10 Panel Data

Andrius Buteikis, andrius.buteikis@mif.vu.lt
http://web.vu.lt/mif/a.buteikis/

Introduction

Panel data combines **cross-sectional** and **time series** data: the *same* individuals (persons, firms, cities, etc.) are observed at several points in time (days, years, before and after treatment etc.).

Panel data allows you to control for variables you cannot observe or measure like:

- cultural (like country or region specific) factors;
- difference in business practices across companies;
- etc.

or variables that change over time but not across entities:

- national policies;
- federal regulations;
- international agreements;
- etc.

This is, it accounts for individual *heterogeneity* (individual effects).

```
##   country year       Y      X1     X2       X3
##        C1 2000  8.8790 4.1087 3.7014  1.1534
##        C1 2001  9.5396 7.4482 2.5272 -0.1381
##        C1 2002 13.1174 5.7196 1.9322  2.2538
##        C2 2000 10.1410 5.8015 2.7820  1.4265
##        C2 2001 10.2586 5.2214 1.9740  0.7049
##        C2 2002 13.4301 3.8883 2.2711  1.8951
##        C3 2000 10.9218 8.5738 2.3750  1.8781
##        C3 2001  7.4699 5.9957 1.3133  1.8216
##        C3 2002  8.6263 1.0668 3.8378  1.6886
```

The above data presentation is termed a stacked time series (one time series is stacked above another). If one cross-section is above another, this is called a **stacked cross section**. If we remove the individual and year attributes and do not make any distinction between cross section and time series, this is called a **pooled data** organization.

The Econometric Model

Consider the **multiple linear regression model** for individual $i = 1, ..., N$ who is observed at several time periods $t = 1, ..., T$:

$$y_{it} = \alpha + x'_{it}\beta + z'_i\gamma + c_i + u_{it}$$

where:

- y_{it} - the dependent variable;
- x_{it} - a $K - dimensional$ row vector of *time-varying* explanatory variables;
- z_i - an $M - dimensional$ row vector of *time-invariant* explanatory variables, excluding the constant;
- α - the intercept;
- β - a $K - dimensional$ column vector of parameters;
- γ - a $M - dimensional$ column vector of parameters;
- c_i - an **individual-specific** effect;
- u_{it} - a *idiosyncratic* error term (observation-specific zero-mean random-error term, analogous to the random-error term of cross-sectional regression analysis).

If each individual in the data set is observed the same number of times, the data set is a **balanced panel**. An **unbalanced panel** data set is one in which individuals may be observed different numbers of times. Some functions are operational only for balanced data.

The T observations for an individual i can be summarized as follows:

$$Y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{bmatrix}_{T\times 1}, \quad X_i = \begin{bmatrix} x_{i1}' \\ \vdots \\ x_{it}' \\ \vdots \\ x_{iT}' \end{bmatrix}_{T\times K}, \quad Z_i = \begin{bmatrix} z_i' \\ \vdots \\ z_i' \\ \vdots \\ z_i' \end{bmatrix}_{T\times M}, \quad u_{it} = \begin{bmatrix} u_i \\ \vdots \\ u_{it} \\ \vdots \\ u_{iT} \end{bmatrix}_{T\times 1}$$

The NT observations for *all* individuals and *all* time periods can be summarized as follows:

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_N \end{bmatrix}_{NT\times 1}, \quad X = \begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{bmatrix}_{NT\times K}, \quad Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_i \\ \vdots \\ Z_N \end{bmatrix}_{NT\times M}, \quad u = \begin{bmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_N \end{bmatrix}_{NT\times 1}$$

The data generation process (DGP) is described by:

1. *PL1: Linearity*

$$y_{it} = \alpha + x_{it}'\beta + z_i'\gamma + c_i + u_{it}, \text{ where } \mathbb{E}[u_{it}] = 0, \text{ and } \mathbb{E}[c_i] = 0$$

   i.e. the model is linear in parameters α, β, γ, effect c_i and error u_{it}.

2. *PL2: Independence*

$$\{X_i, z_i, Y_i\}_{i=1}^N, \text{ i.i.d. (independent and identcally distributed)}$$

   The observations are independent across individuals but not necessarily across time. This is guaranteed by random sampling of individuals.

3. *PL3: Strict Exogeneity*

$$\mathbb{E}[u_{it}|X_i, z_i, c_i] = 0, \text{ (mean independent)}$$

The idiosyncratic error term u_{it} is assumed uncorrelated with the explanatory variables of all past, current and future time periods of the same individual. This is a strong assumption which e.g. **rules out lagged dependent variables**. It also assumes that the idiosyncratic error is uncorrelated with the individual specific effect.

4. *PL4: Error Variance*
   - 4.1 $Var[u_i|X_i, z_i, c_i] = \sigma_u^2 I, \quad \sigma_u^2 > 0$ and finite (i.e. homoscedastic and no serial correlation);
   - 4.2 $Var[u_{it}|X_i, z_i, c_i] = \sigma_{u,it}^2 I, \quad \sigma_{u,it}^2 > 0$ and finite and $Cov[u_{it}, u_{is}|X_i, z_i, c_i] = 0, \quad \forall s \neq t$ (i.e. no serial correlation);
   - 4.3 $Var[u_i|X_i, z_i, c_i] = \Omega_{u,i}$ is positive-defined and finite.

The remaining assumptions are divided into two sets of assumptions:

- The random effects model;
- The fixed effects model.

