



Getting Started in Panel Data Analysis

(ver. 2.1 *beta*)

Oscar Torres-Reyna
Data Consultant
otorres@princeton.edu



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.

One example of panel data



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 (when comparing countries or states within a country –i.e. Utah vs. New York) or difference in business practices across companies.

Panel data help also to control for unobservable variables that change over time but not across entities (i.e. national policies, federal regulations, international agreements, etc.)

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

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

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

- Fixed effects
- Random effects

The rationale behind *fixed effects* model is that you will control for differences across entities (countries, individuals, companies) but not necessarily over time. So the equation for the fixed effects model becomes:

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

Where

- α_i ($i=1\dots n$) 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).

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

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

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 equivalent:

“the slope coefficient on X is the same from one [entity] to the next. The [entity]-specific intercepts in [eq.1] and the binary repressors 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*:

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

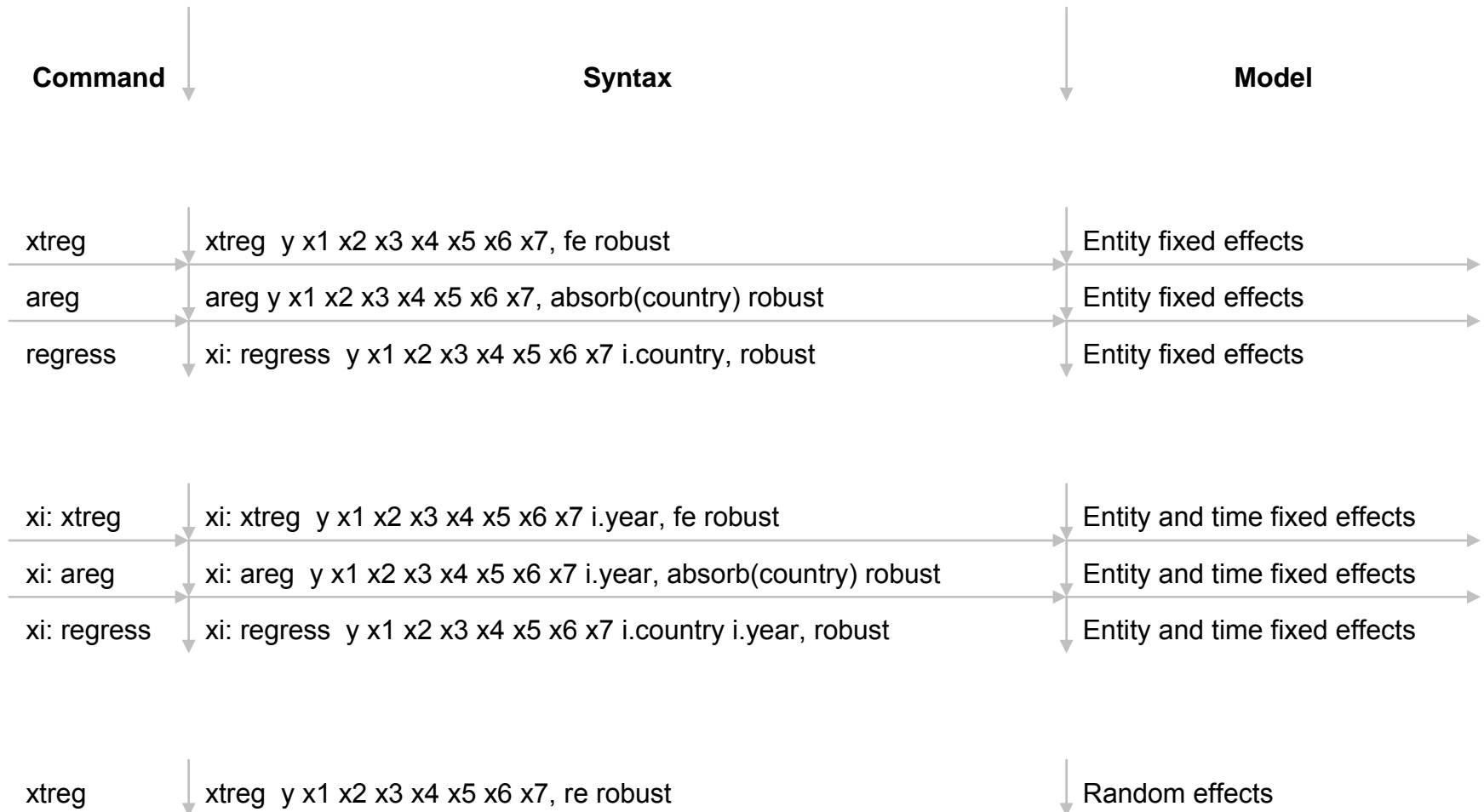
Where

- 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 .

