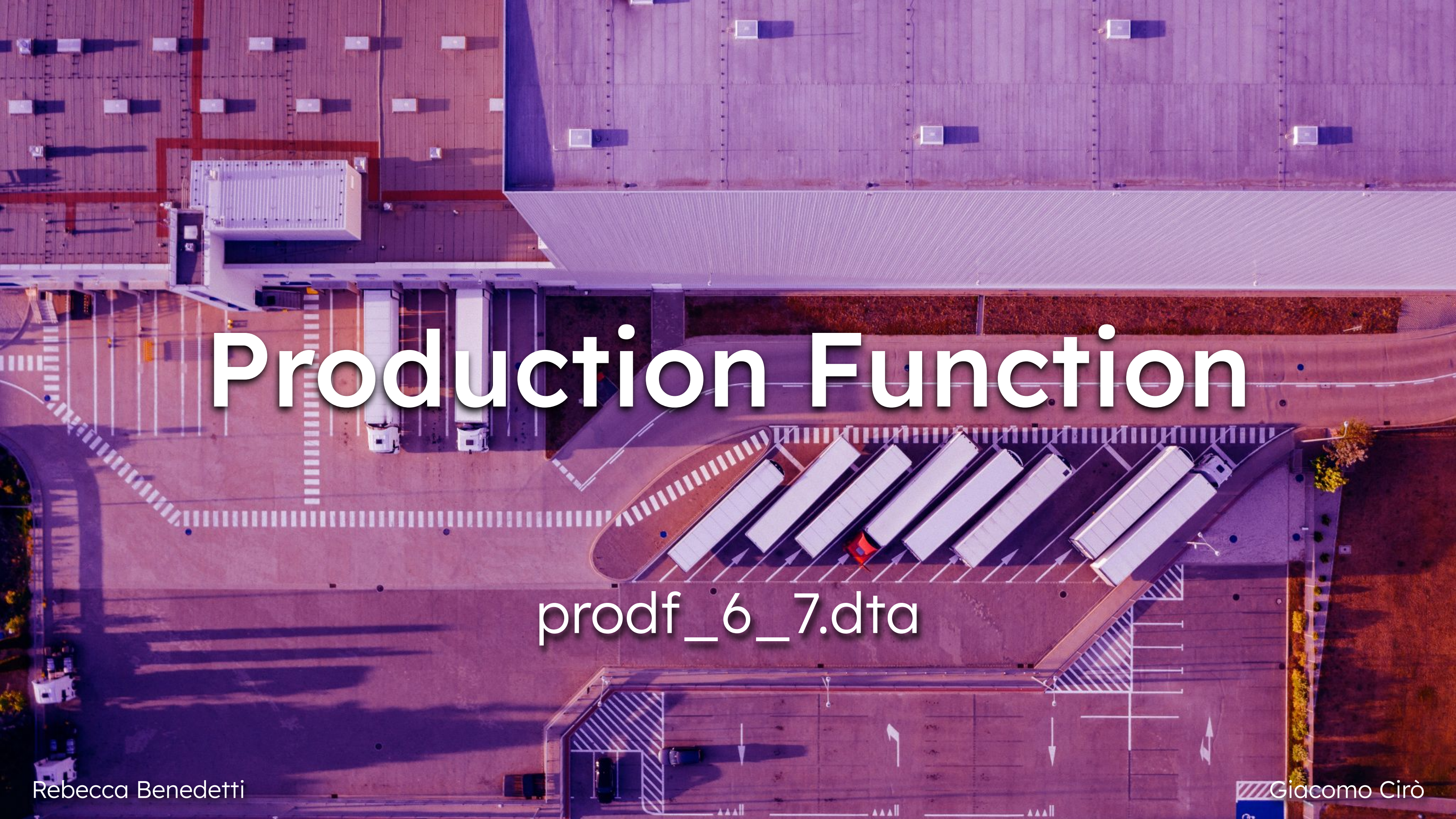


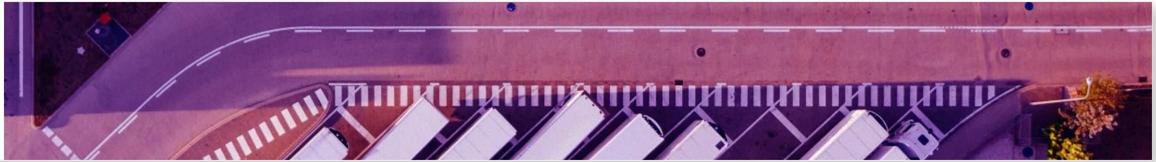
Production Function

prodf_6_7.dta

Rebecca Benedetti

Giacomo Cirò

Introduction



The goal of this project is to estimate a **production function** of the form:

