

Panel data analysis with R

Anabela Rocha (*)
M. Cristina Miranda (*)

* ISCA and CIDMA, University of Aveiro

19 abril de 2024 - Universidade de Évora

- ▶ Introduction to panel data
- ▶ Linear regression models for panel data
- ▶ Robust methods
- ▶ Simulation study with a new robust approach
- ▶ Real data illustration

Panel data - what do we mean by that?

Cross-sectional data gives us information in a particular moment concerning some characteristics of interest from a set of units.

Table 1: Banks - Cross-section data (1st trimester)

	Average balance	Number Employees	Interest rate	Liquidity
A	4861.11	1815	0.000	1000975.5
B	6199.81	1681	0.023	1010559.6
C	4252.28	2469	0.042	997914.6
D	4424.75	1406	0.011	983638.6


In some cases, it is usual to collect the data in multiple time periods 

Table 2: Banks - Cross-section data (1st trimester)

	Average balance	Number Employees	Interest rate	Liquidity
A	4159.14	3198	0.031	998424.7
B	6384.36	1578	0.010	989282.4
C	3744.51	1669	0.004	998610.1
D	5070.14	1555	0.037	994026.9

Table 3: Banks - Cross-section data (2nd trimester)

	Average balance	Number Employees	Interest rate	Liquidity
A	4038.07	2137	0.086	992836.4
B	4707.47	2021	0.027	1002526.5
C	5258.79	2060	0.047	1001520.5
D	3847.87	2782	0.086	996923.4

Table 4: Banks - Cross-section data (3rd trimester)

	Average balance	Number Employees	Interest rate	Liquidity
A	5216.75	3145	0.001	1000747.1
B	4457.51	2482	0.001	992903.1
C	5891.14	1103	0.001	1003975.6
D	5595.98	1851	0.002	1018961.5

Table 5: Banks - panel data

	Trimester	Average balance	Number Employees	Interest rate	Liquidity
A	1st trimester	4159.14	3198	0.031	998424.66
A	2nd trimester	4038.07	2137	0.086	992836.42
A	3rd trimester	5216.75	3145	0.001	1000747.11
B	1st trimester	6384.36	1578	0.01	989282.4
B	2nd trimester	4707.47	2021	0.027	1002526.52
B	3rd trimester	4457.51	2482	0.001	992903.12
C	1st trimester	3744.51	1669	0.004	998610.14
C	2nd trimester	5258.79	2060	0.047	1001520.46
C	3rd trimester	5891.14	1103	0.001	1003975.58
D	1st trimester	5070.14	1555	0.037	994026.87
D	2nd trimester	3847.87	2782	0.086	996923.44
D	3rd trimester	5595.98	1851	0.002	1018961.49

Table 6: Banks - unbalanced panel data

	Trimester	Average balance	Number Employees	Interest rate	Liquidity
A	1st trimester	4159.14	3198	0.031	998424.66
A	2nd trimester	4038.07	–	0.086	992836.42
A	3rd trimester	5216.75	3145	0.001	1000747.11
B	1st trimester	6384.36	1578	0.01	989282.4
B	2nd trimester	4707.47	2021	0.027	1002526.52
B	3rd trimester	4457.51	2482	0.001	992903.12
C	1st trimester	3744.51	1669	0.004	998610.14
C	2nd trimester	5258.79	2060	0.047	1001520.46
C	3rd trimester	–	1103	0.001	1003975.58
D	1st trimester	5070.14	1555	0.037	994026.87
D	2nd trimester	3847.87	2782	0.086	996923.44
D	3rd trimester	5595.98	1851	0.002	1018961.49

Why panel data?

A classical regression model may help to predict the balance banks variation depending on the number of employees. Small banks tend to have lower balance values.

The existence of a small bank with a high balance (due to an unobserved effect like being located in a very rich neighbourhood) will not affect much the model.

But, when we introduce repeated measures. . .

The existing heterogeneity becomes a persistent factor!

OLS assumption of uncorrelated explanatory variables with errors falls down!

Panel data analysis - advantages

- 1- Controlling for heterogeneity
- 2- More informative data - less collinearity, more efficiency
- 3- allows to observe changes over time
- 4- Identify and measure effects that are simply not detectable in pure cross-section or pure time-series data

European Community Household Panel (ECHP)[<https://ec.europa.eu/eurostat/web/microdata/european-community-household-panel>](EUROSTAT 2001): panel survey consisting of annual interviews with a sample of households and individuals.

2020 American Life Panel Survey on Impacts of COVID-19 (Grace Carman and Nataraj 2019) (RAND corporation) American Life Panel ALP (since 2006) assess the effects of the corona virus disease 2019 pandemic on individuals and households across a variety of topics

Typical data set of panel data

Let \mathbf{x} and y be observable random variables, and let c be an unobservable random variable.

The main interest is to study the partial effects of the observable explanatory variables x_j in the population regression function.

$$y_{it} = \beta_0 + \mathbf{x}_{it}^T \boldsymbol{\beta} + c_i + u_{it}, \quad i = 1, \dots, N; t = 1, \dots, T, \quad (1)$$

i - households, individuals, countries, firms, etc.

t - time for each cross-section;

β_0 - scalar;

$\boldsymbol{\beta}$ - vector $K \times 1$;

\mathbf{x}_{it} - the i th observation on K explanatory variables ($K \times 1$ vector);

c_i - unobservable individual-specific effect;

u_{it} - remainder disturbance, idiosyncratic errors.

