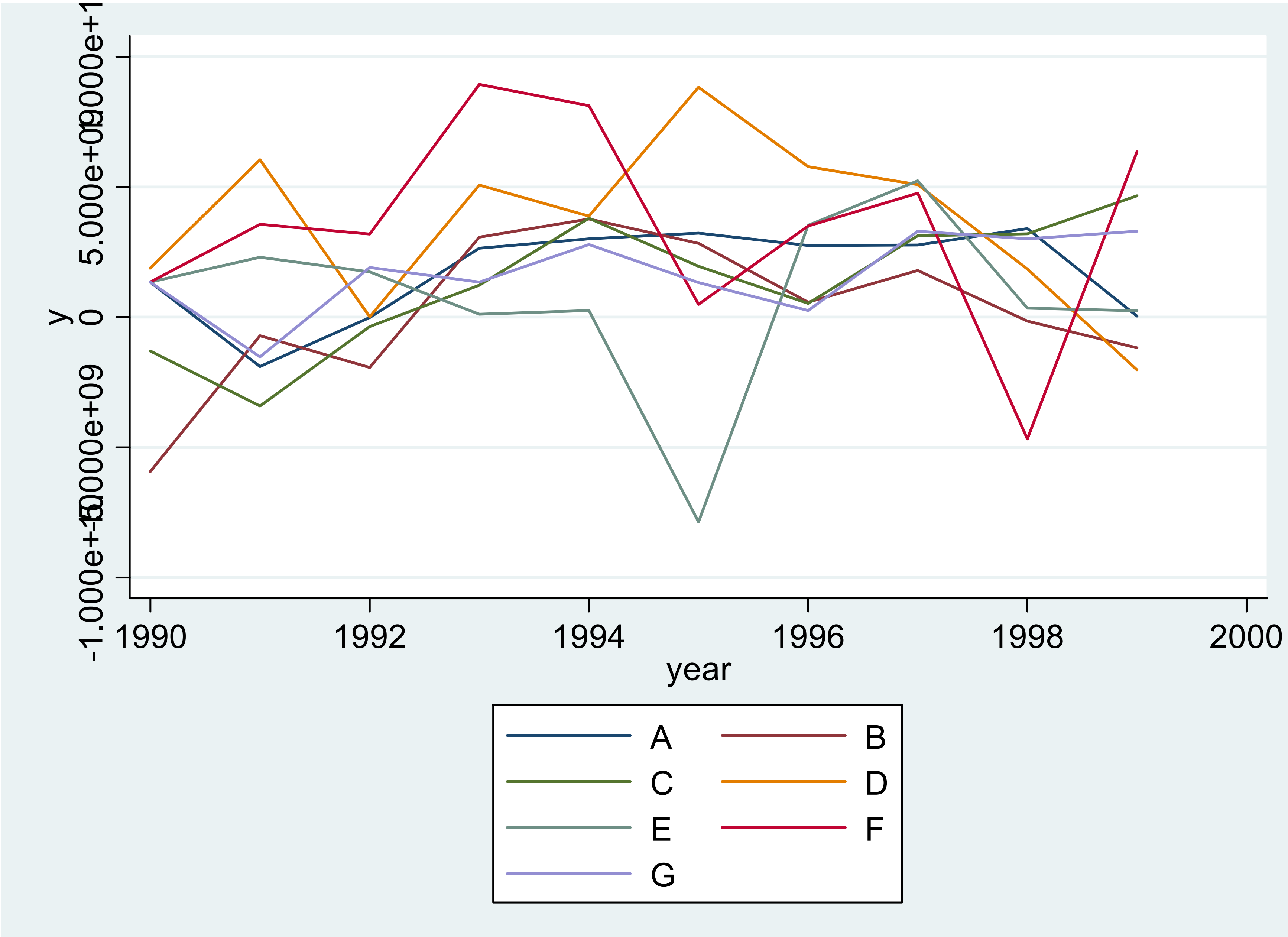
```
encode country, gen(country1)
```

Use '`country1`' instead of '`country`' in the `xtset` command

Exploring panel data

```
use http://dss.princeton.edu/training/Panel101.dta
xtset country year
xtline y
```

```
xtline y, overlay
```



FIXED-EFFECTS MODEL

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

Use fixed-effects (FE) whenever you are only interested in analyzing the impact of variables that vary over time.

FE explore the relationship between predictor and outcome variables within an entity (country, person, company, etc.). Each entity has its own individual characteristics that may or may not influence the predictor variables (for example being a male or female could influence the opinion toward certain issue or the political system of a particular country could have some effect on trade or GDP or the business practices of a company may influence its stock price).

When using FE we assume that something within the individual may impact or bias the predictor or outcome variables and we need to control for this. This is the rationale behind the assumption of the correlation between entity's error term and predictor variables. FE remove the effect of those time-invariant characteristics from the predictor variables so we can assess the predictors' net effect.

Another important assumption of the FE model is that those time-invariant characteristics are unique to the individual and should not be correlated with other individual characteristics. Each entity is different therefore the entity's error term and the constant (which captures individual characteristics) should not be correlated with the others. If the error terms are correlated then FE is no suitable since inferences may not be correct and you need to model that relationship (probably using random-effects), this is the main rationale for the Hausman test (presented later on in this document).

The equation for the fixed effects model becomes:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \qquad \text{[eq.1]}$$

Where

- α_i (i=1….n) is the unknown intercept for each entity (*n* entity-specific intercepts).
- Y_{it} is the dependent variable (DV) where *i* = entity and *t* = time.
- X_{it} represents one independent variable (IV),
- β_1 is the coefficient for that IV,
- u_{it} is the error term

"The key insight is that if the unobserved variable does not change over time, then any changes in the dependent variable must be due to influences other than these fixed characteristics." (Stock and Watson, 2003, p.289-290).

"In the case of time-series cross-sectional data the interpretation of the beta coefficients would be "…for a given country, as *X* varies *across time* by one unit, *Y* increases or decreases by *β* units" (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).

Fixed-effects will not work well with data for which within-cluster variation is minimal or for slow changing variables over time.

