Luca Colaci - 3176608
Rocco Totaro - 3159435
Giovanni M. Parlangeli - 3152699

Econometrics 2nd Group Assignment: Panel Data Estimation

Course
BEMACS '24

Deadline:
11/06/2023

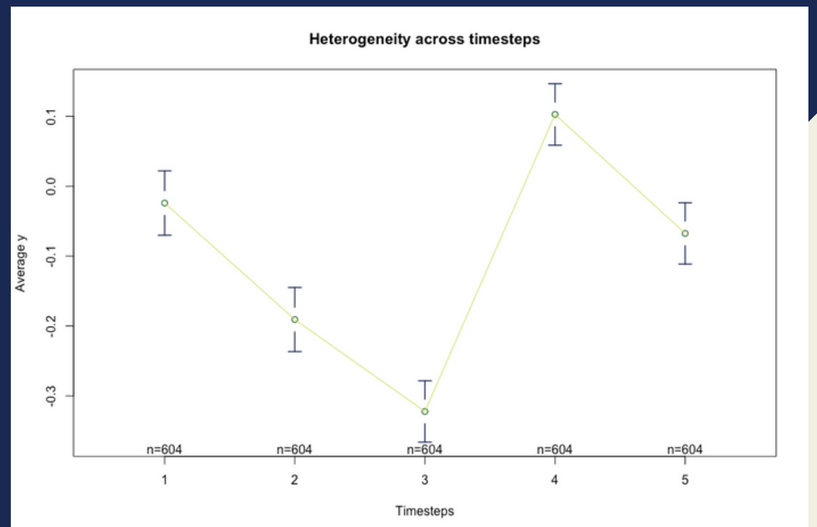
Summary

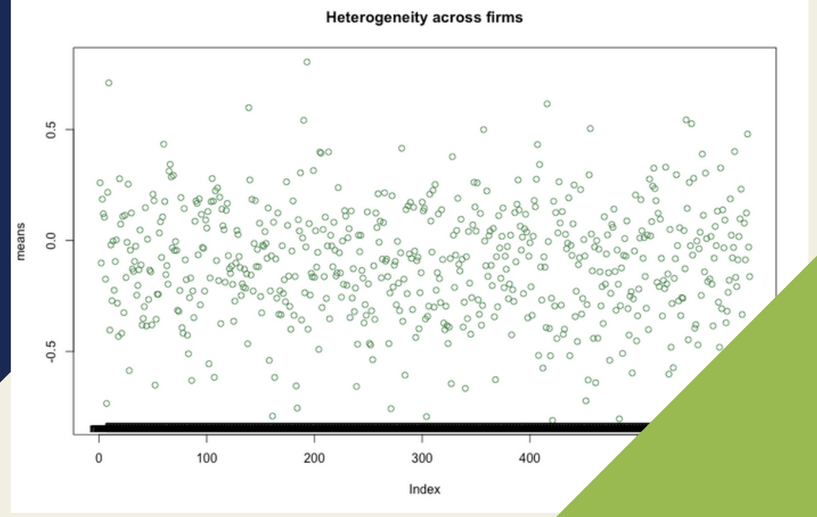
1 - Introduction

Our goal for this group assignment was to determine, given the assigned dataset, which was the best regression method considering the nature of our data, a panel dataset with 604 firms observed in 5 timesteps.

We decided to approach the task as follows:

1. First, run some preliminary checks and tests on our data to have an idea of what to expect from our analysis;
2. Second, implement some regression methodologies and check each one for heterogeneity and serial correlation in residuals;
3. Third, run specificity tests to rule out models, also implementing robust methodologies if the tests run in the previous step raised any red flag.





2 - Exploratory Analysis

We first decided to run a couple of tests to assure that:

- No missing values were present in the data frame;
- No firm had a different number of timesteps than the others and vice versa.

All tests reported a positive output, so our panel data was in perfect condition.

To give a more immediate understanding of the data we were working with, we decided to plot the average output across time and firms to understand which of the two had more significant effects:

- Looking at the timesteps versus y graph we can notice that effects are well defined and contained, with rather small error bars compared to data spread;
- On the other hand, the same graphs with firms grouping reports a more chaotic situation, with means clustered around 0 and error bars (not shown here for clarity) that most of the time incorporate the whole spread of data.