Example - The Grunfeld data

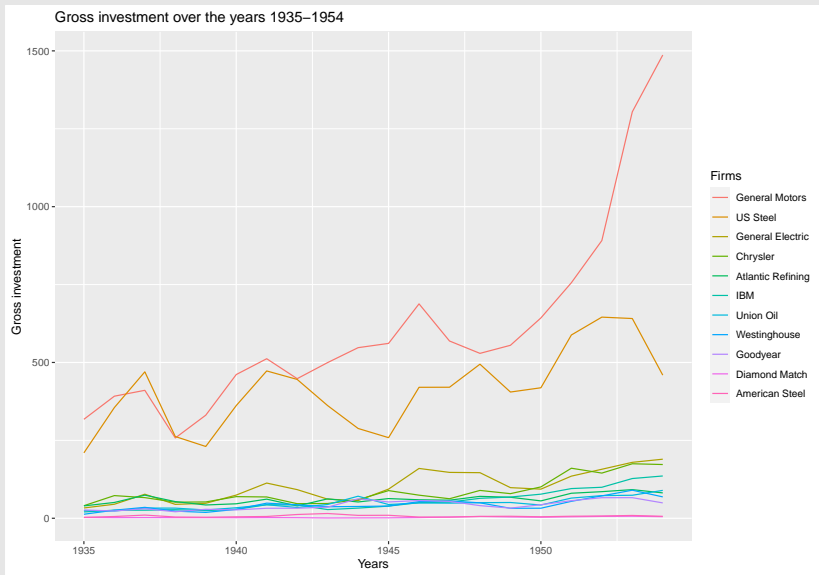
The Grunfeld data set contains investment annual data for 11 firms from US from 1935 until 1954 a total of 220 observations (11 firms \times 20 years) with values of the following variables (in dollars with reference to the year 1947):

- ▶ invest: Gross investment
- ▶ value: market value
- ▶ capital: stock of plant and equipment
- ▶ firm: General Motors, US Steel, General Electric, Chrysler, Atlantic Refining, IBM, Union Oil, Westinghouse, Goodyear, Diamond Match, American Steel
- ▶ year: 1935-1954

Example - The Grunfeld data

	invest	value	capital	firm	year
1	317.6	3078.5	2.8	General Motors	1935
2	391.8	4661.7	52.6	General Motors	1936
3	410.6	5387.1	156.9	General Motors	1937
4	NA	...
26	361.6	2202.9	254.2	US Steel	1940
27	472.8	2380.5	261.4	US Steel	1941
28	445.6	2168.6	298.7	US Steel	1942

Example - The Grunfeld data



Regression model for panel data - vector form

The equation in (1) may be written as

$$\mathbf{y} = \beta_0 \mathbf{I}_{NT} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

where

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ \dots \\ y_{1T} \\ \dots \\ y_{NT} \end{bmatrix} \quad \mathbf{I}_{NT} = \begin{bmatrix} 1 \\ 1 \\ \dots \\ \dots \\ 1 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_{111} & x_{211} & \dots & x_{K11} \\ \dots & \dots & \dots & \dots \\ x_{11T} & x_{21T} & \dots & x_{K1T} \\ x_{121} & x_{221} & \dots & x_{K21} \\ \dots & \dots & \dots & \dots \\ x_{12T} & x_{22T} & \dots & x_{K2T} \\ \dots & \dots & \dots & \dots \\ x_{1NT} & x_{2NT} & \dots & x_{KNT} \end{bmatrix}$$

and \mathbf{u} is a $(NT \times 1)$ matrix with, $\mathbf{u}_i = c_i + \mu_{it}$, $i = 1, \dots, N$; $t = 1, \dots, T$.

By definition of the model in (1),

1- $E[u_t | \mathbf{x}_t, c] = 0$ and

2- $E[\mathbf{x}_t^T u_t] = 0$

for $i = 1, \dots, T$.

Is it possible to consider different approaches to estimate the parameters of a panel data model:

- ▶ Pooled OLS
- ▶ First differences
- ▶ Random effects
- ▶ Fixed effects

When the c variable is uncorrelated with the explanatory variables, data may be treated as independent; if $c_i = c$, constant, OLS may be used to obtain consistent and efficient estimates of the common model parameters (Greene 2011).

This means we will have a single equation for all the data set.

$$y_{it} = \beta_0 + \mathbf{x}_{it}^T \boldsymbol{\beta} + u_{it}, \quad i = 1, \dots, N; t = 1, \dots, T,$$

with the usual assumptions:

- ▶ P1: linearity
- ▶ P2: exogeneity
- ▶ P3: homocedasticity and non-correlation of u
- ▶ P4: non stochastic \mathbf{x}
- ▶ P5: non multicollinearity

Repeated observations are pooled together, regardless the unit they arise from.

- ▶ No problem when there is no heterogeneity among the units.
- ▶ But, if c is correlated with any element of \mathbf{x}_t , then pooled OLS estimates will be biased and inconsistent!

Pooled regression model for Grunfeld data

Ignoring the existence of a firm effect, the regression pooled model will be represented by the equation

$$invest_{it} = \beta_0 + \beta_1 value_{it} + \beta_2 capital_{it} + u_{it}. \quad (2)$$

and we obtain the same parameter OLS estimates for all the unities.