The following pages will offer some examples on how to run this models in Stata. ***There are three ways*** to run fixed effects:

- Using `xtreg`
- Using `areg`
- Using `xi: reg`

Summary of models



To run a fixed effects model in Stata you use the command `xtreg`. But before that you need to tell Stata that you have panel data by using the command `xtset`. Before running `xtreg` 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 (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 and you can still run the model.

The output of the regression when using `xtreg` has the same reading as a the output produced by `regress` (o the regular regression).

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

```
. xtset country year
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

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

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

Dependent variable

Independent variable(s)

Fixed effects option

```
. xtreg y x1 x2 x3 x4 x5 x6 x7, fe robust
```

Robust standard errors (to control for heteroskedasticity)

Total number of cases (rows)

Fixed-effects (within) regression
Group variable: **country**

```
Number of obs      =      490
Number of groups   =       49
Obs per group: min =       10
                  avg  =      10.0
                  max  =       10
```

Total number of groups (entities)

```
R-sq:  within = 0.3018
       between = 0.0723
       overall = 0.1134
```

```
F(7, 434) = 10.68
Prob > F   = 0.0000
```

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.

```
corr(u_i, Xb) = -0.7512
```

(Std. Err. adjusted for clustering on country)

y	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
x1	-3.32e+08	5.73e+08	-0.58	0.562	-1.46e+09 7.94e+08
x2	1.88e+09	5.17e+08	3.62	0.000	8.58e+08 2.89e+09
x3	2.03e+09	6.70e+08	3.02	0.003	7.09e+08 3.34e+09
x4	4.83e+08	1.97e+08	2.45	0.015	9.58e+07 8.71e+08
x5	-6.45e+08	1.55e+08	-4.15	0.000	9.50e+08 -3.39e+08
x6	-4.60e+08	1.61e+08	-2.85	0.005	-7.77e+08 -1.43e+08
x7	1.22e+09	2.86e+08	4.27	0.000	6.60e+08 1.78e+09
_cons	1.76e+09	1.04e+08	16.89	0.000	1.56e+09 1.97e+09
sigma_u	2.883e+09				
sigma_e	2.312e+09				
rho	.60858827				(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)

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*.

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

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

Variance not explained by differences across entities. Also know as the intraclass correlation

$$\rho = \frac{(\sigma_u)^2}{(\sigma_u)^2 + (\sigma_e)^2}$$

σ_u = sd of common residuals u_i
 σ_e = sd of unique residuals e_i

Entity and time fixed effects (using xtreg)

```

Dependent variable: y
Independent variable(s): x1 x2 x3 x4 x5 x6 x7
Time effects: i.year, fe
Fixed effects option: robust
Robust standard errors (to control for heteroskedasticity)

. xi: xtreg y x1 x2 x3 x4 x5 x6 x7 i.year, fe robust
i.year          _lyear_1990-1999      (naturally coded; _lyear_1990 omitted)
    