The Random Effects Model

In the random effects model, the individual-specific effect, c_i is a random variable, that is uncorrelated with the explanatory variable.

1. *RE1: Unrelated effects*

$$\mathbb{E}[c_i|X_i, z_i] = 0$$

RE1 assumes that the individual-specific effect is a random variable that is uncorrelated with the explanatory variables of all past, current and future time periods of the *same* individual.

2. *RE2: Effect Variance*
   - 2.1 $Var[c_i|X_i, z_i] = \sigma_c^2 < \infty$ (homoscedastic, assumes constant variance of the individual specific effect);
   - 2.2 $Var[c_i|X_i, z_i] = \sigma_{c,i}^2 < \infty$ (heteroscedastic).

RE2.1 assumes constant variance of the individual specific effect.

3. *RE3: Indentifiability*
   - 3.1 $rank(W) = K + M + 1 < NT$ and $\mathbb{E}[W_i W_i'] = Q_{WW}$ is positive-defined and finite. The typical element $w_{it}' = [1 \quad x_{it}' \quad z_i']$.
   - 3.2 $rank(W) = K + M + 1 < NT$ and $\mathbb{E}[W_i \Omega_{v,u}^{-1} W_i'] = Q_{WW}$ is positive defined and finite.

RE3 assumes that the regressors including a constant are not perfectly collinear, that all regressors (but the constant) have non-zero variance and not too many extreme values.

The random effects model can be rewritten as:

$$y_{it} = \alpha + x_{it}'\beta + z_i'\gamma + v_{it}$$

where $v_{it} = c_i + u_{it}$. Assuming *PL2, PL4 and RE1*. In special versions of *PL*4.1 and *RE*2.1 leads to:

$$\Omega_v = Var[v|X,Z] = \begin{pmatrix} \Omega_{v,1} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & & & \vdots \\ 0 & & \Omega_{v,i} & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \dots & 0 & \dots & \Omega_{v,N} \end{pmatrix}_{NT \times NT}$$

with element:

$$\Omega_{v,i} = Var[v_i|X_i,z_i] = \begin{pmatrix} \sigma_v^2 & \sigma_c^2 & \dots & \sigma_c^2 \\ \sigma_c^2 & \sigma_v^2 & \dots & \sigma_c^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_c^2 & \sigma_c^2 & \dots & \sigma_v^2 \end{pmatrix}_{T \times T}$$

where $\sigma_v^2 = \sigma_c^2 + \sigma_u^2$.

The Random Effects approach takes c_i to be a group-specific random term (similar to u_{it}) except that for each group, there is but a single draw that enters the regression identically in each period.

The Fixed Effects Model

In the fixed model, the individual-specific effect is a variable that is allowed to be correlated with the explanatory variables.

1. *FE1: Related effects*
$$\mathbb{E}[c_i|X_i, z_i] \neq 0$$
Explicitly states the absence of the unrelatedness assumption in *RE1*.
2. *RE2: Effect Variance*
Because of assumption *FE1*, *FE2* states the absence of the assumption in *RE2*.
3. *RE3: Indentifiability*
$rank(\ddot{X}) = K < NT$ and $\mathbb{E}[\ddot{x}_i' \ddot{x}_i]$ is positive-defined and finite, where the typical element $\ddot{x}_{it} = x_{it} - \bar{x}_i$ and $\bar{x}_i = 1/T \sum_t x_{it}$.

FE3 assumes that the time-varying explanatory variables are not perfectly collinear, that they have non-zero within-variance (i.e. variation over time for a given individual) and not too many extreme values. Hence, X_{it} cannot include a constant or any time-invariant variables. Note that only the parameters β but neither α, nor γ are identifiable in the fixed effects model.

The Fixed Effects approach takes c_i to be a group-specific constant term.

Autoregressive Panel Models (of order 1)

Panel data models are after all regression models, therefore we can analyze dynamic regression models as well, for example:

$$y_{it} = \alpha + y'_{i,t-1}\delta + x'_{it}\beta + z'_i\gamma + c_i + u_{it}$$

However, some specific problems arise when a lag of the dependent variable is included among the regressors in a panel model (this is connected with the fact that y_{it} is found to be correlated with the error term. In this case, the OLS estimators of the coefficients will be biased and inconsistent).

- One strategy for handling this problem is to take first differences for sweeping out the group effects:

$$\Delta y_{it} = \delta \Delta y_{i,t-1} + \Delta x_{it}' \beta + (u_{it} - u_{i,t-1})$$

  To remove the still existing correlation between $\Delta y_{i,t-1}$ and $u_{it} - u_{i,t-1}$, it is suggested to use an instrument for $\Delta y_{i,t-1}$ (it can be $y_{i,t-2}$ or $\Delta y_{i,t-2}$).
- Another, is to use the DPD (Dynamic Panel Data) approach.
  - The so-called Arellano-Bond estimator is based on the notion that the instrumental variables approach noted above does not exploit all of the information available in the sample. By doing so in a Generalized Method of Moments (GMM) context, we may construct more efficient estimates of the dynamic panel data model.
  - The GMM estimator which was suggested by Arellano and Bond is known to be rather inefficient when instruments are weak because making use of the information contained in differences only. Blundell and Bond suggest making use of additional level information besides the differences. The combination of moment restrictions for differences and levels results in an estimator which was called GMM-system-estimator by Arrellano and Bond (or The Blundell-Bond estimator).