In this case we get the fitted regression equation,

$$\hat{invest}_{it} = -38.41 + 0.115value_{it} + 0.228capital_{it}.$$

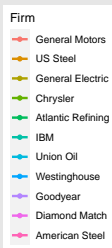
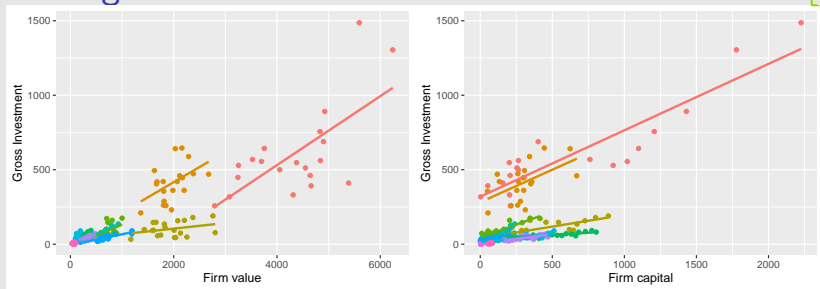
Pooled regression model for Grunfeld data

Table 7: OLS pooling estimates - Grunfeld data

	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	-38.4100540	8.4133709	-4.565358	8.4e-06
value	0.1145344	0.0055188	20.753369	0.0e+00
capital	0.2275141	0.0242283	9.390448	0.0e+00

With this model, both Capital and Value shows positive and significant relationship with Investment of a firm.

Pooled regression model for Grunfeld data



The plots show the heterogeneity for both variables among the firms.

Suppose we have $T = 2$ observed times for the variables of the model and admit c is constant over time.

The variable c is known as an **unobserved effect** or **omitted or latent variable** in panel data.

Examples:

- ▶ skills; field of knowledge; family background; region.
- ▶ number of workers; region; organic structure.

In the two different periods, there is no change in the family background of an individual, or in the organic structure of a firm!

First differences



With $c_i = c$ representing some characteristic constant over time, if we apply the first differences transformation in model (1), this effect will be eliminated.

Applying the first difference transformation results in the model,

$$y_{it} - y_{i(t-1)} = \beta(\mathbf{x}_{it} - \mathbf{x}_{i(t-1)}) + (c - c) + (\mu_{it} - \mu_{i(t-1)}),$$

or

$$\Delta y = \Delta \mathbf{x} \beta + \Delta u.$$

Assuming a random sample, it is possible to apply OLS and obtain a FD estimator,

$$\hat{\beta}_{FD}.$$

OLS estimation implies no correlation between between the error term and the independent variables, so the following conditions must be observed:

F1 - $E[\Delta \mathbf{x}^T \Delta u] = 0$, and

F2- $\text{rank } E[\Delta \mathbf{x}^T \Delta u] = K$.

F1 condition refers to the strict exogeneity assumption (a variable is said to be strictly exogenous if it does not depend on current, past, and future values of the error term u_{it}).

F2 implies that we must not have explanatory variables constant in time (since the variables present in the model result from differences in time it would mean a column of zeros and fail of F2).

First differences model for Grunfeld data

Table 8: First difference estimates - Grunfeld data

	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	-1.6539169	3.2002661	-0.516806	0.605846
value	0.0896966	0.0079583	11.270796	0.000000
capital	0.2905922	0.0506193	5.740738	0.000000

In this case the estimated regression equation is

$$\hat{invest}_{it} = -1.654 + 0.09value_{it} + 0.291capital_{it}.$$

Random effects approach treats the composite error of the model, $\nu_{it} = c_i + u_{it}$ in (1), in a context of generalized least squares (GLS) framework which demands for additional conditions *RE1*, *RE2* and *RE3*:

$$RE1 : \begin{cases} E[u_{it} | \mathbf{x}_i, c] = 0, t = 1, 2, \dots, T; \\ E[c_i | \mathbf{x}_i] = E[c_i] = 0. \end{cases}$$

$$RE2 : \text{rank} E[\mathbf{X}_i^T \boldsymbol{\Omega}^{-1} \mathbf{X}_i] = K$$

where $\boldsymbol{\Omega} = E[\mathbf{v}_i \mathbf{v}_i^T]$, (assumed to be positive definite) and $\mathbf{v}_i = c_i \mathbf{J} + \mathbf{u}_i$, (\mathbf{J} is T dimensional ones vector).

(GLS and feasible GLS are consistent under Assumptions RE.1 and RE.2 (Wooldridge 2010).)

and

$$RE3 : \begin{cases} E[\mathbf{u}_i \mathbf{u}_i^T | \mathbf{x}_i, c] = \sigma_u^2 \mathbf{I} \\ E[c_i^2 | \mathbf{x}_i] = \sigma_c^2 \end{cases}$$

which implies

$$\begin{cases} E[\mathbf{u}_i^2] = \sigma_u^2 \\ E[\mathbf{u}_i \mathbf{u}_s] = 0, \forall t \neq s \\ E[\mathbf{v}_i \mathbf{v}_i^T | \mathbf{x}_i] = E[\mathbf{v}_i \mathbf{v}_i^T]. \end{cases}$$

Units share the same β coefficients and the same intercept; it is assumed that the difference that may exist among individuals (or time periods) lies in their individual specific errors.