```

Fixed-effects (within) regression
Group variable: **country**

R-sq: within = **0.3257**
 between = **0.0648**
 overall = **0.1174**

Number of obs = **490**
 Number of groups = **49**
 Obs per group: min = **10**
 avg = **10.0**
 max = **10**

F(16, 425) = **6.47**
 Prob > F = **0.0000**

corr(u_i, Xb) = **-0.7501**

(Std. Err. adjusted for clustering on country)

y	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
x1	-5.10e+08	5.80e+08	-0.88	0.380	-1.65e+09 6.30e+08
x2	1.91e+09	5.03e+08	3.81	0.000	9.26e+08 2.90e+09
x3	2.00e+09	6.91e+08	2.90	0.004	6.45e+08 3.36e+09
x4	3.99e+08	2.11e+08	1.89	0.059	-1.54e+07 8.13e+08
x5	-6.30e+08	1.55e+08	-4.08	0.000	-9.34e+08 -3.27e+08
x6	-4.78e+08	1.67e+08	-2.86	0.004	-8.07e+08 -1.49e+08
x7	1.20e+09	2.85e+08	4.22	0.000	6.41e+08 1.76e+09
_lyear_1991	-2.85e+08	4.84e+08	-0.59	0.557	-1.24e+09 6.67e+08
_lyear_1992	1.14e+08	4.41e+08	0.26	0.796	-7.53e+08 9.81e+08
_lyear_1993	1.60e+08	4.91e+08	0.33	0.745	-8.06e+08 1.13e+09
_lyear_1994	1.07e+09	4.26e+08	2.51	0.012	2.34e+08 1.91e+09
_lyear_1995	7.01e+08	4.84e+08	1.45	0.148	-2.50e+08 1.65e+09
_lyear_1996	7.41e+08	4.67e+08	1.59	0.113	-1.76e+08 1.66e+09
_lyear_1997	8.36e+08	4.50e+08	1.86	0.064	-4.88e+07 1.72e+09
_lyear_1998	4.60e+08	5.54e+08	0.83	0.407	-6.30e+08 1.55e+09
_lyear_1999	6.83e+08	4.66e+08	1.47	0.143	-2.33e+08 1.60e+09
_cons	1.32e+09	3.54e+08	3.72	0.000	6.20e+08 2.01e+09
sigma_u	2.928e+09				
sigma_e	2.296e+09				
rho	.61914132				(fraction of variance due to u_i)

Total number of cases (rows)

Total number of groups (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.

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

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)

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.

Fixed effects: n entity-specific intercepts (using areg)

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

Dependent variable

Independent variable(s)

Hide the binary variables for each entity

Robust standard errors (to control for heteroskedasticity)

```
. areg y x1 x2 x3 x4 x5 x6 x7, absorb(country) robust
```

Linear regression, absorbing indicators

```
Number of obs = 490
F( 7, 434) = 10.68
Prob > F = 0.0000
R-squared = 0.4931
Adj R-squared = 0.4289
Root MSE = 2.3e+09
```

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. This provides a more honest association between X and Y.

y	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
x1	-3.32e+08	5.73e+08	-0.58	0.562	-1.46e+09 7.94e+08
x2	1.88e+09	5.17e+08	3.62	0.000	8.58e+08 2.89e+09
x3	2.03e+09	6.70e+08	3.02	0.003	7.09e+08 3.34e+09
x4	4.83e+08	1.97e+08	2.45	0.015	9.58e+07 8.71e+08
x5	-6.45e+08	1.55e+08	-4.15	0.000	-9.50e+08 -3.39e+08
x6	-4.60e+08	1.61e+08	-2.85	0.005	-7.77e+08 -1.43e+08
x7	1.22e+09	2.86e+08	4.27	0.000	6.60e+08 1.78e+09
_cons	1.76e+09	1.04e+08	16.89	0.000	1.56e+09 1.97e+09
country	absorbed				(49 categories)

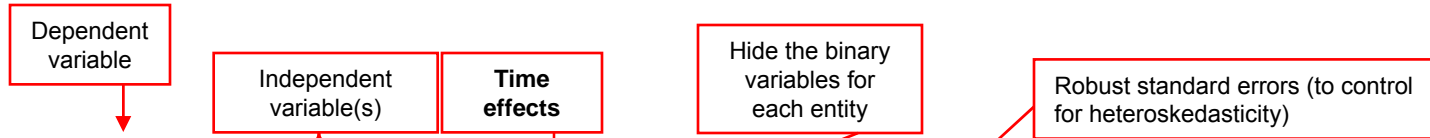
Coefficients of the regressors. Indicate how much Y changes when X increases by one unit.

"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 for our particular Stata configuration." (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)

Entity and time fixed effects (using areg)