$$Y = L^{\beta_1} \cdot K^{\beta_2}$$

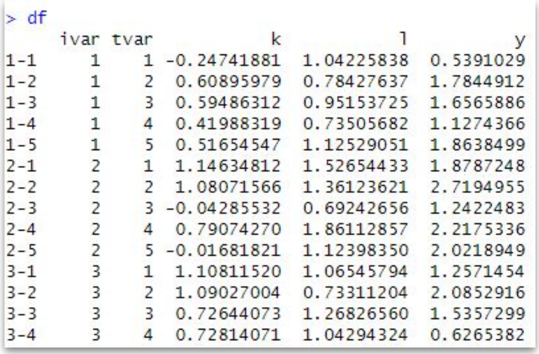
Where Y = **Output**, L = **Labor**, K = **Capital**.

In order to do so, we are going to use the observed values for the variables log(Y), log(L) and log(K) for **604 individuals** over **5 time periods**, thus a total of 3,020 observations.

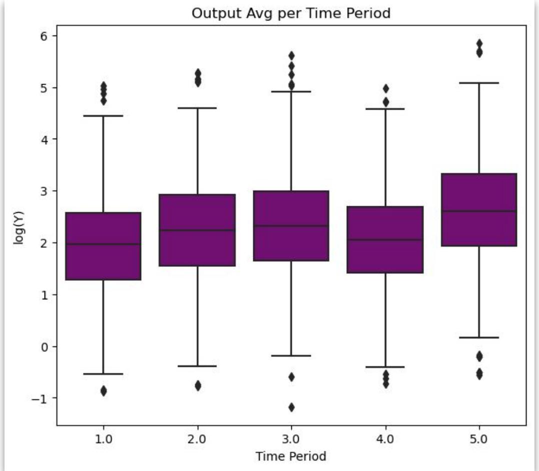
The estimable model is linear and corresponds to the **log-transformed** production function:

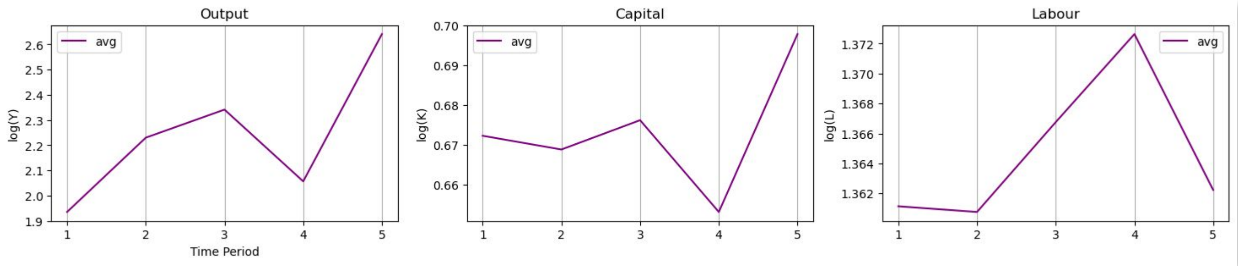
$$log(Y) = \beta_1 log(L) + \beta_2 log(K)$$

Only by looking at simple graphs one can notice a correlation between the variables and an overall uptrend of the output during the period.

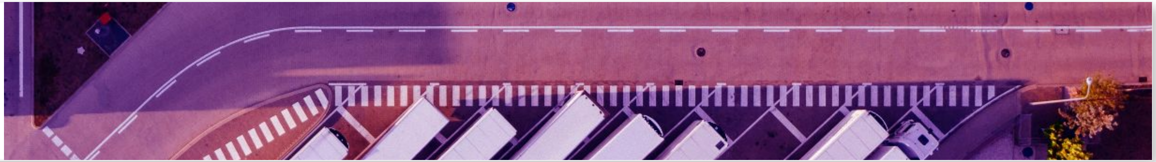


```
> df
    ivar tvar           k          l         y
1-1    1    1 -0.24741881 1.04225838 0.5391029
1-2    1    2  0.60895979 0.78427637 1.7844912
1-3    1    3  0.59486312 0.95153725 1.6565886
1-4    1    4  0.41988319 0.73505682 1.1274366
1-5    1    5  0.51654547 1.12529051 1.8638499
2-1    2    1  1.14634812 1.52654433 1.8787248
2-2    2    2  1.08071566 1.36123621 2.7194955
2-3    2    3 -0.04285532 0.69242656 1.2422483
2-4    2    4  0.79074270 1.86112857 2.2175336
2-5    2    5 -0.01681821 1.12398350 2.0218949
3-1    3    1  1.10811520 1.06545794 1.2571454
3-2    3    2  1.09027004 0.73311204 2.0852916
3-3    3    3  0.72644073 1.26826560 1.5357299
3-4    3    4  0.72814071 1.04294324 0.6265382
```