So, in a random effects approach, the matrix $\mathbf{\Omega}$ may be written as

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \dots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \dots & \sigma_c^2 \\ \dots & \dots & \ddots & \dots \\ \sigma_c^2 & \sigma_c^2 & \dots & \sigma_c^2 + \sigma_u^2 \end{bmatrix}$$

or

$$\mathbf{\Omega} = \sigma_u^2 \mathbf{I} + \sigma_c^2 \mathbf{J} \mathbf{J}^T.$$

The random effects estimator is the FGLS estimator that uses the variance matrix $\hat{\Omega}$, assuming that $\hat{\sigma}_u^2$ and $\hat{\sigma}_c^2$ are consistent estimators of σ_u^2 and σ_c^2 , respectively, i.e.,

$$\hat{\beta}_{RE} = \left(\sum_{i=1}^N \mathbf{x}_i^T \hat{\Omega}^{-1} \mathbf{x}_i \right)^{-1} \left(\sum_{i=1}^N \mathbf{x}_i^T \hat{\Omega}^{-1} \mathbf{y}_i \right),$$

is consistent and efficient under the condition $E[\mathbf{v}_i | x_i] = 0$.

LM - Breusch-Pagan test Pooled vs random

Breusch and Pagan (1980) propose a Lagrange multiplier test for the random effects model based on the OLS residuals e_{it} :

$$H_0 : \sigma_c^2 = 0 \text{ vs } H_1 : \sigma_c^2 \neq 0.$$

The test statistic is LM (Greene 2011),

$$LM = \frac{nT}{2(T-1)} \left[\frac{\sum_{i=1}^n (T\bar{e}_{i\cdot})^2}{\sum_{i=1}^n \sum_{t=1}^T e_{it}^2} - 1 \right];$$

under H_0 $LM \sim \chi_1$.

Rejecting H_0 means the pooled model is not suitable.

Table 9: Random effects estimates - Grunfeld data

	Estimate	Std. Error	z-value	Pr(> z)
(Intercept)	-53.9436014	25.6969760	-2.09922	0.0357975
value	0.1093053	0.0099138	11.02556	0.0000000
capital	0.3080360	0.0163873	18.79724	0.0000000

Obtaining the common part of the estimated equation,

$$\hat{invest}_{it} = -53.944 + 0.109value_{it} + 0.308capital_{it}.$$

with estimated deviations from the intercept which may be obtained from equation

$$\hat{c}_i = \beta_0 + \bar{y}_i - \bar{x}_i \hat{\beta}_{RE},$$

resulting in different equations for each firm, with deviations

	deviations from common intercept
General Motors	-11.26035
US Steel	154.90212
General Electric	-175.64458
Chrysler	26.40907
Atlantic Refining	-58.30666
IBM	30.71117
Union Oil	-11.61131
Westinghouse	-2.82158
Goodyear	-31.75337
Diamond Match	46.49745
American Steel	32.87806

OLS is also possible to use with some adequate transformation of the model:

▶ regression with deviations from the group means (**within** transformation)

▶ regression with the group means (**between** groups transformation)

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)^T \boldsymbol{\beta} + u_{it} - \bar{u}_i,$$

$$\bar{y}_i = \bar{\mathbf{x}}_i^T \boldsymbol{\beta} + u_i - \bar{u}_i,$$

where $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}$, $\bar{\mathbf{x}}_i^T = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_{it}^T$, and $\bar{u}_i = \frac{1}{T} \sum_{t=1}^T u_{it}$.

The **within** transformation results in the **within** or **Fixed Effects** (FE) estimator.

This model assumes differences across units can be captured in **differences in the constant term** and each intercept is a parameter to be estimated.

In this case, we must have strict exogeneity of the explanatory variables conditional on c_i ,

$$FE1 : E(u_{it} | \mathbf{x}_i, c_i) = 0, t = 1, 2, \dots, T$$

but we admit correlation between c_i and \mathbf{x}_i , and this is the main difference between the RE and the FE estimator.

We also must observe the conditions

$$FE2 : \text{rank}E[(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)^T(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)] = K,$$

and

$$FE3 : E[\mathbf{u}_i\mathbf{u}_i^T | \mathbf{x}_i, c_i] = \sigma_u^2 \mathbf{I}_T$$

to assure efficiency (Wooldridge 2010).

Fixed effects (FE) estimator



The **FE** estimator (also known as **within estimator**) may be expressed as

$$\hat{\beta}_{EF} = \left(\sum_{i=1}^N \sum_{t=1}^T (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})^T (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.}) \right)^{-1} \left(\sum_{i=1}^N \sum_{t=1}^T (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})^T (\mathbf{y}_{it} - \bar{\mathbf{y}}_{i.}) \right)$$

Under *FE1* to *FE3*, the FE estimator, $\hat{\beta}_{EF}$ is an unbiased estimator, conditional to \mathbf{X} and the different values of the intercept for each unit may be obtained by

$$\hat{c}_i = \bar{y}_i - \bar{\mathbf{x}}_i \hat{\beta}_{FE}.$$

Alert!

- ▶ Time-constant factors **must not be included** in \mathbf{x}_{it} with FE model.

As we allow for correlation between c_i and \mathbf{x}_i , this means that we cannot tell the difference between the effects from time-constant observable variables and time-constant unobservable c_i !

Variables like gender, location, field of action, etc., (constant over time) should not be included.

- ▶ Is it acceptable considering all the intercepts as equal?

An F statistic comparing the unrestricted model (FE) with a restricted model (pooled model with Q restriction on β_0) may be considered to test if all the intercepts are equal (Baltagi 2021), (Gujarati 2004):

$$F = \frac{SSR_r - SSR_{ur}}{SSR_{ur}} \frac{N(T-1) - K}{Q} \sim F_{Q, N(T-1)-K}$$

Rejecting H_0 means the pooled model is not adequate.