Pooled OLS (POLS) Estimation

The *pooled OLS estimator* ignores the panel structure of the data and simply estimates α, β and γ as:

$$\begin{pmatrix} \hat{\alpha}_{POLS} \\ \hat{\beta}_{POLS} \\ \hat{\gamma}_{POLS} \end{pmatrix} = (W'W)^{-1}W'Y$$

where $W = [\mathbf{1}_{NT\times 1} \quad X \quad Z]$.

- *Random Effects Model*:
  The pooled OLS estimator of α, β and γ is unbiased under the assumptions *PL1*, *PL2*, *PL3* and *RE3* in small samples.However, the pooled OLS estimators is not efficient. More importantly, the usual standard errors of the pooled OLS estimator are incorrect and tests ($t-test$, $F-test$, $z-test$, $Wald-test$) based on them are not valid.
- *Fixed Effects Model*:
  The pooled OLS estimator of α, β and γ are biased and inconsistent, because the variable c_i is omitted and potentially correlated with the other regressors.

Random Effects (RE) Estimation

The random effects estimator is the feasible generalized least squares (GLS) estimator:

$$\begin{pmatrix} \hat{\alpha}_{RE} \\ \hat{\beta}_{RE} \\ \hat{\gamma}_{RE} \end{pmatrix} = (W'\widehat{\Omega}_v^{-1}W)^{-1}W'\widehat{\Omega}_v^{-1}Y$$

where $W = [\mathbf{1}_{NT\times 1} \quad X \quad Z]$.

- *Random Effects Model*:
  We cannot establish small sample properties for the RE estimator. The RE estimator is consistent and asymptotically normally distributed when $N \to \infty$ if T is fixed.
- *Fixed Effects Model*:
  Under the assumptions of the fixed effects model (*FE1*, i.e. *RE1* is violated), the random effects estimator of α, β and γ are biased and inconsistent, because the variable c_i is omitted and potentially correlated with the other regressors.

Fixed Effects (FE) Estimation

Subtracting time averages $\bar{y}_i = 1/T \sum_t y_{it}$ from the initial model:

$$y_{it} = \alpha + x'_{it}\beta + z'_i\gamma + c_i + u_{it}$$

yields the **within model**:

$$\ddot{y}_{it} = \ddot{x}'_{it}\beta + \ddot{u}_{it}$$

where $\ddot{y}_{it} = y_{it} - \bar{y}_i$, $\ddot{x}_{itk} - x_{itk} - \bar{x}_{ik}$ and $\ddot{u}_{it} = u_{it} - \bar{u}_i$. Note that the individual-specific effect c_i, the intercept α and the time-invariant regressors z_i cancel.

The *fixed effects estimator* or the *within estimator* of the slope coefficient β estimates the within model by OLS:

$$\widehat{\beta}_{FE} = \left(\ddot{X}'\ddot{X}\right)^{-1} \ddot{X}\ddot{Y}$$

Note that the parameters α and γ are **not estimated** by the within estimator.

First Differences (FD) Estimator

Subtracting the lagged value $y_{i,t-1}$ from the initial model:

$$y_{it} = \alpha + x'_{it}\beta + z'_i\gamma + c_i + u_{it}$$

yields the *first-difference model*:

$$\dot{y}_{it} = \dot{x}'_{it}\beta + \dot{u}_{it}$$

where $\dot{y}_{it} = y_{it} - y_{i,t-1}$, $\dot{x}_{it} = x_{it} - x_{i,t-1}$ and $\dot{u}_{it} = u_{it} - u_{i,t-1}$. Note that the individual-specific effect c_i, the intercept α and the time-invariant regressors z_i cancel. The *first-difference estimator* (FD) of the slope coefficient β estimates the first-difference model by OLS:

$$\widehat{\beta}_{FD} = \left(\dot{X}'\dot{X}\right)^{-1}\dot{X}'\dot{Y}$$

Note that the parameters α and γ are not estimated by the FD estimator.

In the special case, $T = 2$, the FD estimator is numerically identical to the FE estimator.

The FD estimator is a consistent estimator of β under the same assumptions as the FE estimator. it is less efficient than the FE estimator if u_{it} is not serially correlated (PL4.1).

Random Effects vs. Fixed Effects Estimation

The random effects model can be consistently estimated by both the RE estimator or the FE estimator. We would prefer the RE estimator if we can be sure that the individual-specific effect really is an unrelated effect (*RE1*)

The Hausman test can be used to differentiate between fixed effects model and random effects model in panel data. In this case, Random effects (RE) is preferred under the null hypothesis due to higher efficiency (i.e. has the smallest asymptotic variance, at least compared to the FE estimator), while under the alternative Fixed effects (FE) is at least consistent and thus preferred:

	H_0 is true	H_1 is true
$\widehat{\beta}_{RE}$	Consistent Efficient	Inconsistent
$\widehat{\beta}_{FE}$	Consistent Inefficient	Consistent