Another way to see the fixed effects model is by using binary variables. So the equation for the fixed effects model becomes:

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \ldots + \beta_k X_{k,it} + \gamma_2 E_2 + \ldots + \gamma_n E_n + u_{it} \qquad \text{[eq.2]}$$

Where

- –Y_{it} is the dependent variable (DV) where i = entity and t = time.
- –$X_{k,it}$ represents independent variables (IV),
- –β_k is the coefficient for the IVs,
- –u_{it} is the error term
- –E_n is the entity n. Since they are binary (dummies) you have n-1 entities included in the model.
- –γ_2 Is the coefficient for the binary repressors (entities)

Both eq.1 and eq.2 are equivalents:

> "the slope coefficient on *X* is the same from one [entity] to the next. The [entity]-specific intercepts in [eq.1] and the binary regressors in [eq.2] have the same source: the unobserved variable Z_i that varies across states but not over time." (Stock and Watson, 2003, p.280)

You could add time effects to the entity effects model to have a *time and entity fixed effects regression model*:

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \ldots + \beta_k X_{k,it} + \gamma_2 E_2 + \ldots + \gamma_n E_n + \delta_2 T_2 + \ldots + \delta_t T_t + u_{it} \quad \text{[eq.3]}$$

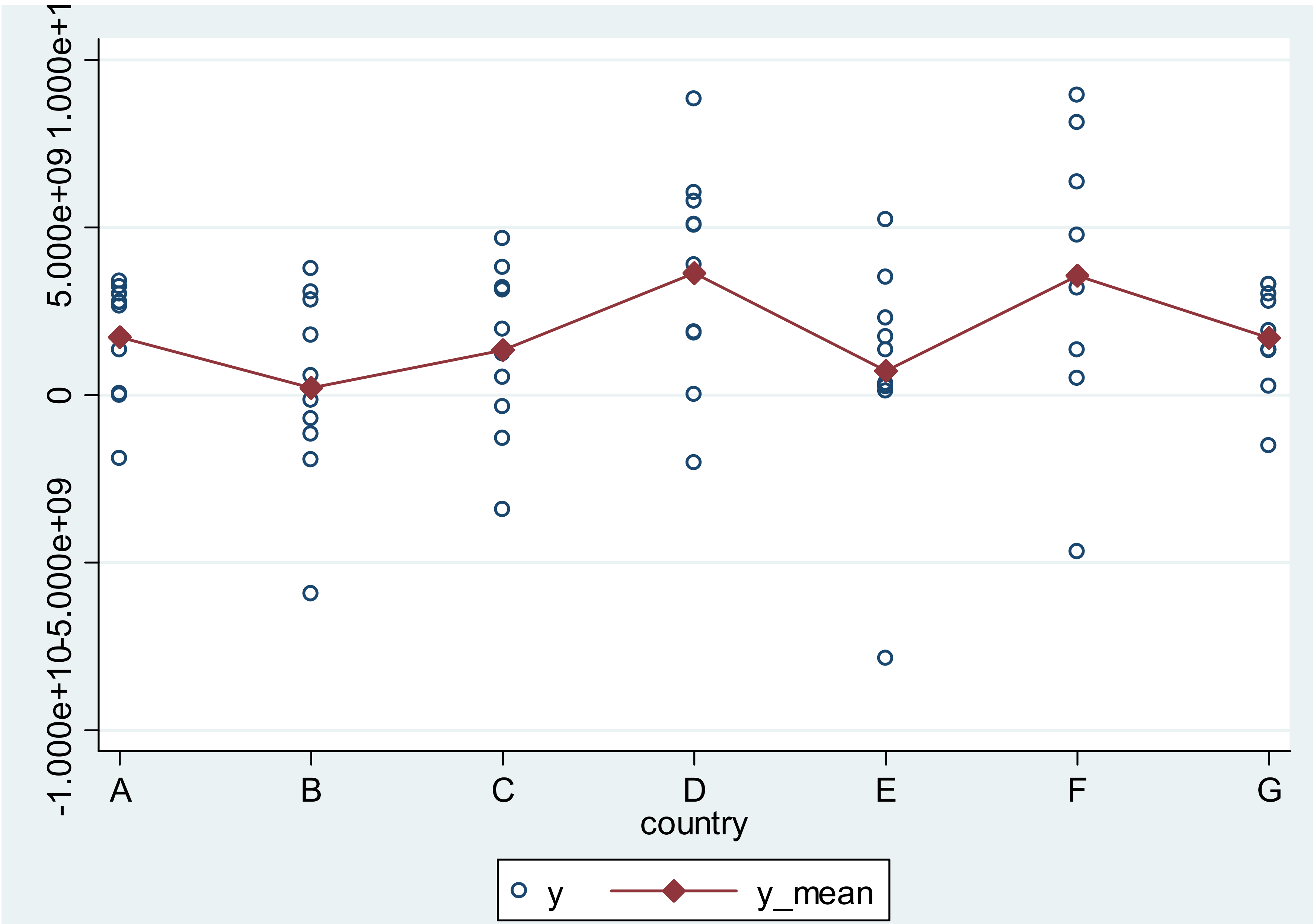
Where

- Y_{it} is the dependent variable (DV) where i = entity and t = time.
- $X_{k,it}$ represents independent variables (IV),
- β_k is the coefficient for the IVs,
- u_{it} is the error term
- E_n is the entity n. Since they are binary (dummies) you have n-1 entities included in the model.
- γ_2 is the coefficient for the binary regressors (entities) .
- T_t is time as binary variable (dummy), so we have t-1 time periods.
- δ_t is the coefficient for the binary time regressors .

Control for time effects whenever unexpected variation or special events my affect the outcome variable.

Fixed effects: Heterogeneity across countries (or entities)