F-test for pooled vs FE (Grunfeld data)

With the Grunfeld data, we obtain the indication that a common intercept is not adequate for the data.

The F observed value is 5.7218251 with 20 , 187 degrees of freedom and a p-value close to zero.

Table 10: Fixed effects estimates - Grunfeld data

	Estimate	Std. Error	t-value	Pr(> t)
value	0.1101291	0.0112998	9.746075	0
capital	0.3100334	0.0165405	18.743924	0

For the fixed effects we obtain the estimated β coefficients,

$$\hat{invest}_{it} = 0.11value_{it} + 0.31capital_{it},$$

And also, the intercept for each firm. For example,

$$\hat{invest}_{it} = -70.299 + 0.11value_{it} + 0.31capital_{it},$$

for the General Motors firm, and

$$\hat{invest}_{it} = 101.905 + 0.11value_{it} + 0.31capital_{it},$$

for the US Steel.

Which model to choose? Fixed or random effects?

Hausman test

In most situations it is unlikely to have zero correlations between the random effects and the explanatory variables. In that case, β will be inconsistent.

The Hausman test (Hausman, Hausman, and Jerry 1978) helps to evaluate if that correlation exists.

Under the hypothesis of no correlation, both estimators are consistent. So, it is expected both to produce similar estimates.

The Hausman test focus in the difference $[\hat{\beta}_{FE} - \hat{\beta}_{RE}]$

The Hausman test

The Hausman statistic test is

$$W = [\hat{\beta}_{FE} - \hat{\beta}_{RE}]^T \hat{\Psi}^{-1} [\hat{\beta}_{FE} - \hat{\beta}_{RE}],$$

and $W \sim \chi_{k-1}^2$, where $\hat{\Psi} = \text{Var}[\hat{\beta}_{FE} - \hat{\beta}_{RE}]$,

- ▶ Not to reject H_0 means both estimators are consistent and, in that case, $\hat{\beta}_{RE}$ is a better option because it is more efficient.
- ▶ Rejecting H_0 means the $\hat{\beta}_{FE}$ should be chosen because $\hat{\beta}_{RE}$ is not consistent.

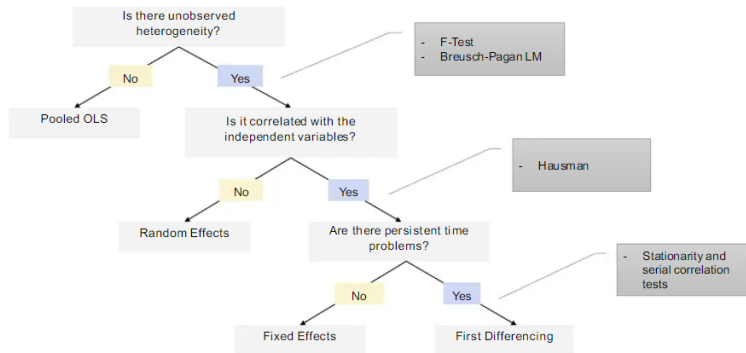
The Hausman test - Grunfeld data

The Hausman test for the Grunfeld data points not to reject the H_0 hypothesis with a p - *value* = 0.1375503.

This means that the $\hat{\beta}_{RE}$ estimator should be considered since in this case, both estimators are consistent but $\hat{\beta}_{RE}$ is more efficient.

Which model to choose?

A useful summary scheme shows the possible paths searching for an adequate model (Mesquita, Fernandes, and Filho 2021).



Source: Mesquita et al. (2021)

- ▶ In real panel data, it is common to have the failure of the assumed hypotheses and frequently, atypical observations occur.
- ▶ Outliers are observations with low probability of belonging to the same distribution of the majority of the data.
- ▶ Classical estimation of PDM (FGLS) may be seriously affected by the presence of outliers.
- ▶ It's important to detect the presence of outliers in panel data, but visual observation is not easy, given the multivariate nature of this type of data.

Robust estimation must be less affected by the presence of outliers.

Some robust procedures have been proposed for PDM: Bramati and Croux (2007), Aquaro and Cížek (2018), Dhaene and Zhu (2017), Ji, A., et al. (2023). These authors adapted robust regression methods (P. J. Rousseeuw and Leroy 1987) to PDM.

The implementation of robust estimation methods is not as accessible as the implementation of classical estimation methods for this type of models, and, perhaps for this reason, there are few papers with application of robust methodologies in the fields of economics and finance.

We intend to make a contribution to reverse this trend, proposing a robust estimator for PDM that may be more easy to apply.

Our proposal is to robustify the FGLS estimator, recalling that its implementation includes three steps.

- ▶ Estimate the Pooled Model parameters using Least Squares and collect the residuals (error estimates).
- ▶ Estimate the errors covariances matrix Ω by sample covariances matrix of the residuals of the former step.
- ▶ Estimate the model parameters by FGLS with the estimated covariance matrix obtained in the former step.

We propose RFGLS (Robust Feasible Generalized Least Squares) estimator where we implemented robust techniques in FGLS steps.