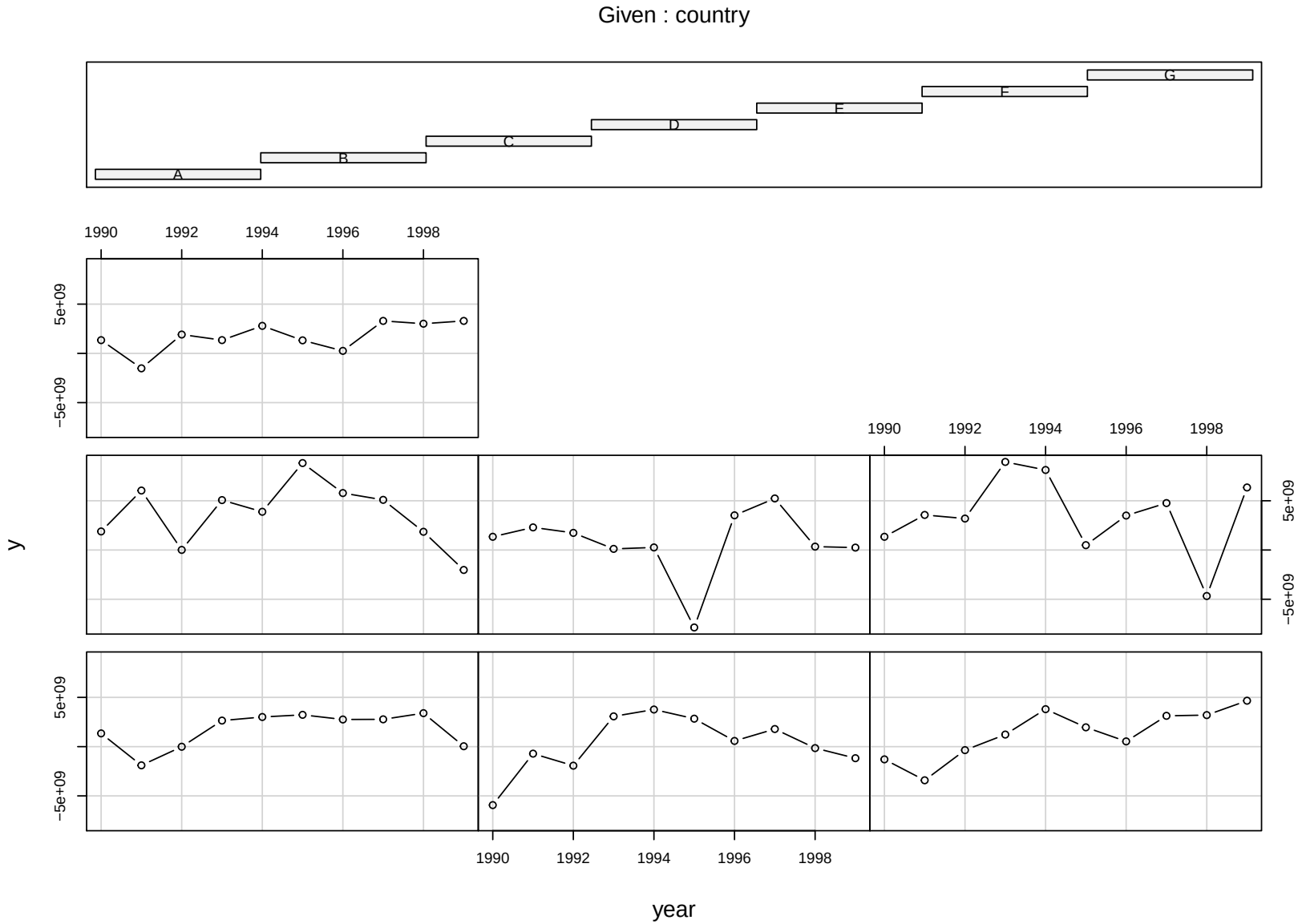
Example

Example panel data in a stacked cross-section form, where y - is the output and $x1$ is the predictive variable. This is a balanced panel example:

```
suppressPackageStartupMessages({
  library("readxl")
  library("plm")
  library("gplots")
})
txt1 <- "http://web.vu.lt/mif/a.buteikis/wp-content/"
txt2 <- "uploads/2018/05/pp.xlsx"
tmp = tempfile(fileext = ".xlsx")
download.file(url = paste0(txt1, txt2),
              destfile = tmp, mode = "wb")
ex.dt <- data.frame(read_excel(path = tmp))
```

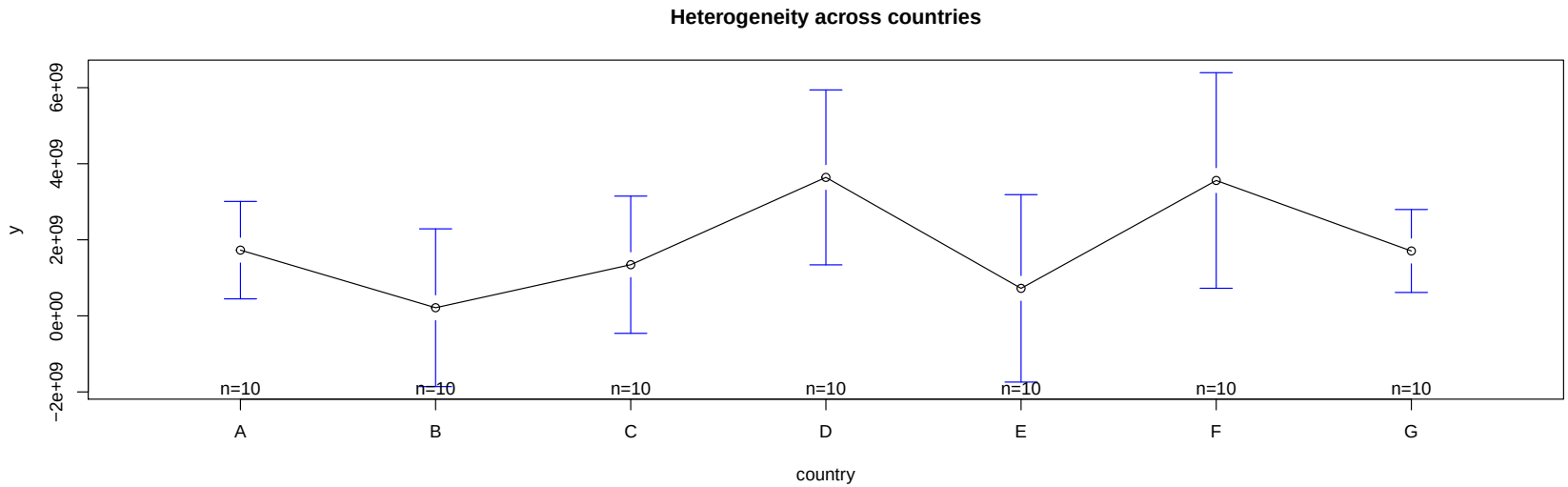
We can explore our data (left-most graph in the bottom line is for country A, next to the right is for B etc.):