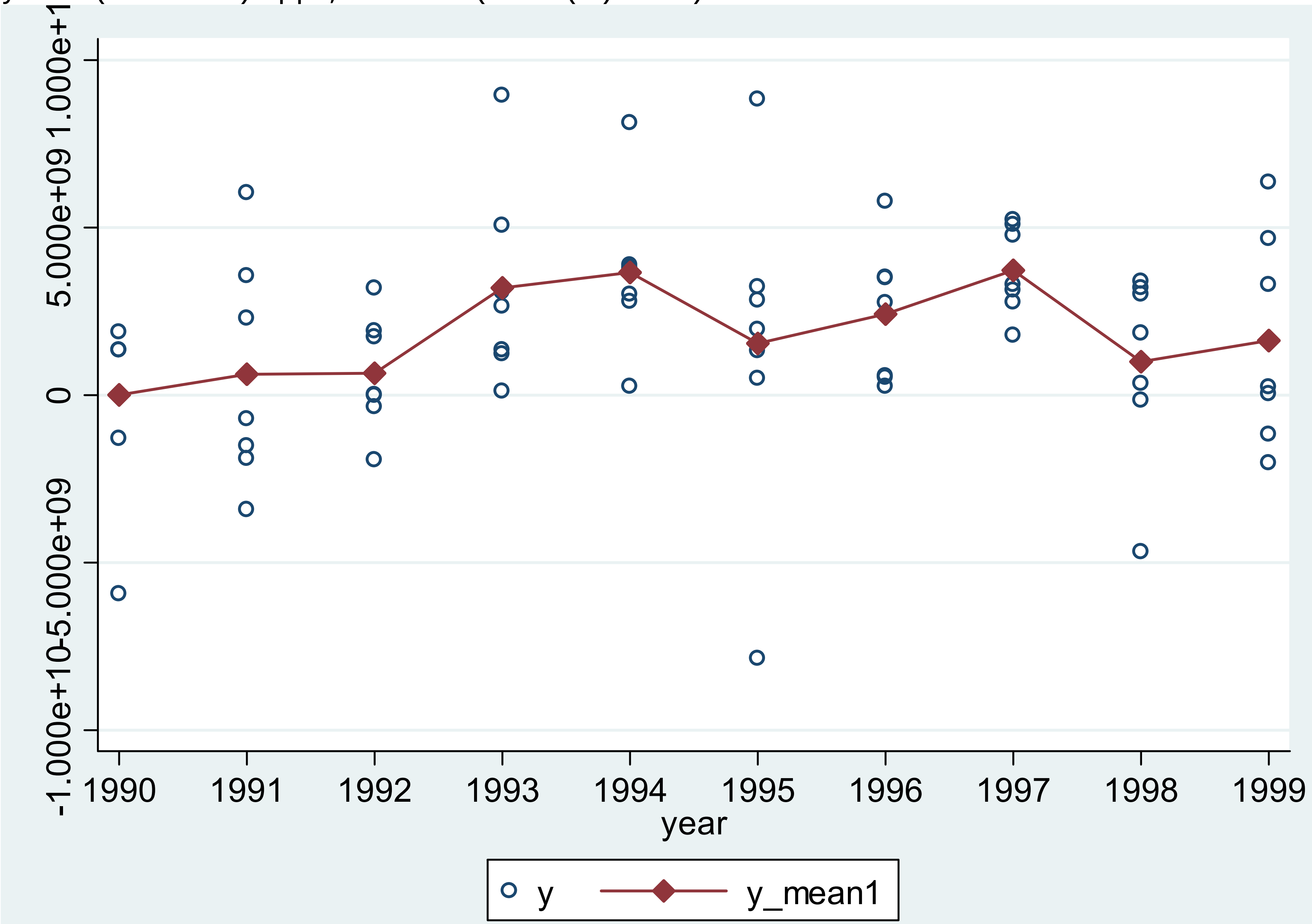
```
bysort country: egen y_mean=mean(y)
twoway scatter y country, msymbol(circle_hollow) || connected y_mean country,
msymbol(diamond) || , xlabel(1 "A" 2 "B" 3 "C" 4 "D" 5 "E" 6 "F" 7 "G")
```



Heterogeneity: unobserved variables that do not change over time

Fixed effects: Heterogeneity across years

```
bysort year: egen y_mean1=mean(y)
twoway scatter y year, msymbol(circle_hollow) || connected y_mean1 year,
msymbol(diamond) || , xlabel(1990(1)1999)
```



Heterogeneity: unobserved variables that do not change over time

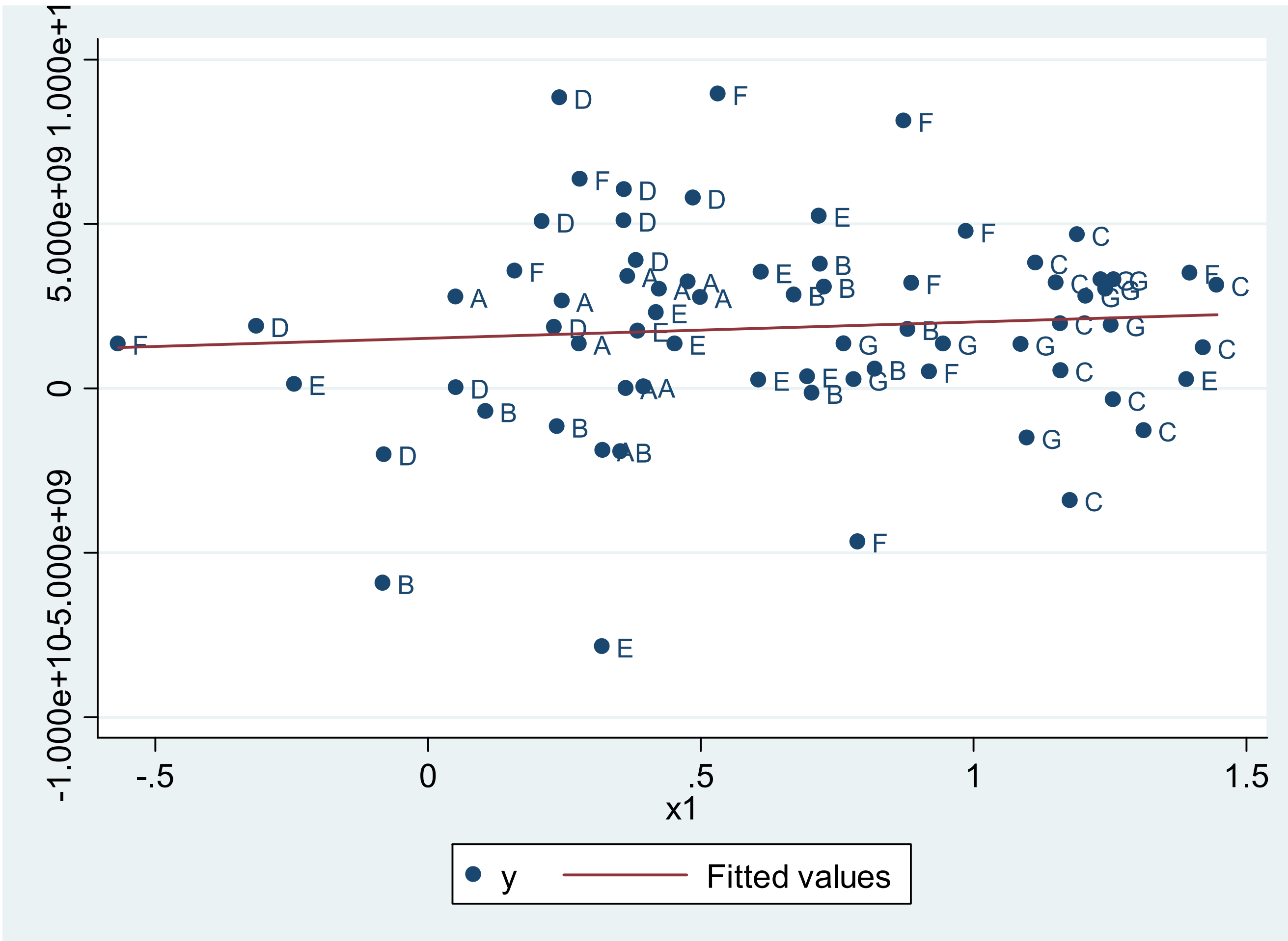
OLS regression

```
. regress y x1

      Source |       SS       df       MS              Number of obs =      70
-------------+------------------------------           F(  1,    68) =    0.40
       Model |  3.7039e+18     1  3.7039e+18           Prob > F      =  0.5272
    Residual |  6.2359e+20    68  9.1705e+18           R-squared     =  0.0059
-------------+------------------------------           Adj R-squared = -0.0087
       Total |  6.2729e+20    69  9.0912e+18           Root MSE      =  3.0e+09

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   4.95e+08   7.79e+08     0.64   0.527    -1.06e+09    2.05e+09
       _cons |   1.52e+09   6.21e+08     2.45   0.017     2.85e+08    2.76e+09
------------------------------------------------------------------------------
```

```
twoway scatter y x1,
mlabel(country) || lfit y x1,
clstyle(p2)
```