1. Estimate the Pooled Model parameters using robust regression method - we used Least Trimmed Squares (LTS) (P. J. Rousseeuw 1984) and compute the corresponding residuals.
2. Estimate the errors covariance matrix applying a robust covariance matrix estimator - we use Minimum Volume Ellipsoide (MVE) (P. Rousseeuw 1985) applied to LTS residuals.
3. Filter the original data matrix - we used univariate-and-bivariate filter (UBF) (Agostinelli et al. 2015) and Leung et al., (2017).
4. Replace the identified outliers in the former step with the means of each variable.
5. Estimate the model parameters by FGLS from the filtered data matrix and use the robust estimated covariance matrix obtained at the second step.

Robust methods involved

- ▶ Least Trimmed Squares (**LTS**) estimator (P. J. Rousseeuw 1984):

This is a robust regression method that estimates the model parameters through the minimization of the sum of the h smallest squared residuals.

In this way, the method does not consider the $n-h$ largest absolute residuals in the estimation process.

LTS estimator tolerates $n-h$ outliers and, for $n=h$ corresponds to OLS estimator.

The LTS estimator has low efficiency but provides a good initial estimate.

- ▶ Minimum Volume Ellipsoide estimator (**MVE**) (P. Rousseeuw 1985):

This is a a robust location and scale estimation method that searches, among all ellipsoids that contain at least half of the sample points, for the one with the minimum volume.

The location MVE estimator corresponds to the mean of this ellipsoid points and the scale MVE estimator corresponds to the covariance matrix of the points of this ellipsoid.

The MVE estimator admits a high percentage of outliers without producing arbitrary values of the estimate, but has low efficiency.

Robust methods involved

Considering datasets in matrix form: rows are the cases and columns are the variables, different types of outliers may occur.

Many real datasets contain outliers and it is important to be able to detect him!

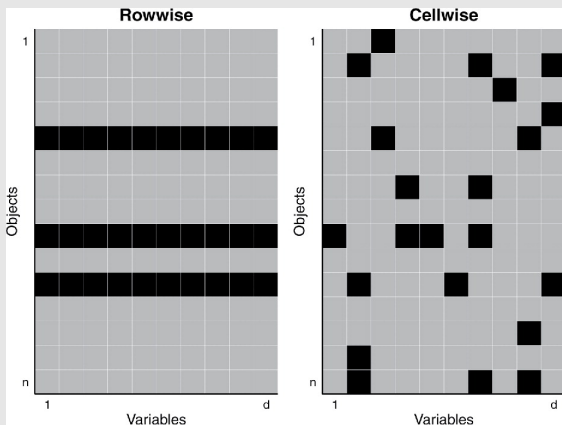


Figure 1: Rousseeuw & Bossche (2018).

- ▶ Traditionally, outlier refers to a case, a row of the data matrix - **casewise outlier**.
- ▶ Recently, Alqallaf et al. (2009) proposed **cellwise outlier**, when most of the data cells in a row are similar but some of them are atypical.
- ▶ Robust fitting methods are less sensitive to casewise outliers and allow to detect these type of outliers (Maronna, Martin, and Yohai 2006).
- ▶ Recent work has been done on how to identify cellwise outliers and how to deal with him in the estimation and fitting processe.

- ▶ Univariate-and-bivariate filter (UBF) (Agostinelli et al. 2015) and (Leung, Yohai, and Zamar 2017)

Univariate filter - flags cellwise outliers by comparing the standardized empirical distribution of each marginal with a high quantile of the standard normal distribution.

Bivariate filter - flags casewise outliers by comparing the squares of the pairwise robust Mahalanobis distances with a high quantile of a chi-square with 2 d.f. distribution. A cell is additionally flagged if the number of flagged pairs exceeds a large quantile of the binomial model, considering that the number of flagged pairs associated with each cell approximately follows a binomial model.

To evaluate the performance of proposed robust estimator, RFGLS, we run a simulation study.

- ▶ We generate some data set and next we contaminated them, including outliers in two ways and considering different percentages of contamination.
- ▶ The simulation settings were defined taking into account the works of Bramati and Croux (2007) and Aquaro and Cížek (2018).

R packages used: *plm*, *robustbase*, *rrcov* and *GSE*.

- ▶ The dependent variables values are generated according to the RE model, defined in 1.
- ▶ The errors values are generated according with a $N(0, 1)$ distribution.
- ▶ For the parameter vector, we consider $\beta = (-1, 0, 1)$ and we generate c_i values according with a $N(0, 1)$ distribution.
- ▶ The explanatory variables values are generated from a multivariate (dimension three) standard normal.
- ▶ We generated data panels with 240 observations resulting from two cases of the dimensions: $N = 8$ and $T = 30$ and $N = 12$ and $T = 20$.

In the sample generation process, we consider different scenarios of contaminated samples:

- ▶ without contamination
- ▶ with 5% contamination
- ▶ with 10% contamination

The contamination is generated completely randomly over all observations of the panel of data.

The contamination was introduced including outliers in two different ways:

- ▶ only on y (to originate vertical outliers), by adding to some of the y initially generated, a term generated according with $N(50, 1)$;
- ▶ on y and x (to originate bad leverage points), by replacing the explanatory variables values, corresponding to the observations already contaminated in y , by points coming from a k -variate $N(10, I)$ distribution, with $k = 3$.

We run $M = 100$ replications for each of the 10 sampling schemes, resulting in a total of 20 scenarios and 2000 runs.

For each run, we estimated the β coefficient of the model using FGLS and RFGLS estimators.