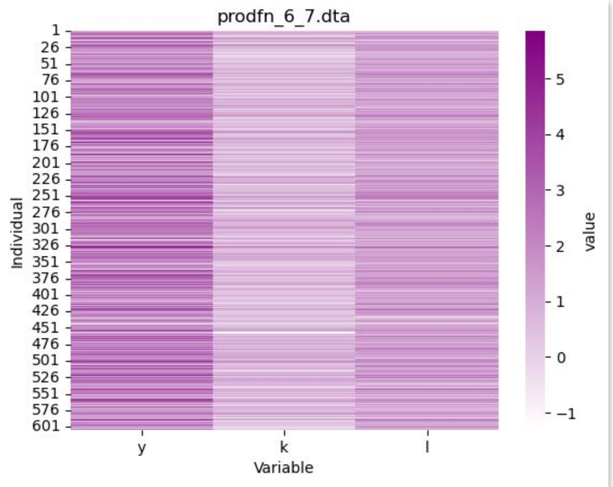


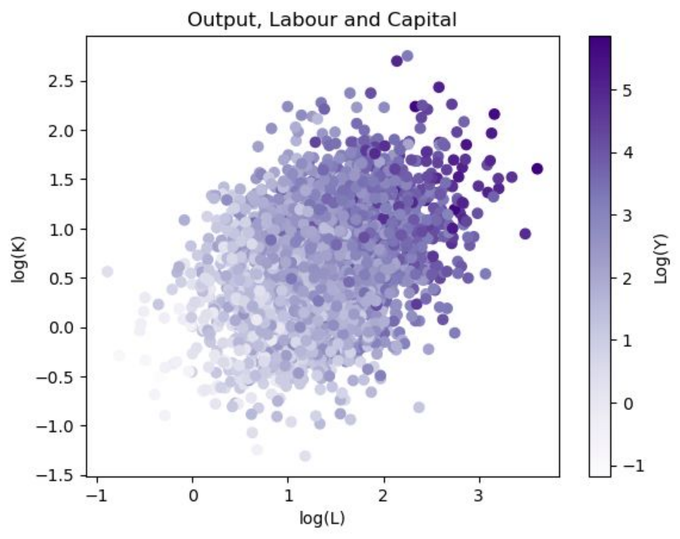
prodfn_6_7.dta

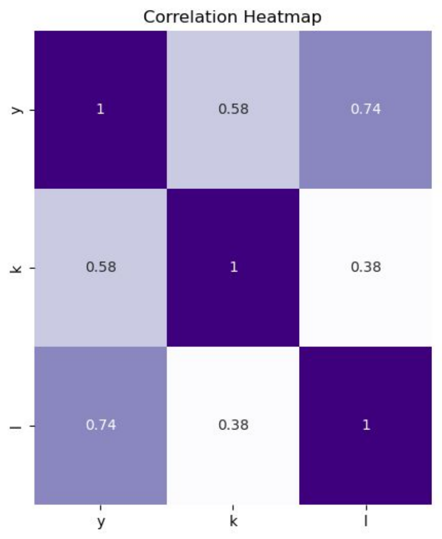


The dataset we are provided with, *prodfn_6_7.dta* , is **balanced** and it does not present missing or null values.

The values range from -1 to 5 for all the three variables.



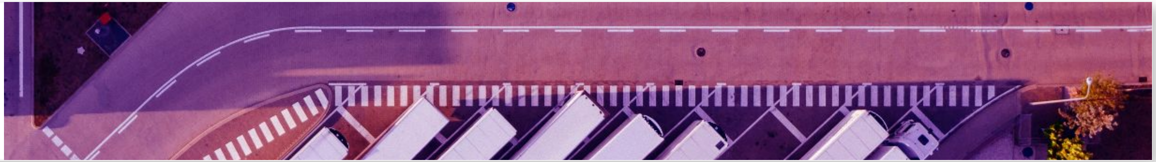




The **scatterplot** and the **correlations** heatmap confirm what was suggested by the initial plots: it is clear that both the regressors are strongly positively correlated with the dependent variable.