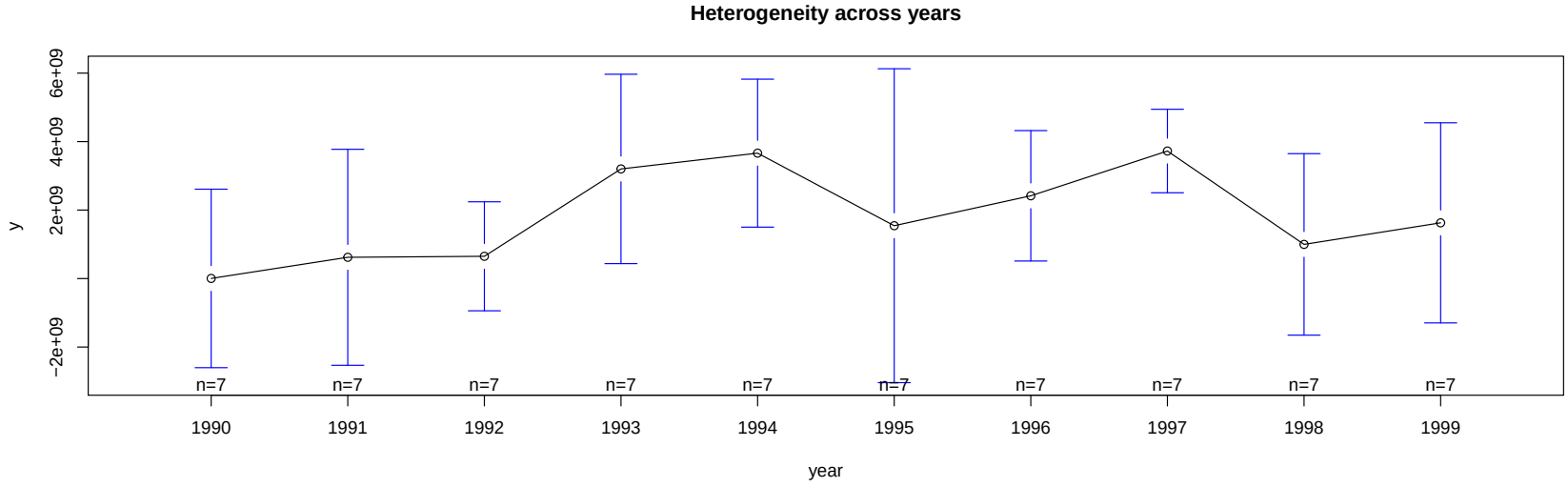
```
coplot(y ~ year | country, type = "b", data = ex.dt)
```



```
par(mfrow = c(1,2))
plotmeans(y ~ country, main="Heterogeneity across countries",
          data = ex.dt)
plotmeans(y ~ year, main="Heterogeneity across years",
          data = ex.dt)
```

In the plots below: The country effect on the mean of y (left) and the *year* (progress) effect on y (right) (95% confidence interval around the means is included).





The main purpose of the panel data analysis is to quantify the `x1` effect on `y`.

Some modelling using 'lm'

We start either with the pooled model:

$$y_{it} = \alpha + \beta x1_{it} + \epsilon_{it}, \quad i = 1, ..., N, \quad t = 1, ..., T$$

```
pooled = lm(y ~ x1, data = ex.dt)
round(summary(pooled)$coef, 4)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1524319070  621072624  2.4543   0.0167
## x1           494988914  778861261  0.6355   0.5272
```

or with OLS models, restricted to *individual* countries. For example:

```
countryB=lm(y ~ x1, data = ex.dt[ex.dt$country=="B",])
round(summary(countryB)$coef, 4)
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3453776646 1059056889 -3.2612   0.0115
## x1           7139367737 1755562974  4.0667   0.0036
```

note the difference in the coefficients.

In the Pooled OLS case, we assume the same coefficients **across all countries** whereas in the second case, we estimate a model on the data from country B only.