Due to these findings, we respect to see more relevant effects from time than firms.

3 - Model Estimation

```
> re3 = plm(re_formula, data = pdata, model = 'random', effect = 'twoway')
> summary(re3)
Twoways effects Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = re_formula, data = pdata, effect = "twoway", model = "random")

Balanced Panel: n = 604, T = 5, N = 3020

Effects:
                  var  std.dev share
idiosyncratic 0.243681 0.493641 0.870
individual    0.002449 0.049488 0.009
time          0.033929 0.184199 0.121
theta: 0.02422 (id) 0.8916 (time) 0.02419 (total)

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-1.6833185 -0.3417065 -0.0082019  0.3387188  1.4757212

Coefficients:
            Estimate Std. Error z-value Pr(>|z|)
(Intercept) -0.097243   0.082855 -1.1737   0.2405
k            0.471306   0.018071 26.0802   <2e-16 ***
l            0.377926   0.030371 12.4436   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    935.24
Residual Sum of Squares: 734.58
R-Squared:      0.21456
Adj. R-Squared: 0.21404
Chisq: 824.161 on 2 DF, p-value: < 2.22e-16
```

We then implemented a comprehensive list of models that we thought suited our dataset, that is:

- Pooled OLS model;
- Fixed Effect models (*individual*, *time* and *two-way effect*);
- Random Effect models (individual, time and two-way effect);
- Between model.

All models regressed successfully with various precision scores, the worst being the individual FE model with a negative adjusted R^2.

During the various regressions, we also run some tests to look for heteroscedasticity (*BP test for panel data*) and serial correlation (*Woolridge*): most of the tests failed to reject the null hypothesis, therefore excluding the presence of heteroscedasticity and serial correlation in residuals.

Only two models raised red flags:

- Individual RE model rejected the null of no serial correlation, therefore we created an Arellano variance-covariance matrix to use in future specificity tests;
- Pooled OLS model also rejected no serial correlation, so we created a second Arellano variance-covariance matrix.

On the side, we reported the regression results of a particularly successful model, the two-way RE model.

4 - Specificity Tests

We now have to run tests in order to rule out regression models: where we have multiple variations of the same model (that is the case, for example, of the multiple instances of RE and FE models caused by the number of possible effects) tests will be run for each of them, interpreting results individually.

01

LATENT HETEROGENEITY

This test aims at comparing the performances of FE models against the Pooled OLS model. It tests if effects are jointly 0 through an F-test, with significant effects as the alternative hypothesis.

02

HAUSMAN TEST

This test aims at comparing performances of FE models against RE ones by looking at consistency of certain additional assumptions: if these hold in both situations we're in the null hypothesis, accepting RE, otherwise only FE comes out as consistent and is therefore accepted.

03

FINAL MODEL ANALYSIS

Last, we will compare the between model with the RE model variations in order to determine exactly which is the best-suited model for our data.

```
> # Testing POLS vs FE models
> pFtest(fe, pols)

        F test for individual effects

data:  fe_formula
F = 0.94035, df1 = 603, df2 = 2414, p-value = 0.8254
alternative hypothesis: significant effects

> pFtest(fe2, pols)

        F test for time effects

data:  fe_formula
F = 69.114, df1 = 4, df2 = 3013, p-value < 2.2e-16
alternative hypothesis: significant effects

> pFtest(fe3, pols)

        F test for twoways effects

data:  fe_formula
F = 1.5, df1 = 607, df2 = 2410, p-value = 2.449e-11
alternative hypothesis: significant effects
```

4.1 - Latent Heterogeneity

The latent heterogeneity test aims at evaluating the significance of individual/time/two-way effect compared to a simple OLS model. Unfortunately, we were not able to find a robust version of this test available, so we'll have biased results.

We performed the test on all variations of the FE model, with the following results:

- When testing POLS against the individual FE model, the tests fail to reject the "insignificant effects" hypothesis, preferring the latter model;
- Comparing the time FE model makes the test reject the null, signalling significant effects from the time component;
- A similar result is obtained when testing the two-way FE model, with the null rejected once again.