Indeed, the color gets darker as we move to the top right corner of the scatterplot, that is the (log of) output increases as both the (log of) labor and (log of) capital increase.

Panel Data Models



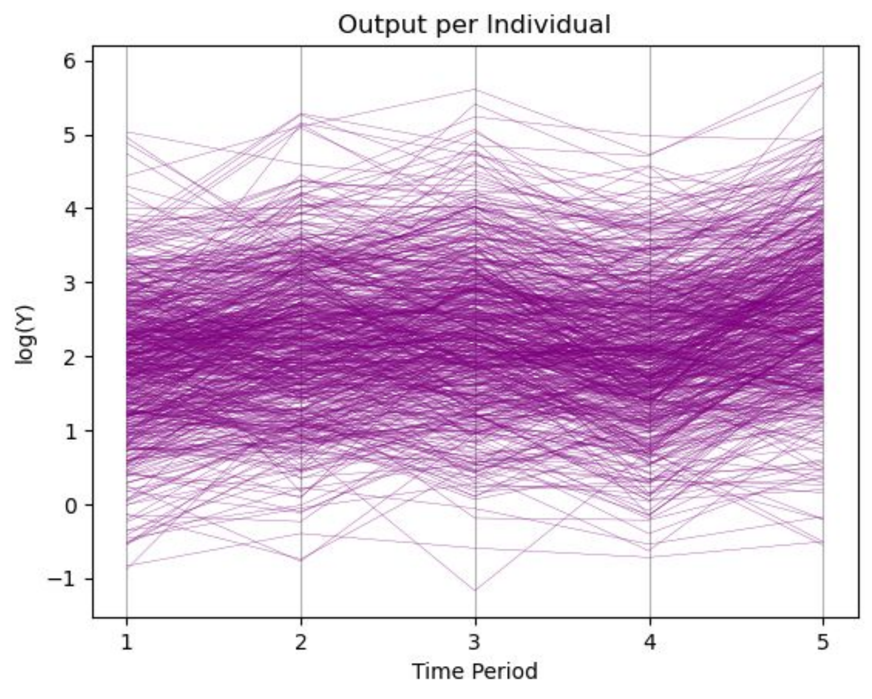
Panel data techniques are powerful methods that allow to extract all the available information from datasets where the observations are indexed with both an individual variable and a time variable, focusing on individual-specific and time-constant **latent heterogeneity** (LH).

Clearly, identifying the correct and optimal model to use is key in order to grasp all the information that such data sets can provide.

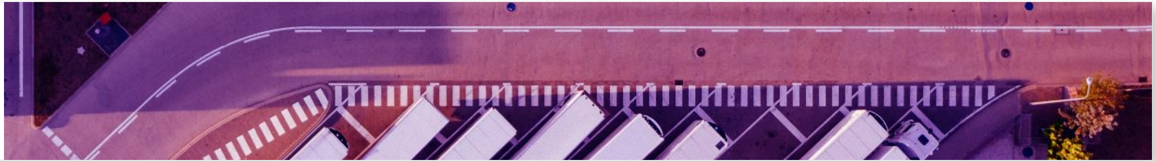
When dealing with panel data, **two** main **models** can be used:

- ***Fixed*** Effect, when no restrictions are imposed on the LH component;
- ***Random*** Effect, when the LH component is assumed to be uncorrelated to the individual and is considered a random group-specific noise.

The plot on the right highlights the value of the output over the time periods for each of the 604 individuals, as previously noticed, a mild uptrend appears.



Fixed Effects



First of all, we ran a classic **one-way** fixed effects model.

As expected, both coefficient estimates are **positive**, which means that a positive increase in the regressor leads to a positive shock in the output, for both capital and labor.