Both approaches have some drawbacks - the pooled model does not take into account heterogeneity across countries while individual model are based on small number of observations and do not consider common features of the countries (all they interact and experience the same influence of the progress). One possibility to take this common environment into account is to use the fixed effects (FE) model. To allow for the country effect, we introduce dummy variables $D_i =$ `factor(country)`:

$$y_{it} = \alpha + \beta x1_{it} + \sum_{i=1}^{N} \nu_i D_i + \epsilon_{it} = \begin{cases} \alpha + \beta x1_{1t} + \nu_1 + \epsilon_{1t}, & i = 1 \\ \alpha + \beta x1_{2t} + \nu_2 + \epsilon_{2t}, & i = 2 \\ \dots & \\ \alpha + \beta x1_{Nt} + \nu_N + \epsilon_{Nt}, & i = N \end{cases}$$

```r
fixed = lm(y ~ x1 + factor(country) - 1, data = ex.dt)
round(summary(fixed)$coef, 4)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## x1                     2475617827 1106675594  2.2370   0.0289
## factor(country)A        880542404  961807052  0.9155   0.3635
## factor(country)B      -1057858363 1051067684 -1.0065   0.3181
## factor(country)C      -1722810755 1631513751 -1.0560   0.2951
## factor(country)D       3162826897  909459150  3.4777   0.0009
## factor(country)E       -602622000 1064291684 -0.5662   0.5733
## factor(country)F       2010731793 1122809097  1.7908   0.0782
## factor(country)G       -984717493 1492723118 -0.6597   0.5119
```

Note that now, when we take into account `country`, the coefficient at `x1` is significant and quite different from that of the pooled model.

To compare the models visually, we use:

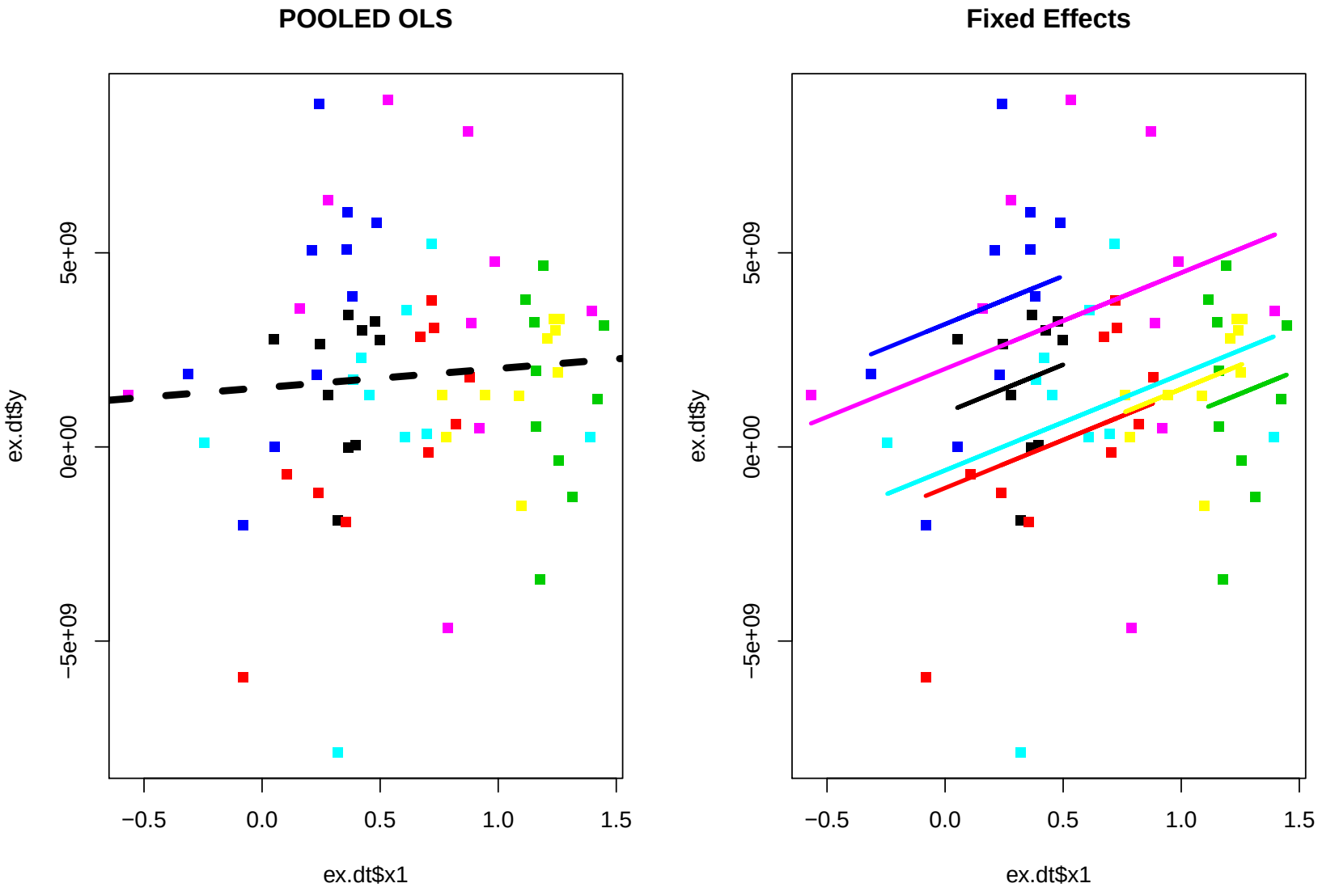
```
par(mfrow = c(1, 2))
plot(ex.dt$x1, ex.dt$y, col = as.factor(ex.dt$country),
     pch = 15, main = "POOLED OLS")
abline(pooled, lwd = 5, lty = 2)

plot(ex.dt$x1, ex.dt$y, col = as.factor(ex.dt$country),
     pch = 15, main = "Fixed Effects")
lines(ex.dt$x1[ex.dt$country=="A"],
      predict(fixed, newdata=ex.dt[ex.dt$country=="A",]),
      col=1,lwd=3)
lines(ex.dt$x1[ex.dt$country=="B"],
      predict(fixed, newdata=ex.dt[ex.dt$country=="B",]),
      col=2,lwd=3)
```