We have used the Root Mean Squared Error (RMSE) criteria to evaluate the performance of the two approaches each sampling scheme,

$$RMSE = \sqrt{\frac{1}{M} \sum_{j=1}^M \|\hat{\beta}^j - \beta\|^2}.$$

Since RMSE is a measure of the estimation error, the estimator's performance will be better the lower the RMSE value.

Table 11: RMSE for FGLS and RFGLS, $T = 30$, $N = 8$ and contamination 0%, 5% and 10%

	C0	C5		C10	
		conty	conty and x	conty	conty and x
FGLS	0.15	1.26	2.70	1.71	2.81
RFGLS	0.25	0.24	0.19	0.27	0.21

- ▶ We can see that the RMSE values are always smaller for the RFGLS estimator, except in the case where there is no data contamination.
- ▶ This means that the robust estimator always generates more accurate estimates in the contamination situations considered.

- ▶ We also note that in presence of bad leverage points (contamination on y and x) the results obtained by RFGLS are particularly good in contrast to the results obtained with FGLS which are negatively affected in these cases.
- ▶ **Only** in the case of well-behaved samples, **without outliers**, does the FGLS estimator outperform the RFLGS.

Table 12: RMSE for FGLS and RFGLS, $T = 20$, $N = 12$ and contamination 0%, 5% and 10%

	C0	C5		C10	
		conty	conty and x	conty	conty and x
FGLS	0.15	1.21	2.70	1.81	2.81
RFGLS	0.24	0.25	0.21	0.24	0.18

Notice the same pattern.

- ▶ The obtained estimates with RFGLS are more precise, with lower RMSE, than those obtained with FGLS in all contamination cases considered.
- ▶ In summary, for the considered cases, we conclude that without contamination, the efficiency of the robust estimator is not as good as the one of the classic estimator.
- ▶ The RMSE results for the robust estimator improve as expected in the presence of outliers. This happens both for vertical outliers and bad leverage points, for all dimensions and percentages of contamination considered.
- ▶ The proposed robust estimator presents particularly positive results that compare better to the classical estimator in the presence of bad leverage points.

Grunfeld data - outliers

To explore the existence of outliers in the Grunfeld data set we have applied the Robust Mahalanobis distance (available with **chemometrics** package) which allowed to identify three atypical firms.

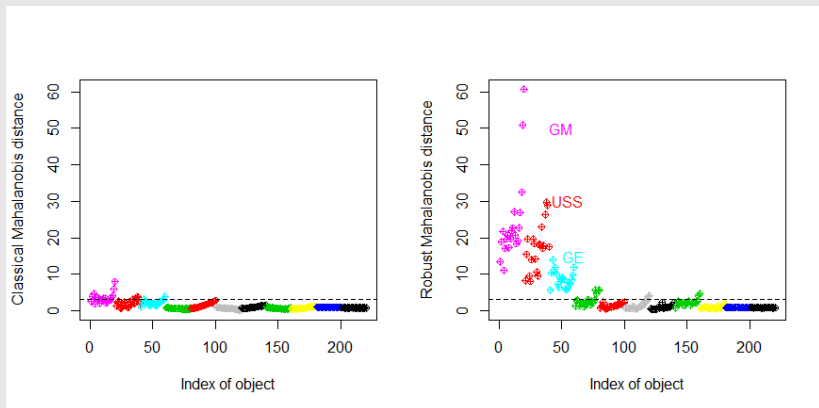


Figure 2: Grunfeld outliers

We have estimated the parameters for the model with the Grunfeld data set using RFGLS and FGLS methods.

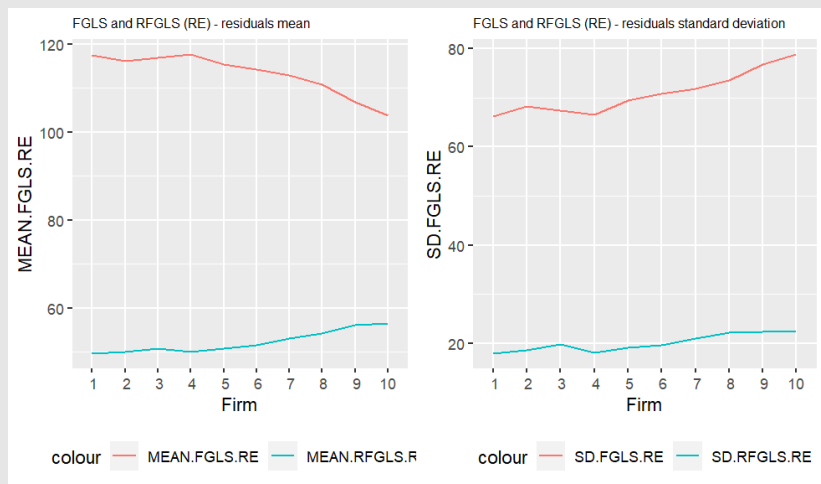
Table 13: RFGLS and FGLS estimatives

	x
beta1.RFGLS	0.0469826
beta2.RFGLS	0.1208576
beta1.FGLS	0.1097104
beta2.FGLS	0.3073739

To evaluate the performance of the two estimators in this case, we calculated the multivariate residuals for each of the fitted models and carried out a residual analysis. Then, we compared each of the mean and standard deviation obtained values.

Grunfeld data

Figure 3 shows that the robust method performs better. The residuals obtained with the robust method RFGLS present smaller mean (left) and standard deviation (right) values for every company.



Both mean and standard deviation were lower for the robust method.