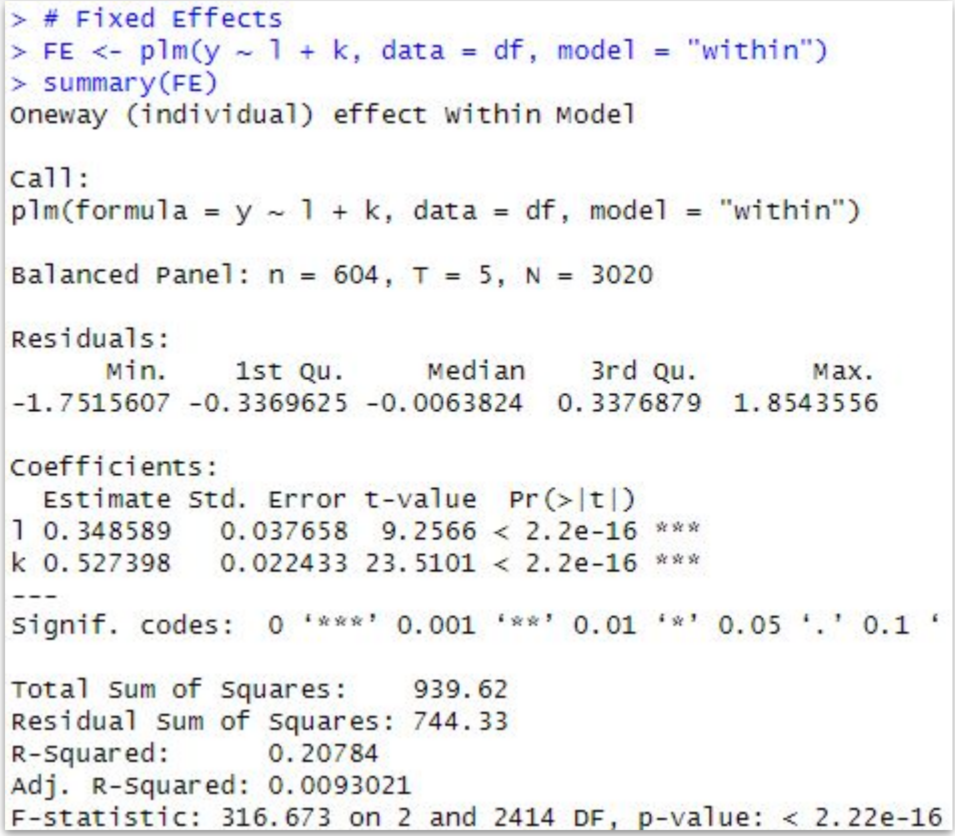
Given the log-log nature of the regression, the coefficients can further be interpreted as the **elasticity**.

In particular, a 1% increase in **capital** leads to a **0.53%** increase in the output and a 1% increase in **labor** results in **0.35%** increase in the output.

Both regressors and the overall model are **statistically significant**. In fact, the probability of each coefficient being equal to zero both singularly or jointly is statistically zero, as reported by the respective **t-tests** and the overall **f-test**.

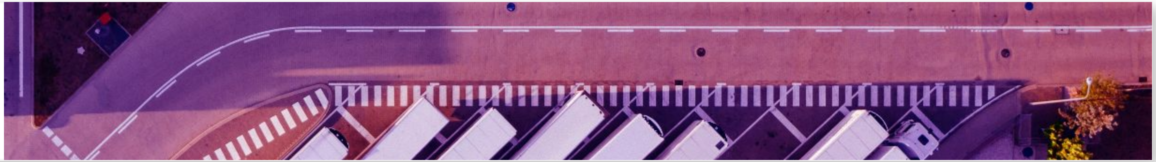
The **R-squared** value is **0.2078**, suggesting that approximately 20.78% of the variability in the dependent variable 'Y' can be explained by the included independent variables within individuals.

The **adjusted** R-squared value, which considers the number of independent variables and the degrees of freedom in the model, is 0.0093021. It provides a more conservative measure of the model's explanatory power.



```
> # Fixed Effects
> FE <- plm(y ~ l + k, data = df, model = "within")
> summary(FE)
Oneway (individual) effect Within Model

Call:
plm(formula = y ~ l + k, data = df, model = "within")

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

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-1.7515607 -0.3369625 -0.0063824  0.3376879  1.8543556

Coefficients:
  Estimate Std. Error t-value  Pr(>|t|)
l 0.348589   0.037658  9.2566 < 2.2e-16 ***
k 0.527398   0.022433 23.5101 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Total Sum of Squares:    939.62
Residual Sum of Squares: 744.33
R-Squared:      0.20784
Adj. R-Squared: 0.0093021
F-statistic: 316.673 on 2 and 2414 DF, p-value: < 2.22e-16
```

Random Effects



The random effects model indicates that a portion of the total variation in Y is due to **idiosyncratic** effects (**0.308**) and **individual** effects (**0.014**).