```
. xi: areg y x1 x2 x3 x4 x5 x6 x7 i.year, absorb(country) robust
i.year _l year_1990-1999 (naturally coded; _l year_1990 omitted)
```

Linear regression, absorbing indicators

Number of obs = 490
 F(16, 425) = 6.47
 Prob > F = 0.0000
 R-squared = 0.5105
 Adj R-squared = 0.4368
 Root MSE = 2.3e+09

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. This provides a more honest association between X and Y.

y	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
x1	-5.10e+08	5.80e+08	-0.88	0.380	-1.65e+09	6.30e+08
x2	1.91e+09	5.03e+08	3.81	0.000	9.26e+08	2.90e+09
x3	2.00e+09	6.91e+08	2.90	0.004	6.45e+08	3.36e+09
x4	3.99e+08	2.11e+08	1.89	0.059	-1.54e+07	8.13e+08
x5	-6.30e+08	1.55e+08	-4.08	0.000	-9.34e+08	-3.27e+08
x6	-4.78e+08	1.67e+08	-2.86	0.004	-8.07e+08	-1.49e+08
x7	1.20e+09	2.85e+08	4.22	0.000	6.41e+08	1.76e+09
_l year_1991	-2.85e+08	4.84e+08	-0.59	0.557	-1.24e+09	6.67e+08
_l year_1992	1.14e+08	4.41e+08	0.26	0.796	-7.53e+08	9.81e+08
_l year_1993	1.60e+08	4.91e+08	0.33	0.745	-8.06e+08	1.13e+09
_l year_1994	1.07e+09	4.26e+08	2.51	0.012	2.34e+08	1.91e+09
_l year_1995	7.01e+08	4.84e+08	1.45	0.148	-2.50e+08	1.65e+09
_l year_1996	7.41e+08	4.67e+08	1.59	0.113	-1.76e+08	1.66e+09
_l year_1997	8.36e+08	4.50e+08	1.86	0.064	-4.88e+07	1.72e+09
_l year_1998	4.60e+08	5.54e+08	0.83	0.407	-6.30e+08	1.55e+09
_l year_1999	6.83e+08	4.66e+08	1.47	0.143	-2.33e+08	1.60e+09
_cons	1.32e+09	3.54e+08	3.72	0.000	6.20e+08	2.01e+09
country	absorbed				(49 categories)	

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

“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 for our particular Stata configuration.” (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)

Fixed effects: common intercept and n-1 binary regressors (using dummies and regress)

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

Dependent variable

Independent variable(s)

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

Robust standard errors (to control for heteroskedasticity)

```
. xi: regress y x1 x2 x3 x4 x5 x6 x7 i.country, robust
i.country          _lcountry_1-49      (naturally coded; _lcountry_1 omitted)
```

Linear regression

Number of obs = 490
 F(55, 434) = 7.60
 Prob > F = 0.0000
 R-squared = 0.4931
 Root MSE = 2.3e+09

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

	y	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
	x1	-3.32e+08	5.73e+08	-0.58	0.562	-1.46e+09 7.94e+08
	x2	1.88e+09	5.17e+08	3.62	0.000	8.58e+08 2.89e+09
	x3	2.03e+09	6.70e+08	3.02	0.003	7.09e+08 3.34e+09
	x4	4.83e+08	1.97e+08	2.45	0.015	9.58e+07 8.71e+08
	x5	-6.45e+08	1.55e+08	-4.15	0.000	-9.50e+08 -3.39e+08
	x6	-4.60e+08	1.61e+08	-2.85	0.005	-7.77e+08 -1.43e+08
	x7	1.22e+09	2.86e+08	4.27	0.000	6.60e+08 1.78e+09
	_lcountry_2	2.97e+09	1.33e+09	2.24	0.026	3.65e+08 5.57e+09
	_lcountry_3	-1.16e+09	1.07e+09	-1.08	0.280	-3.26e+09 9.45e+08
	_lcountry_4	-2.71e+08	1.52e+09	-0.18	0.859	-3.26e+09 2.72e+09
	_lcountry_5	-6.15e+09	1.94e+09	-3.16	0.002	-9.97e+09 -2.33e+09
	_lcountry_6	2.63e+09	1.81e+09	1.45	0.147	-9.28e+08 6.19e+09
	_lcountry_7	-4.40e+09	2.72e+09	-1.62	0.107	-9.76e+09 9.50e+08

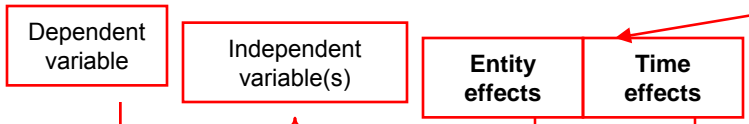
Coefficients of the regressors indicate how much Y changes when X increases by one unit.

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)

Entity and time fixed effects (using regress)

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



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