Our tests, therefore, show poor effects from the individual component but good results on its counterparts, rejecting the null two out of three times.

4.2 - Hausman Tests

We decided to run all possible combinations of tests, therefore evaluating the three effects variations of RE and FE models against each other. Results came back positive for all tests, with failure to reject the null hypothesis in all 9 tests. We can therefore be pretty confident in saying that RE models are preferred for this particular dataset, a decision backed up by the tests' confidence.

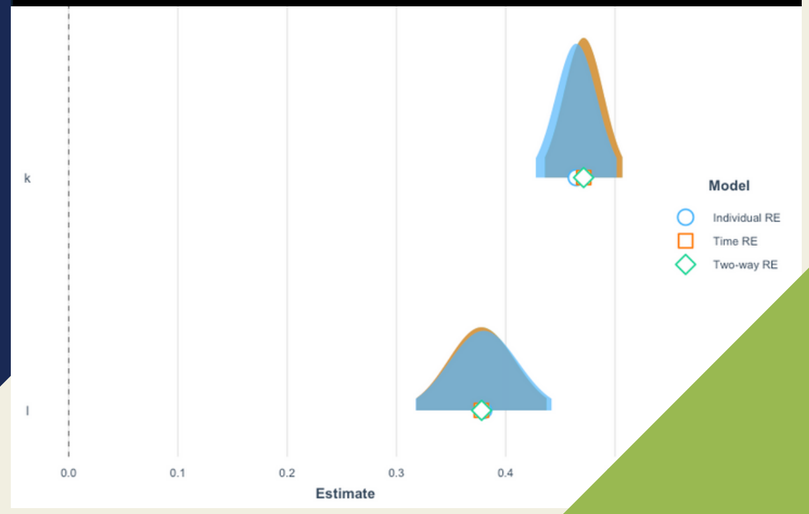
For the first batch of tests (individual RE vs. FE variations) the test was robustified with an Arellano matrix for the RE model.

```
> # Set 2: Time RE vs FE
> phtest(fe, re2)

        Hausman Test

data:  fe_formula
chisq = 1.1763, df = 2, p-value = 0.5554
alternative hypothesis: one model is inconsistent

> phtest(fe2, re2)

        Hausman Test

data:  fe_formula
chisq = 0.084484, df = 2, p-value = 0.9586
alternative hypothesis: one model is inconsistent

> phtest(fe3, re2)

        Hausman Test

data:  fe_formula
chisq = 0.22833, df = 2, p-value = 0.8921
alternative hypothesis: one model is inconsistent

>
> # In all cases the RE model is preferred over the FE
```

```
> # Testing RE vs FE models
> # Set 1: Individual RE vs FE (we use a robust variance-covariance matrix)
> phtest(fe, re, cov.fe = vcov_re)

        Hausman Test

data:  fe_formula
chisq = 0.34518, df = 2, p-value = 0.8415
alternative hypothesis: one model is inconsistent

> phtest(fe2, re, cov.fe = vcov_re)

        Hausman Test

data:  fe_formula
chisq = 1.7354, df = 2, p-value = 0.4199
alternative hypothesis: one model is inconsistent

> phtest(fe3, re, cov.fe = vcov_re)

        Hausman Test

data:  fe_formula
chisq = 0.46699, df = 2, p-value = 0.7918
alternative hypothesis: one model is inconsistent
```

```
> # Set 3: Two-way RE vs FE
> phtest(fe, re3)

        Hausman Test

data:  fe_formula
chisq = 1.1497, df = 2, p-value = 0.5628
alternative hypothesis: one model is inconsistent

> phtest(fe2, re3)

        Hausman Test

data:  fe_formula
chisq = 0.97299, df = 2, p-value = 0.6148
alternative hypothesis: one model is inconsistent

> phtest(fe3, re3)

        Hausman Test

data:  fe_formula
chisq = 0.21111, df = 2, p-value = 0.8998
alternative hypothesis: one model is inconsistent

>
> # In all cases the RE model is preferred over the FE
```

```
> phtest(be, re)

        Hausman Test

data:  be_formula
chisq = 0.38741, df = 2, p-value = 0.8239
alternative hypothesis: one model is inconsistent

> phtest(be, re2)

        Hausman Test

data:  be_formula
chisq = 0.21474, df = 2, p-value = 0.8982
alternative hypothesis: one model is inconsistent
```



4.3 - Final Model Analysis

We were left with determining the best model by ruling out one among the Between model and the RE model variations. After a thorough online research, we found out that the two can be compared by using the Hausman test if the RE model is not a two-way one.