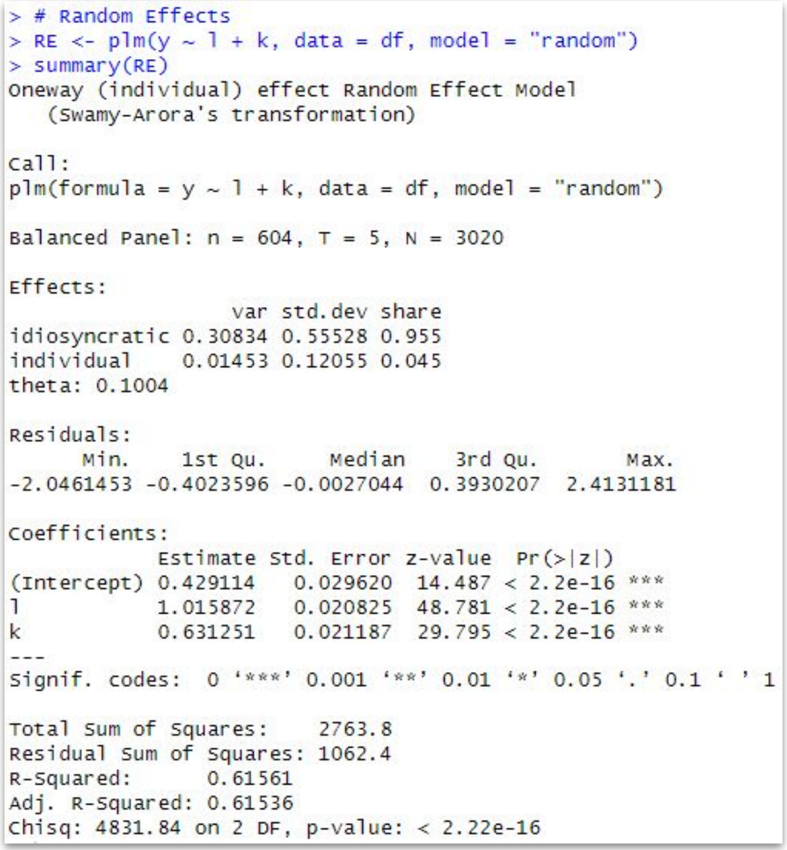
The **proportion** of the total variance attributed to individual-specific effects is estimated to be **0.1004**.

The coefficient estimates are positive: a 1% increase in **labor** leads to a **1.01%** increase in the output and a 1% increase in **capital** results in **0.63%** increase in the output.

Both coefficients and the intercept are **statistically** different from zero at any conventional level of **significance**, as reported by the significance codes (‘***’).

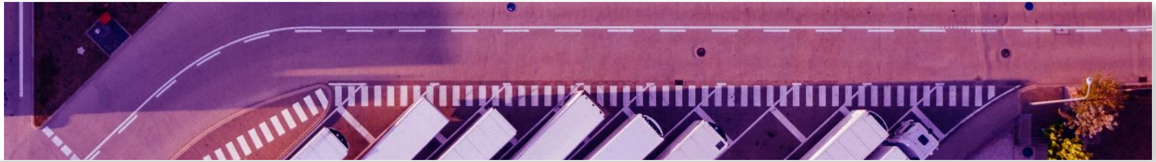
Similarly, the calculated **chi-square** statistic is 4831.84 and the associated p-value is extremely small (< 2.22e-16), therefore the regressors are **jointly significant**.

The **R-squared** value is **0.61561**, indicating that approximately 61.5% of the total variability in ‘Y’ is explained by the independent variables “l” and “k”. The adjusted R-squared value, which takes into account the number of independent variables and the degrees of freedom, is slightly lower (0.61536)



```
> # Random Effects
> RE <- plm(y ~ l + k, data = df, model = "random")
> summary(RE)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = y ~ l + k, data = df, model = "random")

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

Effects:
                  var std.dev share
idiosyncratic 0.30834 0.55528 0.955
individual    0.01453 0.12055 0.045
theta: 0.1004

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-2.0461453 -0.4023596 -0.0027044  0.3930207  2.4131181

Coefficients:
            Estimate Std. Error z-value  Pr(>|z|)
(Intercept) 0.429114   0.029620  14.487 < 2.2e-16 ***
l           1.015872   0.020825  48.781 < 2.2e-16 ***
k           0.631251   0.021187  29.795 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    2763.8
Residual Sum of Squares: 1062.4
R-Squared:      0.61561
Adj. R-Squared: 0.61536
Chisq: 4831.84 on 2 DF, p-value: < 2.22e-16
```