```
. xi: regress y x1 x2 x3 x4 x5 x6 x7 i.country i.year, robust
i.country      _lcountry_1-49 (naturally coded; _lcountry_1 omitted)
i.year         _lyear_1990-1999 (naturally coded; _lyear_1990 omitted)

Linear regression on
```

Robust standard errors (to control for heteroskedasticity)

```
Number of obs =      490
F( 64,  425) =      6.91
Prob > F       =      0.0000
R-squared      =      0.5105
Root MSE      =      2.3e+09
```

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.

	y	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
x1		-5.10e+08	5.80e+08	-0.88	0.380	-1.65e+09 6.30e+08
x2		1.91e+09	5.03e+08	3.81	0.000	9.26e+08 2.90e+09
x3		2.00e+09	6.91e+08	2.90	0.004	6.45e+08 3.36e+09
x4		3.99e+08	2.11e+08	1.89	0.059	-1.54e+07 8.13e+08
x5		-6.30e+08	1.55e+08	-4.08	0.000	-9.34e+08 -3.27e+08
x6		-4.78e+08	1.67e+08	-2.86	0.004	-8.07e+08 -1.49e+08
x7		1.20e+09	2.85e+08	4.22	0.000	6.41e+08 1.76e+09
_lcountry_2		2.94e+09	1.28e+09	2.31	0.022	4.35e+08 5.45e+09
_lcountry_3		-1.21e+09	1.05e+09	-1.15	0.251	-3.27e+09 8.56e+08
_lcountry_4		-4.87e+08	1.48e+09	-0.33	0.741	-3.39e+09 2.41e+09
_lcountry_5		-6.30e+09	1.99e+09	-3.17	0.002	-1.02e+10 -2.39e+09
_lcountry_6		2.42e+09	1.81e+09	1.34	0.182	-1.14e+09 5.97e+09
_lcountry_7		-5.02e+09	2.81e+09	-1.79	0.075	-1.05e+10 5.00e+08

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

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)

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 x4 x5 x6 x7, fe robust
estimates store fixed
xi: regress y x1 x2 x3 x4 x5 x6 x7 i.country, robust
estimates store ols
areg y x1 x2 x3 x4 x5 x6 x7, absorb(country) robust
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)
```

All three commands provide the same results

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

Variable	fixed	ols	areg
x1	-3.323e+08	-3.323e+08	-3.323e+08
x2	1.875e+09***	1.875e+09***	1.875e+09***
x3	2.027e+09**	2.027e+09**	2.027e+09**
x4	4.834e+08*	4.834e+08*	4.834e+08*
x5	-6.445e+08***	-6.445e+08***	-6.445e+08***
x6	-4.599e+08**	-4.599e+08**	-4.599e+08**
x7	1.222e+09***	1.222e+09***	1.222e+09***
_l country_2		2.969e+09*	
_l country_3		-1.157e+09	
_l country_4		-2.710e+08	
_l country_5		-6.149e+09**	
_l country_6		2.631e+09	
_l country_7		-4.403e+09	

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 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:

$$Y_{it} = \beta X_{it} + \alpha + u_{it} + \varepsilon_{it} \quad [\text{eq.4}]$$

Random effects

In Stata you can estimate a random effects model using `xtreg` and the option `re`. The reading of the output is the same as that for fixed effects but notice the difference in coefficients.

```

    . xtreg y x1 x2 x3 x4 x5 x6 x7, re robust
  
```

Annotations:
 - **Dependent variable**: y
 - **Independent variable(s)**: x1 x2 x3 x4 x5 x6 x7
 - **Random effects option**: re
 - **Robust standard errors (to control for heteroskedasticity)**: robust

```

Random-effects GLS regression
Group variable: country

R-sq:  within = 0.2319
       between = 0.2461
       overall = 0.2321

Number of obs   = 490
Number of groups = 49
Obs per group:  min = 10
                avg  = 10.0
                max  = 10

Random effects u_i ~ Gaussian
corr(u_i, X)      = 0 (assumed)

Wald chi2(8)    = 294.61
Prob > chi2     = 0.0000
  
```

Differences across units are uncorrelated with the regressors

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.