And we continue adding `lines()`'s for countries A through G:

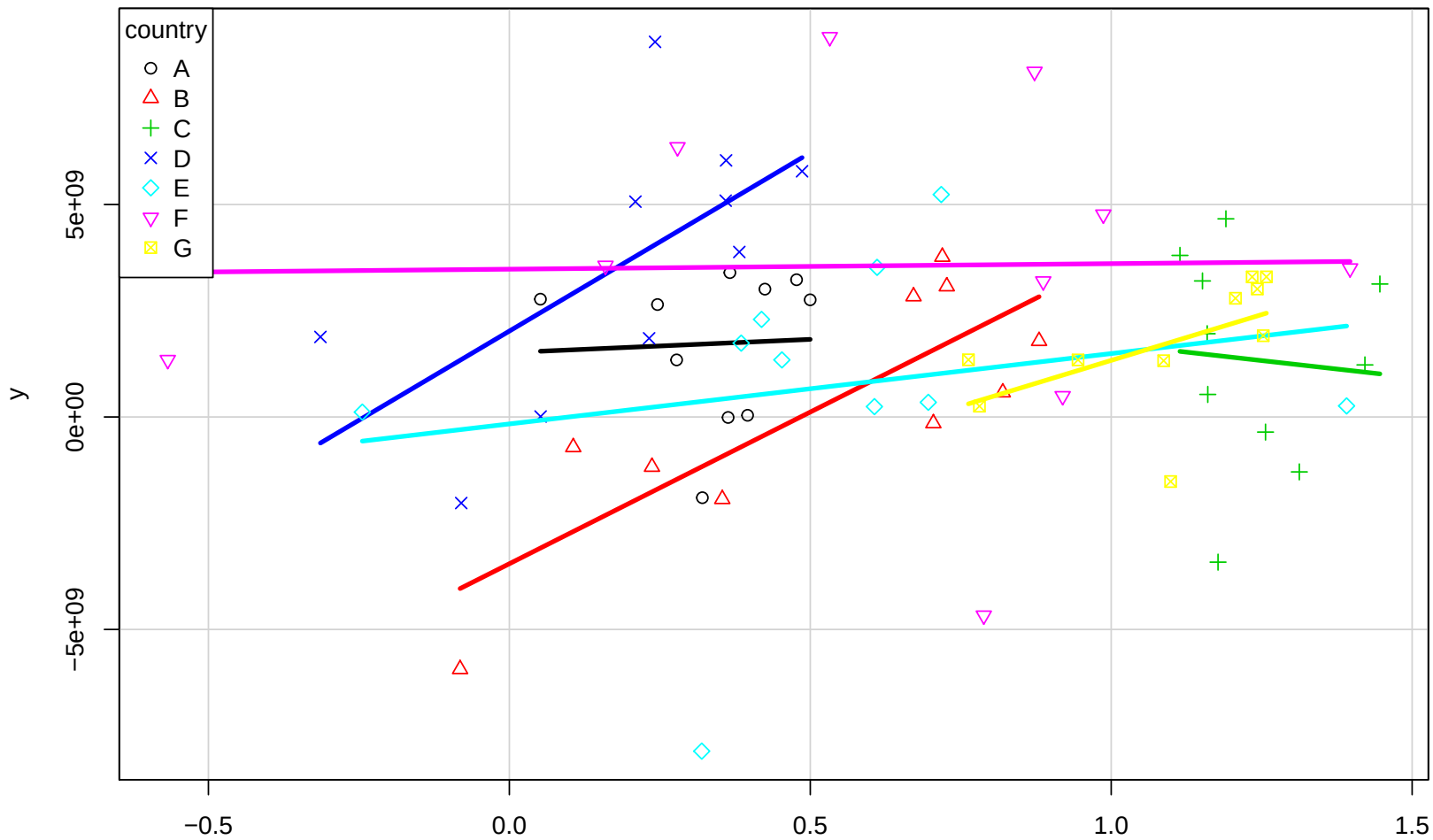
```
lines(ex.dt$x1[ex.dt$country=="G"],
      predict(fixed, newdata=ex.dt[ex.dt$country=="G",]),
      col=7,lwd=3)
```

Pooled model (left) and fixed effects model (right):



Individual countries models:

```
car::scatterplot(y ~ x1 | country, data = ex.dt,
                 legend.coords="topleft",
                 smoother = FALSE, lwd=3, reset.par = FALSE)
```



Using 'plm' function to estimate panel data models

To use a more systematic approach, we shall apply the `plm` (linear model for panel data) package.

```
suppressPackageStartupMessages({library(plm)})
```

- Pooled OLS:

```
pooled = plm(y ~ x1, data = ex.dt, index = c("country", "year"),
             model = "pooling")
round(summary(pooled)$coef, 4)
```

```
##                Estimate Std. Error t-value Pr(>|t|)
## (Intercept) 1524319070  621072624  2.4543   0.0167
## x1           494988914  778861261  0.6355   0.5272
```

- Fixed Effects Estimator:

```
fixed = plm(y ~ x1, data = ex.dt, index = c("country", "year"),
            model="within")
round(summary(fixed)$coef, 4)
```

```
##      Estimate Std. Error t-value Pr(>|t|)
## x1 2475617827 1106675594   2.237   0.0289
```

We can also extract the individual-specific (fixed) effects:

```
fixef(fixed)
```

```
##           A           B           C           D           E
##   880542404 -1057858363 -1722810755  3162826897  -602622000
##           G
##  -984717493
```

Should we even include fixed effects in our model? We can test this by comparing Fixed Effects (with individual-specific effects) and Pooled models (no individual effects).