- ▶ Panel data is a suitable representation for economical and financial data. Financial and economic real data often contain outliers and violate the assumptions usually assumed in the model.
- ▶ Robust methods are recommended for this type of data analysis. We propose a robust estimator for panel data model which results from the robustification of the FGLS estimator, called RFGLS.
- ▶ The RFGLS estimator performed well with contaminated simulated data. The proposed robust estimator RFGLS improve regardind to FGLS as expected in the presence of several type of outliers.

- ▶ Either for vertical outliers or for leverage points and for all dimensions and percentages of contamination considered, RFGLS performs better than FGLS, according to RMSE. The RFGLS produces particularly good estimates for panels of data with bad leverage points.
- ▶ The RFGLS estimator did not perform as well without contaminated data, as expected.
- ▶ The robust estimated model for the Grunfeld data is less affected by the identified outliers (for the three firms, GM, USS, GE) than the classical estimated model.

To be continued. . .

- ▶ More simulations with different scenarios will allow a better evaluation of the robustness properties of the proposed estimator.
- ▶ Contaminate data panels by introducing the contamination in a concentrated way in certain groups (concentrated contamination) is another case already studied by other authors which would also be interesting to analyse.
- ▶ Include a robust regression method in the last step may improve the robustness properties of the robust estimator.

Acknowledgements

This work is supported by The Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology

(FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020.

<https://doi.org/10.54499/UIDB/04106/2020> and

<https://doi.org/10.54499/UIDP/04106/2020>.



- Agostinelli, Claudio, Andy Leung, Victor J. Yohai, and Ruben H. Zamar. 2015. "Rejoinder on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination." *Test* 24 (3): 484–88.
<https://doi.org/10.1007/S11749-015-0457-Z/TABLES/2>.
- Alqallaf, Fatemah, Stefan Van Aelst, Victor J. Yohai, and Ruben H. Zamar. 2009. "Propagation of Outliers in Multivariate Data." *The Annals of Statistics* 37 (1): 311–31.
<http://www.jstor.org/stable/25464750>.
- Baltagi, B. H. 2021. *Econometric Analysis of Panel Data*. Springer Texts in Business and Economics. Springer International Publishing.
<https://books.google.pt/books?id=h6QjEAAAQBAJ>.

References II



- Bramati, Maria Caterina, Christophe Croux, Maria Caterina Bramati, and Christophe Croux. 2007. "Robust estimators for the fixed effects panel data model." *Econometrics Journal* 10 (3): 521–40. <https://econpapers.repec.org/RePEc:ect:emjrn1:v:10:y:2007:i:3:p:521-540>.
- Breusch, T. S., and A. R. Pagan. 1980. "The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics." *The Review of Economic Studies* 47 (1): 239. <https://doi.org/10.2307/2297111>.
- Dhaene, Geert, Yu Zhu, Geert Dhaene, and Yu Zhu. 2017. "Median-based estimation of dynamic panel models with fixed effects." *Computational Statistics & Data Analysis* 113 (C): 398–423. <https://doi.org/10.1016/J.CSDA.2016.05.021>.
- Ek, P., and M. Aquaro. 2018. "Robust estimation and moment selection in dynamic fixed-effects panel data models." *Computational Statistics* 33 (2): 675–708. <https://doi.org/10.1007/S00180-017-0782-7/TABLES/4>.

- EUROSTAT. 2001. "European Community Household Panel - Eurostat."
<https://ec.europa.eu/eurostat/web/microdata/european-community-household-panel>.
- Grace Carman, Katherine, and Shanthi Nataraj. 2019. "2020 American Life Panel Survey on Impacts of COVID-19: Technical Documentation."
<https://kclpure.kcl.ac.uk/portal/files/127048253/>.
- Greene, W. H. 2011. *Econometric Analysis*. Edited by Pearson. 7th Edition.
- Gujarati, Damodar N. 2004. *Basic Econometrics* . 4th ed. Tata McGraw Hill.
- Hausman, Jerry A., Hausman, and Jerry. 1978. "Specification Tests in Econometrics." *Econometrica* 46 (6): 1251–71.

- Leung, Andy, Victor Yohai, and Ruben Zamar. 2017. "Multivariate location and scatter matrix estimation under cellwise and casewise contamination." *Computational Statistics & Data Analysis* 111 (July): 59–76. <https://doi.org/10.1016/J.CSDA.2017.02.007>.
- Maronna, Ricardo A., R. Douglas Martin, and Víctor J. Yohai. 2006. *Robust Statistics: Theory and Methods*. Wiley. <https://doi.org/10.1002/0470010940>.
- Mesquita, Rafael, Antônio Alves Tôres Fernandes, and Dalson Britto Figueiredo Filho. 2021. "Uma Introdução à Regressão Com Dados de Painel." *Revista Política Hoje* 29 (July): 434–507. <https://doi.org/10.51359/1808-8708.2021.246522>.
- Rousseeuw, Peter. 1985. "Multivariate estimation with high breakdown point." *Mathematical Statistics and Applications*, 283–97. https://doi.org/10.1007/978-94-009-5438-0_20.

- Rousseeuw, Peter J. 1984. "Least Median of Squares Regression." *Journal of the American Statistical Association* 79 (388): 871. <https://doi.org/10.2307/2288718>.
- Rousseeuw, Peter J., and Annick M. Leroy. 1987. "Robust Regression and Outlier Detection," Wiley series in probability and statistics, October. <https://doi.org/10.1002/0471725382>.
- Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. The MIT Press.