Pooled OLS

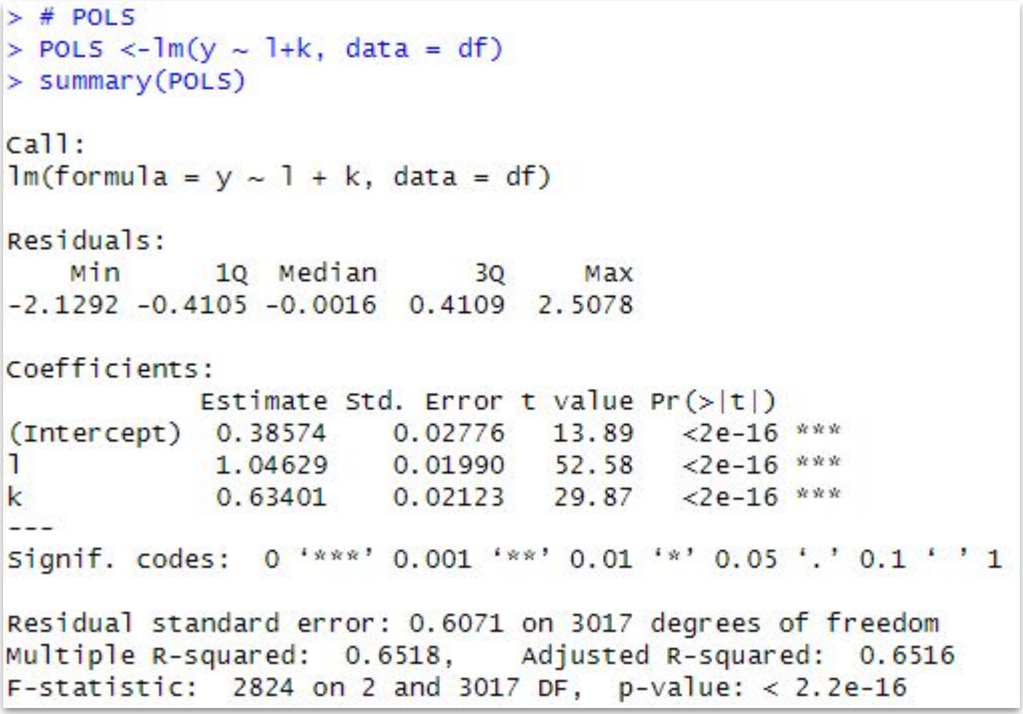


Even though the f-test on the FE suggests the presence of **LH**, we implemented the pooled OLS model, which assumes that the unobserved individual effects are **absent** or not relevant.

As in the FE and Re models, both coefficient estimates are positive: a unit percentage increase in **labor** leads to a **1.04%** increase in the output and a unit percentage increase in **capital** results in **0.63%** increase in the output.

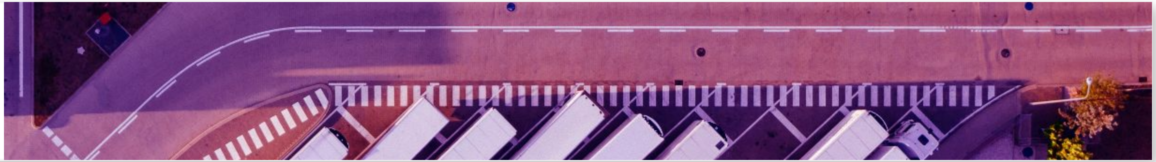
As indicated by the respective **t-tests** and **f-test**, the probability of each coefficient being equal to zero both singularly or jointly is **statistically** zero, hence the model is **significant** and so are the two regressors.

The **R-squared** value is **0.651**, which means that approximately 65.1% of the variance in the dependent variable 'Y' is explained by the independent variables.



```
> # POLS
> POLS <-lm(y ~ l+k, data = df)
> summary(POLS)

Call:
lm(formula = y ~ l + k, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1292 -0.4105 -0.0016  0.4109  2.5078 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.38574    0.02776   13.89   <2e-16 ***
l            1.04629    0.01990   52.58   <2e-16 ***
k            0.63401    0.02123   29.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6071 on 3017 degrees of freedom
Multiple R-squared:  0.6518,	Adjusted R-squared:  0.6516 
F-statistic:  2824 on 2 and 3017 DF,  p-value: < 2.2e-16
```