Fixed Effects using least squares dummy variable model (LSDV)

```
. xi: regress y x1 i.country
i.country         _Icountry_1-7       (naturally coded; _Icountry_1 omitted)

      Source |       SS       df       MS              Number of obs =      70
-------------+------------------------------           F(  7,    62) =    2.61
       Model |  1.4276e+20     7  2.0394e+19           Prob > F      =  0.0199
    Residual |  4.8454e+20    62  7.8151e+18           R-squared     =  0.2276
-------------+------------------------------           Adj R-squared =  0.1404
       Total |  6.2729e+20    69  9.0912e+18           Root MSE      =  2.8e+09

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.48e+09   1.11e+09     2.24   0.029     2.63e+08    4.69e+09
 _Icountry_2 |  -1.94e+09   1.26e+09    -1.53   0.130    -4.47e+09    5.89e+08
 _Icountry_3 |  -2.60e+09   1.60e+09    -1.63   0.108    -5.79e+09    5.87e+08
 _Icountry_4 |   2.28e+09   1.26e+09     1.81   0.075    -2.39e+08    4.80e+09
 _Icountry_5 |  -1.48e+09   1.27e+09    -1.17   0.247    -4.02e+09    1.05e+09
 _Icountry_6 |   1.13e+09   1.29e+09     0.88   0.384    -1.45e+09    3.71e+09
 _Icountry_7 |  -1.87e+09   1.50e+09    -1.25   0.218    -4.86e+09    1.13e+09
       _cons |   8.81e+08   9.62e+08     0.92   0.363    -1.04e+09    2.80e+09
------------------------------------------------------------------------------
```

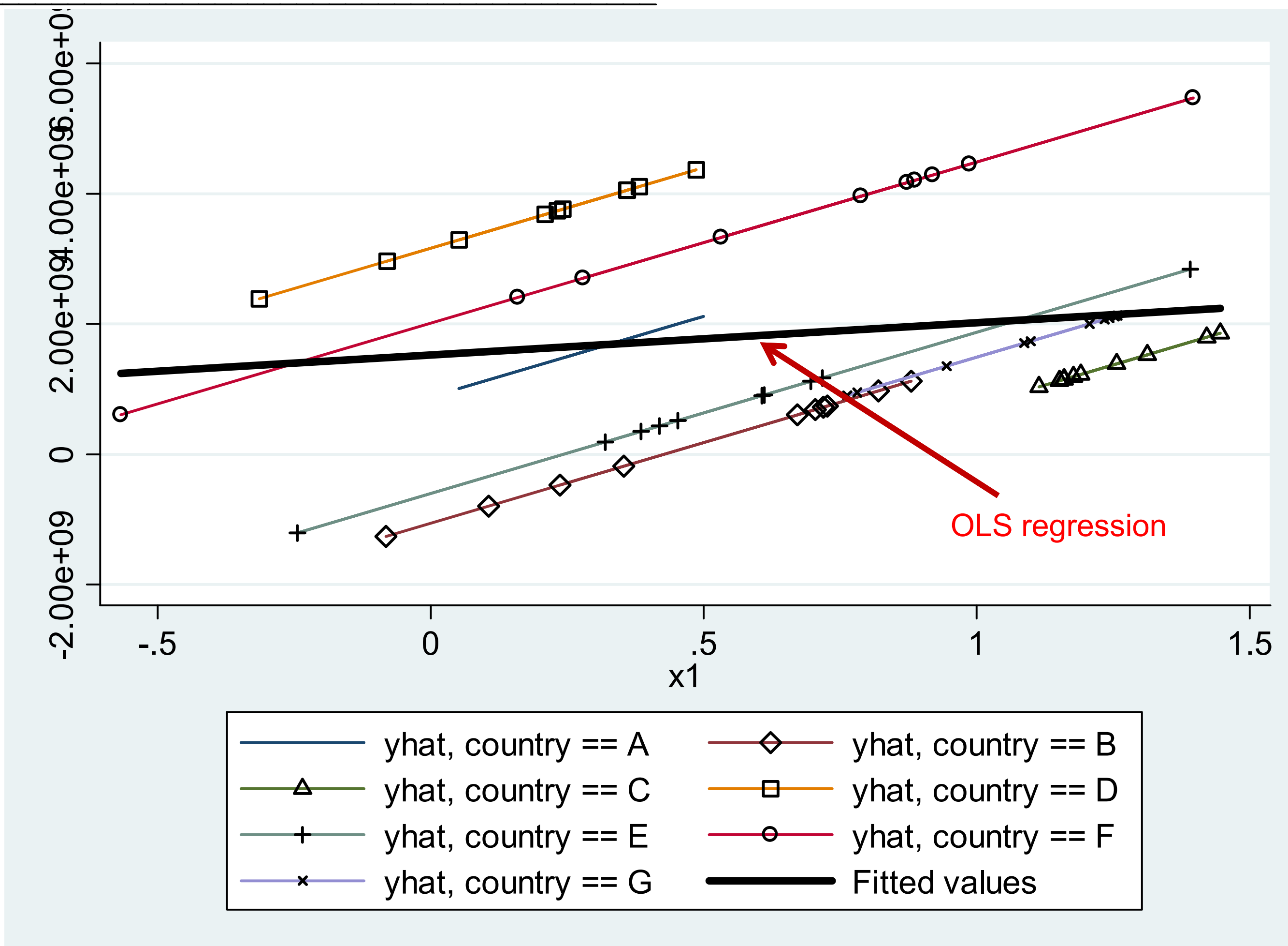
```
xi: regress y x1 i.country

predict yhat

separate y, by(country)

separate yhat, by(country)

twoway connected yhat1-yhat7
x1, msymbol(none
diamond_hollow triangle_hollow
square_hollow + circle_hollow
x) msize(medium) mcolor(black
black black black black black
black) || lfit y x1,
clwidth(thick) clcolor(black)
```