We will use the $F - test$ to test the null hypothesis:

$$H_0 : \text{all the constants (i.e. the fixed effects) equal 0}$$

```
pFtest(fixed, pooled)
```

```
##
##  F test for individual effects
##
## data:  y ~ x1
## F = 2.9655, df1 = 6, df2 = 62, p-value = 0.01307
## alternative hypothesis: significant effects
```

Because `p-value = 0.01307 < 0.05`, we reject the null hypothesis. So, the fixed effects model is better.

As we know, we have two cases of individual-specific effect model specification: FE and RE.

Usually, there are too many parameters in the FE model and the loss of degrees of freedom can be avoided if c_i in $y_{it} = \alpha + x_{it}'\beta + z_i'\gamma + c_i + \epsilon_{it}$ are assumed random. More specifically, $c_i \sim i.i.d(0, \sigma_v^2)$, $\epsilon_{it} \sim i.i.d.(0, \sigma_\epsilon^2)$, $\mathbb{E}(c_i + \epsilon_{it}|X) = \sigma_v^2 + \sigma_\epsilon^2$, $\mathbb{E}(c_i + \epsilon_{it})(c_j + \epsilon_{jt})|X) = \sigma_v^2$.

The just presented conditions mean that the random effects (RE) model fits into the framework of a generalized LS model with autocorrelated within a group disturbances. In particular, the parameters of the RE model can be estimated consistently, though not efficiently, by OLS.

The RE model is an appropriate specification if we are drawing N individuals randomly from a large population (this is usually the case for household panel studies; in this case, N is usually large and a fixed effects model would lead to an enormous loss of degrees of freedom).

- Random Effects Estimator:

```
random = plm(y ~ x1, data = ex.dt, index = c("country", "year"),
             model = "random")
round(summary(random)$coef, 4)
```

```
##                Estimate Std. Error t-value Pr(>|t|)
## (Intercept) 1037014284  790626206  1.3116   0.1941
## x1          1247001782  902145601  1.3823   0.1714
```

Interpretation of the coefficient is tricky since it includes both the effects inside a country and between countries. In the case of time series-cross sectional data, it represents the average effect of X over Y when X changes across time and between countries by one unit.

Which of the three models to use? One hint is given by the `Effects` summary:

```
summary(random)$ercomp
```

```
##                        var   std.dev share
## idiosyncratic 7.815e+18 2.796e+09 0.873
## individual    1.133e+18 1.065e+09 0.127
## theta: 0.3611
```

Here `idiosyncratic` $= \widehat{\sigma}_\epsilon^2$, `individual` $= \widehat{\sigma}_c^2$.

Specifically, by the quantity `theta`$= \theta = 1 - \dfrac{\sigma_\epsilon}{\sqrt{\sigma_\epsilon^2 + T\sigma_v^2}}$:

- We always have $0 \leq \theta \leq 1$;
- If $\theta = 0$, the model becomes a pooled model;
- If $\theta = 1$, the model becomes a FE model.

As a rule, σ_c^2 is much bigger than σ_ϵ^2, thus θ, or, more specifically, $\widehat{\theta} = 1 - \sqrt{\widehat{\sigma}_\epsilon^2/(\widehat{\sigma}_\epsilon^2 + T\widehat{\sigma}_c^2)}$ must be close enough to 1.

The same applies, when T is big: in both cases, FE and RE models are close.

To formally decide between fixed or random effects, you can run a Hausman test where the null hypothesis is that the preferred model is RE vs. the FE alternative.

It basically tests whether the unique errors ν_i are correlated with the regressors (the null hypothesis is they are not), thus if the `p-value` is small, for example <0.05, then use fixed effects, if not use random effects.

```
phtest(fixed, random)
```

```
##
##  Hausman Test
##
## data:  y ~ x1
## chisq = 3.674, df = 1, p-value = 0.05527
## alternative hypothesis: one model is inconsistent
```

We note that `p-value = 0.05527` which is very close to `0.05`. On the other hand, since we have to choose one, because `p-value > 0.05`, we do not reject the null hypothesis. As such both models are consistent, but the RE estimator is also efficient, so we choose the RE model.

Summary

In short, in order to estimate a panel data model:

$$y_{it} = \alpha + x_{it}'\beta + z_i'\gamma + c_i + u_{it}$$

we use the Least-Squared (OLS) Methods:

- In the Pooled regression case, we assume $c_i = 0$;
- In the Fixed Effect case, we treat c_i as individual dummy variables;
- In the Random Effect case, the errors $c_i + \epsilon_{it}$ are autocorrelated, thus, we apply a Generalized Least-Squares (GLS) method.