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)

(Std. Err. adjusted for clustering on country)

	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
y						
x1	1.77e+08	2.04e+08	0.87	0.386	-2.22e+08	5.76e+08
x2	4.48e+08	1.65e+08	2.72	0.007	1.25e+08	7.71e+08
x3	5.14e+08	1.70e+08	3.03	0.002	1.81e+08	8.46e+08
x4	3.97e+08	1.12e+08	3.53	0.000	1.76e+08	6.17e+08
x5	-5.87e+08	1.33e+08	-4.40	0.000	-8.48e+08	-3.26e+08
x6	-1.34e+08	1.10e+08	-1.22	0.222	-3.49e+08	8.09e+07
x7	1.16e+09	4.20e+08	2.75	0.006	3.33e+08	1.98e+09
_cons	1.76e+09	1.62e+08	10.90	0.000	1.45e+09	2.08e+09
sigma_u	7.482e+08					
sigma_e	2.312e+09					
rho	.09476001	(fraction of variance due to u_i)				

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).

To do this, 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 x2 x3 x4 x5 x6 x7, fe`
- `estimates store fixed`
- `xtreg y x1 x2 x3 x4 x5 x6 x7, re`
- `estimates store random`
- `hausman fixed random`

```
. hausman fixed random
```

	Coefficients		(b-B) Difference	sqrt(diag(V_b-V_B)) S. E.
	(b) fixed	(B) random		
x1	-3.32e+08	1.77e+08	-5.09e+08	3.30e+08
x2	1.88e+09	4.48e+08	1.43e+09	3.16e+08
x3	2.03e+09	5.14e+08	1.51e+09	4.12e+08
x4	4.83e+08	3.97e+08	8.65e+07	1.46e+08
x5	-6.45e+08	-5.87e+08	-5.76e+07	6.11e+07
x6	-4.60e+08	-1.34e+08	-3.26e+08	4.76e+07
x7	1.22e+09	1.16e+09	6.60e+07	.

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(7) = (b-B)' [(V_b-V_B)^(-1)] (b-B)
 = 44.91
 Prob>chi2 = 0.0000
 (V_b-V_B is not positive definite)

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

If you get this message then the test may not be conclusive

Random and fixed are not that much different if you have panels with lots of years. Random is usually preferred when you have large number of entities.

Moving average for panel data

Use the command `egenmore`, you may have to install it first by typing

```
scc install egenmore
```

For the lags to work you may need to `xtset` your data by typing

```
xtset [name of panel variable] [time variable]
```

For example:

```
xtset country year
```

For a three year moving average type

```
egen moveave = filter(x1), lags(0/3) normalise
```

Where `x1` is the variable of interest.

Type `help egenmore` for more details.

Source: <http://www.stata.com/support/faqs/stat/moving.html>

Useful links / Recommended books

- DSS Online Training Section <http://dss.princeton.edu/training/>
- UCLA Resources to learn and use STATA <http://www.ats.ucla.edu/stat/stata/>
- DSS help-sheets for STATA http://dss/online_help/stats_packages/stata/stata.htm
- *Introduction to Stata* (PDF), Christopher F. Baum, Boston College, USA. “A 67-page description of Stata, its key features and benefits, and other useful information.”
<http://fmwww.bc.edu/GStat/docs/StataIntro.pdf>
- STATA FAQ website <http://stata.com/support/faqs/>
- Princeton DSS Libguides <http://libguides.princeton.edu/dss>

Books

- *Introduction to econometrics* / James H. Stock, Mark W. Watson. 2nd ed., Boston: Pearson Addison Wesley, 2007.
- *Data analysis using regression and multilevel/hierarchical models* / Andrew Gelman, Jennifer Hill. Cambridge ; New York : Cambridge University Press, 2007.
- *Econometric analysis* / William H. Greene. 6th ed., Upper Saddle River, N.J. : Prentice Hall, 2008.
- *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
- *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