NOTE: In Stata 11 you do not need "xi:" when adding dummy variables

The least square dummy variable model (LSDV) provides a good way to understand fixed effects.

The effect of x1 is mediated by the differences across countries.

By adding the dummy for each country we are estimating the pure effect of x1 (by controlling for the unobserved heterogeneity).

Each dummy is absorbing the effects particular to each country.

```
regress y x1
estimates store ols
xi: regress y x1 i.country
estimates store ols_dum
estimates table ols ols_dum, star stats(N)
```

```
. estimates table ols ols_dum, star stats(N)
```

Variable	ols	ols_dum
x1	4.950e+08	2.476e+09*
_Icountry_2		-1.938e+09
_Icountry_3		-2.603e+09
_Icountry_4		2.282e+09
_Icountry_5		-1.483e+09
_Icountry_6		1.130e+09
_Icountry_7		-1.865e+09
_cons	1.524e+09*	8.805e+08
N	70	70

legend: * p<0.05; ** p<0.01; *** p<0.001

Comparing the fixed effects using dummies with `xtreg` we get the same results.

```
. xtreg  y  x1, fe
```

Using xtreg

```
Fixed-effects (within) regression               Number of obs      =        70
Group variable: country                         Number of groups   =         7

R-sq:  within  = 0.0747                         Obs per group: min =        10
       between = 0.0763                                        avg =      10.0
       overall = 0.0059                                        max =        10

                                                F(1,62)            =      5.00
corr(u_i, Xb)  = -0.5468                        Prob > F           =    0.0289

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.48e+09   1.11e+09     2.24   0.029     2.63e+08    4.69e+09
       _cons |   2.41e+08   7.91e+08     0.30   0.762    -1.34e+09    1.82e+09
-------------+----------------------------------------------------------------
     sigma_u |  1.818e+09
     sigma_e |  2.796e+09
         rho |  .29726926   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

OLS regression

```
. xi: regress y x1 i.country
i.country         _Icountry_1-7       (naturally coded; _Icountry_1 omitted)

      Source |       SS       df       MS              Number of obs =      70
-------------+------------------------------           F(  7,    62) =    2.61
       Model |  1.4276e+20     7  2.0394e+19           Prob > F      =  0.0199
    Residual |  4.8454e+20    62  7.8151e+18           R-squared     =  0.2276
-------------+------------------------------           Adj R-squared =  0.1404
       Total |  6.2729e+20    69  9.0912e+18           Root MSE      =  2.8e+09

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.48e+09   1.11e+09     2.24   0.029     2.63e+08    4.69e+09
 _Icountry_2 |  -1.94e+09   1.26e+09    -1.53   0.130    -4.47e+09    5.89e+08
 _Icountry_3 |  -2.60e+09   1.60e+09    -1.63   0.108    -5.79e+09    5.87e+08
 _Icountry_4 |   2.28e+09   1.26e+09     1.81   0.075    -2.39e+08    4.80e+09
 _Icountry_5 |  -1.48e+09   1.27e+09    -1.17   0.247    -4.02e+09    1.05e+09
 _Icountry_6 |   1.13e+09   1.29e+09     0.88   0.384    -1.45e+09    3.71e+09
 _Icountry_7 |  -1.87e+09   1.50e+09    -1.25   0.218    -4.86e+09    1.13e+09
       _cons |   8.81e+08   9.62e+08     0.92   0.363    -1.04e+09    2.80e+09
------------------------------------------------------------------------------
```

Fixed effects: *n* entity-specific intercepts (using `xtreg`)

$Y_{it} = \beta_1 X_{it} + \ldots + \beta_k X_{kt} + \alpha_i + e_{it}$ [see eq.1]

NOTE: Add the option 'robust' to control for heteroskedasticity

Outcome variable

Predictor variable(s)

Fixed effects option

Total number of cases (rows)

Total number of groups (entities)

```
. xtreg  y  x1, fe

Fixed-effects (within) regression               Number of obs      =        70
Group variable: country                         Number of groups   =         7

R-sq:  within  = 0.0747                         Obs per group: min =        10
       between = 0.0763                                        avg =      10.0
       overall = 0.0059                                        max =        10

                                                F(1,62)            =      5.00
corr(u_i, Xb)  = -0.5468                        Prob > F           =    0.0289

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.48e+09   1.11e+09     2.24   0.029     2.63e+08    4.69e+09
       _cons |   2.41e+08   7.91e+08     0.30   0.762    -1.34e+09    1.82e+09
-------------+----------------------------------------------------------------
     sigma_u |  1.818e+09
     sigma_e |  2.796e+09
         rho |  .29726926   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The 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 a test (F) to see whether all the coefficients in the model are different than zero.

Coefficients of the regressors. Indicate how much *Y* changes when *X* increases by one unit.

Two-tail p-values test the hypothesis that each coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (95%, you could choose also an alpha of 0.10), if this is the case then you can say that the variable has a significant influence on your dependent variable (y)

29.7% of the variance is due to differences across panels.