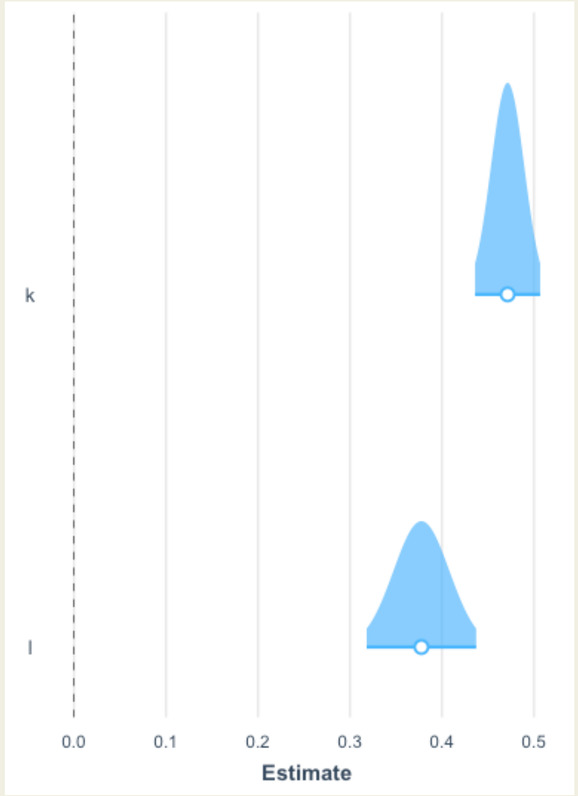
Test results all return high p-values and were therefore not capable to reject the null hypothesis; this allows us to rule out the Between model with confidence.

Last but not least, a final decision had to be made about which of the three Random Effect models we had to choose. Our choice fell on the Time Effect RE after considerating:

- It has the highest R^2 value among the three;
- Captures time effects, the most relevant effect from our initial analysis;
- We can see from model summaries that individual effects accounts for a very tiny share of the effect, so it can be discarded.

5 - Findings & Conclusions

```
Oneway (time) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = re_formula, data = pdata, effect = "time", model = "random")

Balanced Panel: n = 604, T = 5, N = 3020

Effects:
                 var std.dev share
idiosyncratic 0.24598 0.49596 0.879
time          0.03393 0.18419 0.121
theta: 0.8911

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-1.6903235 -0.3412717 -0.0078171  0.3401750  1.4875564

Coefficients:
             Estimate Std. Error z-value Pr(>|z|)
(Intercept) -0.097241   0.082855 -1.1736   0.2405
k            0.471396   0.018068 26.0899   <2e-16 ***
l            0.377689   0.030369 12.4368   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    944.73
Residual Sum of Squares: 741.95
R-Squared:      0.21464
Adj. R-Squared: 0.21412
Chisq: 824.569 on 2 DF, p-value: < 2.22e-16
```



The chosen model shows an R^2 of 0.215 (adjusted: 0.214), with the second lowest value for the Residual Sum of Squares (741).
It shows high significance on all coefficients and the joint test of no significance rejects the null at a high degree of confidence.

Tests regarding the presence of heteroscedasticity and/or no serial correlation all rejected the null, so there is no need to correct the variance-covariance matrix for them.

On the right, we decided to report the regression summary and a plot displaying coefficients estimates and approximate distributions.

Econometrics 2nd Group Assignment: Panel Data Estimation

We thank you for your attention.

Contacts

Luca Colaci - 3176608

luca.colaci@studbocconi.it

Rocco Totaro - 3159435

rocco.totaro@studbocconi.it

Giovanni Maria Parlangeli - 3152699

giovanni.parlangeli@studbocconi.it