'rho' is known as the intraclass correlation

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

sigma_u = sd of residuals within groups u_i

sigma_e = sd of residuals (overall error term) e_i

t-values test the hypothesis that each coefficient is different from 0. To reject this, the t-value has to be higher than 1.96 (for a 95% confidence). If this is the case then you can say that the variable has a significant influence on your dependent variable (y). The higher the t-value the higher the relevance of the variable.

For more info see Hamilton, Lawrence, *Statistics with STATA*.

Another way to estimate fixed effects: *n* entity-specific intercepts (using **areg**)

$$Y_{it} = \beta_1 X_{it} + \ldots + \beta_k X_{kt} + \alpha_i + e_{it} \qquad \text{[see eq.1]}$$

Outcome variable

Predictor variable(s)

Hide the binary variables for each entity

```
. areg y x1, absorb(country)

Linear regression, absorbing indicators          Number of obs =        70
                                                 F(  1,    62) =      5.00
                                                 Prob > F      =    0.0289
                                                 R-squared     =    0.2276
                                                 Adj R-squared =    0.1404
                                                 Root MSE      =   2.8e+09

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.48e+09   1.11e+09     2.24   0.029     2.63e+08    4.69e+09
       _cons |   2.41e+08   7.91e+08     0.30   0.762    -1.34e+09    1.82e+09
-------------+----------------------------------------------------------------
     country |          F(6, 62) =      2.965   0.013          (7 categories)
```

If this number is < 0.05 then your model is ok. This is a test (F) to see whether all the coefficients in the model are different than zero.

R-square shows the amount of variance of Y explained by X

Adj R-square shows the same as R-sqr but adjusted by the number of cases and number of variables. When the number of variables is small and the number of cases is very large then Adj R-square is closer to R-square.

Coefficients of the regressors. Indicate how much *Y* changes when *X* increases by one unit.

NOTE: Add the option 'robust' to control for heteroskedasticity

"Although its output is less informative than regression with explicit dummy variables, areg does have two advantages. It speeds up exploratory work, providing quick feedback about whether a dummy variable approach is worthwhile. Secondly, when the variable of interest has many values, creating dummies for each of them could lead to too many variables or too large a model …." (Hamilton, 2006, p.180)

t-values test the hypothesis that each coefficient is different from 0. To reject this, the t-value has to be higher than 1.96 (for a 95% confidence). If this is the case then you can say that the variable has a significant influence on your dependent variable (y). The higher the t-value the higher the relevance of the variable.

Two-tail p-values test the hypothesis that each coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (95%, you could choose also an alpha of 0.10), if this is the case then you can say that the variable has a significant influence on your dependent variable (y)

Another way to estimate fixed effects: common intercept and n-1 binary regressors (using `dummies` and `regress`)

Notice the "xi:" (interaction expansion) to automatically generate dummy variables

Outcome variable

Predictor variable(s)

Notice the "i." before the indicator variable for entities

If this number is < 0.05 then your model is ok. This is a test (F) to see whether all the coefficients in the model are different than zero.

```
. xi: regress y x1 i.country
i.country         _Icountry_1-7       (naturally coded; _Icountry_1 omitted)

      Source |       SS       df       MS              Number of obs =      70
-------------+------------------------------           F(  7,    62) =    2.61
       Model |  1.4276e+20     7  2.0394e+19           Prob > F      =  0.0199
    Residual |  4.8454e+20    62  7.8151e+18           R-squared     =  0.2276
-------------+------------------------------           Adj R-squared =  0.1404
       Total |  6.2729e+20    69  9.0912e+18           Root MSE      =  2.8e+09

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   2.48e+09   1.11e+09     2.24   0.029     2.63e+08    4.69e+09
 _Icountry_2 |  -1.94e+09   1.26e+09    -1.53   0.130    -4.47e+09    5.89e+08
 _Icountry_3 |  -2.60e+09   1.60e+09    -1.63   0.108    -5.79e+09    5.87e+08
 _Icountry_4 |   2.28e+09   1.26e+09     1.81   0.075    -2.39e+08    4.80e+09
 _Icountry_5 |  -1.48e+09   1.27e+09    -1.17   0.247    -4.02e+09    1.05e+09
 _Icountry_6 |   1.13e+09   1.29e+09     0.88   0.384    -1.45e+09    3.71e+09
 _Icountry_7 |  -1.87e+09   1.50e+09    -1.25   0.218    -4.86e+09    1.13e+09
       _cons |   8.81e+08   9.62e+08     0.92   0.363    -1.04e+09    2.80e+09
------------------------------------------------------------------------------
```

Coefficients of the regressors indicate how much *Y* changes when *X* increases by one unit.

<u>R-square</u> shows the amount of variance of Y explained by X

NOTE: Add the option 'robust' to control for heteroskedasticity

t-values test the hypothesis that each coefficient is different from 0. To reject this, the t-value has to be higher than 1.96 (for a 95% confidence). If this is the case then you can say that the variable has a significant influence on your dependent variable (y). The higher the t-value the higher the relevance of the variable.

Two-tail p-values test the hypothesis that each coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (95%, you could choose also an alpha of 0.10), if this is the case then you can say that the variable has a significant influence on your dependent variable (y)

NOTE: In Stata 11 you do not need "xi:" when adding dummy variables

Fixed effects: comparing `xtreg (with fe)`, `regress (OLS with dummies)` and `areg`

To compare the previous methods type "`estimates store [name]`" after running each regression, at the end use the command "`estimates table…`" (see below):

```
xtreg  y x1 x2 x3, fe
estimates store fixed
xi: regress  y x1 x2 x3 i.country
estimates store ols
areg y x1 x2 x3, absorb(country)
estimates store areg
estimates table fixed ols areg, star stats(N r2 r2_a)
```

```
. estimates table fixed ols areg, star stats(N r2 r2_a)
```

Variable	fixed	ols	areg
x1	2.425e+09*	2.425e+09*	2.425e+09*
x2	1.823e+09	1.823e+09	1.823e+09
x3	3.097e+08	3.097e+08	3.097e+08
_Icountry_2		-5.961e+09	
_Icountry_3		-1.598e+09	
_Icountry_4		-2.091e+09	
_Icountry_5		-5.732e+09	
_Icountry_6		8.026e+08	
_Icountry_7		-1.375e+09	
_cons	-2.060e+08	2.073e+09	-2.060e+08
N	70	70	70
r2	.10092442	.24948198	.24948198
r2_a	-.03393692	.13690428	.13690428

legend: * p<0.05; ** p<0.01; *** p<0.001

All three commands provide the same results

Tip: When reporting the R-square use the one provided by either `regress` or `areg`.

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, <u>fixed-effects models are designed to study the causes of changes within a person [or entity]</u>. 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)

Random effects

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 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.

The random effects model is:

$$Y_{it} = \beta X_{it} + \alpha + \underbrace{u_{it}}_{\text{Between-entity error}} + \underbrace{\varepsilon_{it}}_{\text{Within-entity error}} \qquad \text{[eq.4]}$$

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.

You can estimate a random effects model using `xtreg` and the option `re`.

NOTE: Add the option 'robust' to control for heteroskedasticity

Outcome variable

Predictor variable(s)

Random effects option

If this number is < 0.05 then your model is ok. This is a test (F) to see whether all the coefficients in the model are different than zero.

Differences across units are uncorrelated with the regressors

```
. xtreg y x1, re

Random-effects GLS regression                   Number of obs      =        70
Group variable: country                         Number of groups   =         7

R-sq:  within  = 0.0747                         Obs per group: min =        10
       between = 0.0763                                        avg =      10.0
       overall = 0.0059                                        max =        10

Random effects u_i ~ Gaussian                   Wald chi2(1)       =      1.91
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.1669

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.25e+09   9.02e+08     1.38   0.167    -5.21e+08    3.02e+09
       _cons |   1.04e+09   7.91e+08     1.31   0.190    -5.13e+08    2.59e+09
-------------+----------------------------------------------------------------
     sigma_u |  1.065e+09
     sigma_e |  2.796e+09
         rho |  .12664193   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Two-tail p-values test the hypothesis that each coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (95%, you could choose also an alpha of 0.10), if this is the case then you can say that the variable has a significant influence on your dependent variable (y)

Interpretation of the coefficients is tricky since they include both the within-entity and between-entity effects. In the case of TSCS data represents the average effect of *X* over *Y* when *X* changes across time and between countries by one unit.

FIXED OR RANDOM?

Fixed or Random: Hausman test

To decide between fixed or random effects you can run a Hausman test where the null hypothesis is that the preferred model is random effects vs. the alternative the fixed effects (see Green, 2008, chapter 9). It basically tests whether the unique errors (u_i) are correlated with the regressors, the null hypothesis is they are not.

Run a fixed effects model and save the estimates, then run a random model and save the estimates, then perform the test. See below.

```
xtreg  y x1, fe
estimates store fixed
xtreg  y x1, re
estimates store random
hausman fixed random
```

```
. hausman fixed random

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed        random       Difference          S.E.
-------------+----------------------------------------------------------------
          x1 |    2.48e+09     1.25e+09        1.23e+09        6.41e+08
------------------------------------------------------------------------------
                          b = consistent under Ho and Ha; obtained from xtreg
           B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        3.67
                Prob>chi2 =      0.0553
```

If this is < 0.05 (i.e. significant) use fixed effects.

OTHER TESTS/ DIAGNOSTICS

Testing for time-fixed effects

To see if time fixed effects are needed when running a FE model use the command `testparm`. It is a joint test to see if the dummies for all years are equal to 0, if they are then no time fixed effects are needed.

`testparm _Iyear*`

In Stata 11 type:

`tesparm i.year`

```
. xi: xtreg  y  x1 i.year, fe
i.year            _Iyear_1990-1999    (naturally coded; _Iyear_1990 omitted)

Fixed-effects (within) regression               Number of obs      =        70
Group variable: country                         Number of groups   =         7

R-sq:  within  = 0.2323                         Obs per group: min =        10
       between = 0.0763                                        avg =      10.0
       overall = 0.1395                                        max =        10

                                                F(10,53)           =      1.60
corr(u_i, Xb)  = -0.2014                        Prob > F           =    0.1311

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.39e+09   1.32e+09     1.05   0.297    -1.26e+09    4.04e+09
 _Iyear_1991 |   2.96e+08   1.50e+09     0.20   0.844    -2.72e+09    3.31e+09
 _Iyear_1992 |   1.45e+08   1.55e+09     0.09   0.925    -2.96e+09    3.25e+09
 _Iyear_1993 |   2.87e+09   1.50e+09     1.91   0.061    -1.42e+08    5.89e+09
 _Iyear_1994 |   2.85e+09   1.66e+09     1.71   0.092    -4.84e+08    6.18e+09
 _Iyear_1995 |   9.74e+08   1.57e+09     0.62   0.537    -2.17e+09    4.12e+09
 _Iyear_1996 |   1.67e+09   1.63e+09     1.03   0.310    -1.60e+09    4.95e+09
 _Iyear_1997 |   2.99e+09   1.63e+09     1.84   0.072    -2.72e+08    6.26e+09
 _Iyear_1998 |   3.67e+08   1.59e+09     0.23   0.818    -2.82e+09    3.55e+09
 _Iyear_1999 |   1.26e+09   1.51e+09     0.83   0.409    -1.77e+09    4.29e+09
       _cons |  -3.98e+08   1.11e+09    -0.36   0.721    -2.62e+09    1.83e+09
-------------+----------------------------------------------------------------
     sigma_u |  1.547e+09
     sigma_e |  2.754e+09
         rho |  .23985725   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(6, 53) =     2.45               Prob > F = 0.0362

. testparm _Iyear*

 ( 1)  _Iyear_1991 = 0
 ( 2)  _Iyear_1992 = 0
 ( 3)  _Iyear_1993 = 0
 ( 4)  _Iyear_1994 = 0
 ( 5)  _Iyear_1995 = 0
 ( 6)  _Iyear_1996 = 0
 ( 7)  _Iyear_1997 = 0
 ( 8)  _Iyear_1998 = 0
 ( 9)  _Iyear_1999 = 0

       F(  9,    53) =    1.21
            Prob > F =    0.3094
```

We failed to reject the null that all years coefficients are jointly equal to zero therefore no time fixed-effects are needed.

Testing for random effects: Breusch-Pagan Lagrange multiplier (LM)

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 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  y  x1, re
xttest0
```

```
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

        y[country,t] = Xb + u[country] + e[country,t]

        Estimated results:
                         |       Var     sd = sqrt(Var)
                ---------+-----------------------------
                       y |   9.09e+18       3.02e+09
                       e |   7.82e+18       2.80e+09
                       u |   1.13e+18       1.06e+09

        Test:   Var(u) = 0
                             chi2(1) =     2.67
                         Prob > chi2 =     0.1023
```

Here we failed to reject the null and conclude that random effects is not appropriate. This is, no evidence of significant differences across countries, therefore you can run a simple OLS regression.

Testing for cross-sectional dependence/contemporaneous correlation: using Breusch-Pagan LM test of independence

According to Baltagi, cross-sectional dependence is a problem in macro panels with long time series (over 20-30 years). This is not much of a problem in micro panels (few years and large number of cases).

The null hypothesis in the B-P/LM test of independence is that residuals across entities are not correlated. The command to run this test is `xttest2` (run it after `xtreg, fe`):

```
xtreg y x1, fe
xttest2
```

```
. xttest2

Correlation matrix of residuals:

          __e1     __e2     __e3     __e4     __e5     __e6     __e7
__e1    1.0000
__e2    0.3615   1.0000
__e3    0.5391   0.4146   1.0000
__e4    0.3209   0.4660  -0.3015   1.0000
__e5   -0.2032  -0.3764  -0.3590  -0.3080   1.0000
__e6   -0.2572   0.2432  -0.0491  -0.1065   0.2321   1.0000
__e7    0.6403   0.0793   0.8206  -0.4358  -0.0818   0.0355   1.0000

Breusch-Pagan LM test of independence: chi2(21) =    28.914, Pr = 0.1161
Based on 10 complete observations over panel units
```

No cross-sectional dependence

Type `xttest2` for more info. If not available try installing it by typing `ssc install xttest2`

Testing for cross-sectional dependence/contemporaneous correlation: Using 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 y x1, fe
xtcsd, pesaran abs
```

```
. xtcsd, pesaran abs


Pesaran's test of cross sectional independence =     1.155, Pr = 0.2479

Average absolute value of the off-diagonal elements =     0.316
```

No cross-sectional dependence

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

A test for heteroskedasticiy is avalable for the fixed- effects model using the command `xttest3`.

This is a user-written program, to install it type `ssc install xtest3.`

```
xtreg y x1, fe
xttest3
```

```
. xttest3

Modified Wald test for groupwise heteroskedasticity
in fixed effect regression model

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

chi2 (7)  =        42.77
Prob>chi2 =       0.0000
```

Presence of heteroskedasticity

The null is homoskedasticity (or constant variance). Above we reject the null and conclude heteroskedasticity. Type `help xtest3` for more details.

NOTE: Use the option '`robust`' to control for heteroskedasticiy (in both fixed and random effects).

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`

`xtserial y x1`

```
. xtserial y x1

Wooldridge test for autocorrelation in panel data
H0: no first-order autocorrelation
    F(  1,       6) =      0.214
           Prob > F =      0.6603
```

No serial correlation

The null is no serial correlation. Above we fail to reject the null and conclude the data does not have first-order autocorrelation. Type `help xtserial` for more details.

Stata 11 has a series of unit root tests using the command `xtunitroot`, it included the following series of tests (type `help xtunitroot` for more info on how to run the tests):

"xtunitroot performs a variety of tests for unit roots (or stationarity) in panel datasets. The Levin-Lin-Chu (2002), Harris-Tzavalis (1999), Breitung (2000; Breitung and Das 2005), Im-Pesaran-Shin (2003), and Fisher-type (Choi 2001) tests have as the null hypothesis that all the panels contain a unit root. The Hadri (2000) Lagrange multiplier (LM) test has as the null hypothesis that all the panels are (trend) stationary. The top of the output for each test makes explicit the null and alternative hypotheses. Options allow you to include panel-specific means (fixed effects) and time trends in the model of the data-generating process" [Source: http://www.stata.com/help.cgi?xtunitroot or type `help xtunitroot`]

Stata 10 does not have this command but can run user-written programs to run the same tests. You will have to find them and install them in your Stata program (remember, these are only for Stata 9.2/10). To find the add-ons type:

```
findit panel unit root test
```

A window will pop-up, find the desired test, click on the blue link, then click where it says "(click here to install)"

For more info on unit roots please check: http://dss.princeton.edu/training/TS101.pdf

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

Summary of basic models (FE/RE)

Command	Syntax
	Entity fixed effects
xtreg	xtreg *y x1 x2 x3 x4 x5 x6 x7*, fe
areg	areg *y x1 x2 x3 x4 x5 x6 x7*, absorb(country)
regress	xi: regress *y x1 x2 x3 x4 x5 x6 x7 i.country*,
	Entity and time fixed effects
xi: xtreg	xi: xtreg *y x1 x2 x3 x4 x5 x6 x7 i.year*, fe
xi: areg	xi: areg *y x1 x2 x3 x4 x5 x6 x7 i.year*, absorb(country)
xi: regress	xi: regress *y x1 x2 x3 x4 x5 x6 x7 i.country i.year*
	Random effects
xtreg	xtreg *y x1 x2 x3 x4 x5 x6 x7*, re robust

NOTE: In Stata 11 you do not need "`xi:`" when adding dummy variables using `regress` or `areg`

Useful links / Recommended books / References

- DSS Online Training Section http://dss.princeton.edu/training/
- UCLA Resources http://www.ats.ucla.edu/stat/
- Princeton DSS Libguides http://libguides.princeton.edu/dss

Books/References

- "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
- *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, 2nd ed., Stata Press, 2009.
- *Designing Social Inquiry: Scientific Inference in Qualitative Research / Gary King, Robert O. Keohane, Sidney Verba, Princeton University Press*, 1994.
- *Econometric analysis* / William H. Greene. 6th ed., Upper Saddle River, N.J. : Prentice Hall, 2008.
- *Econometric Analysis of Panel Data*, Badi H. Baltagi, Wiley, 2008
- *Introduction to econometrics* / James H. Stock, Mark W. Watson. 2nd ed., Boston: Pearson Addison Wesley, 2007.
- *Statistical Analysis: an interdisciplinary introduction to univariate & multivariate methods /* Sam Kachigan, New York : Radius Press, c1986
- *Statistics with Stata (updated for version 9) /* Lawrence Hamilton, Thomson Books/Cole, 2006
- *Unifying Political Methodology: The Likelihood Theory of Statistical Inference* / Gary King, Cambridge University Press, 1989