Two-way FE

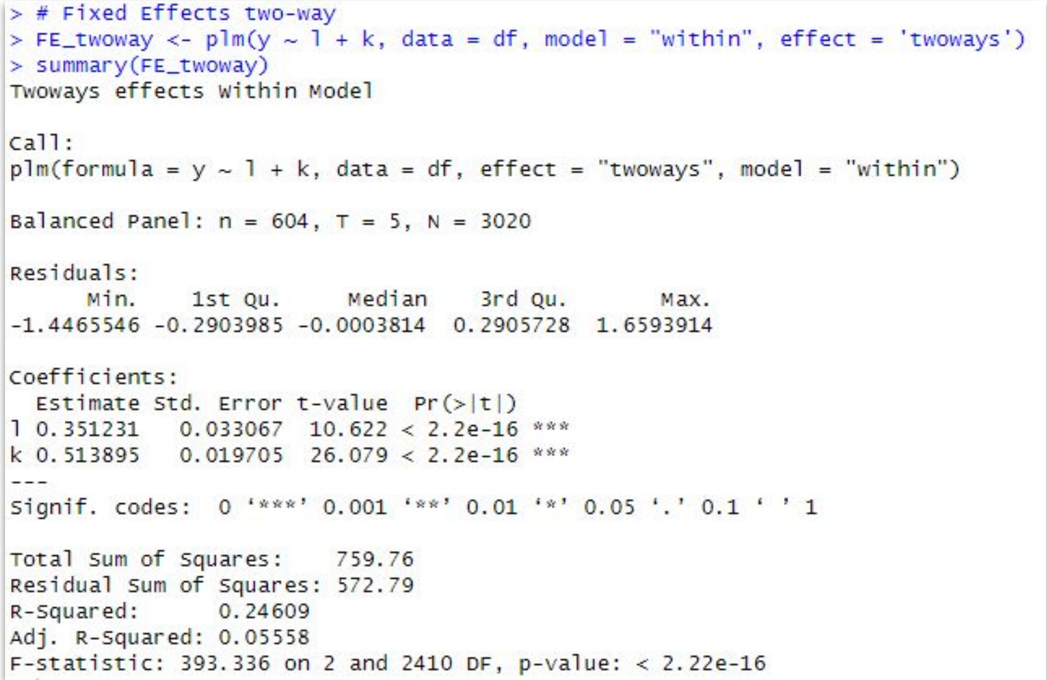


We implemented the two-way fixed effect model and noticed that the portion of total variation captured by the model increased, suggesting the presence of **time**-constant **effects** as well.

Both the coefficient estimates are still **positive** and similar to the previously estimated ones.

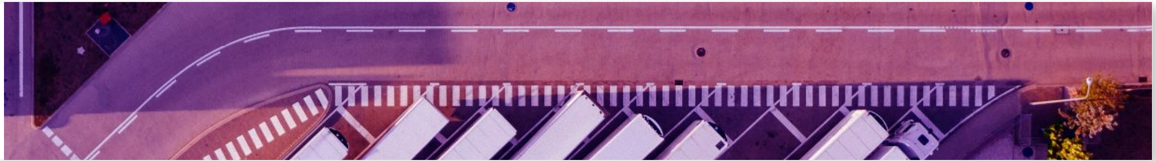
This model is **statistically significant** as shown by the f-test and the t-tests.

The **R-squared** value is **0.24609,** a slight improvement compared to the previous 0.2078 of the one-way fixed effect model.

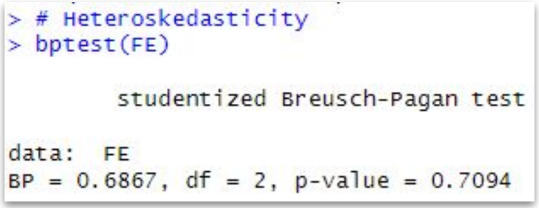


```
> # Fixed Effects two-way
> FE_twoway <- plm(y ~ l + k, data = df, model = "within", effect = 'twoways')
> summary(FE_twoway)
Twoways effects Within Model

Call:
plm(formula = y ~ l + k, data = df, effect = "twoways", model = "within")

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

Residuals:
     Min.    1st Qu.     Median    3rd Qu.      Max. 
-1.4465546 -0.2903985 -0.0003814  0.2905728  1.6593914 

Coefficients:
  Estimate Std. Error t-value  Pr(>|t|)    
l 0.351231   0.033067  10.622 < 2.2e-16 ***
k 0.513895   0.019705  26.079 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    759.76
Residual Sum of Squares: 572.79
R-Squared:      0.24609
Adj. R-Squared: 0.05558
F-statistic: 393.336 on 2 and 2410 DF, p-value: < 2.22e-16
```

Diagnostic Checks



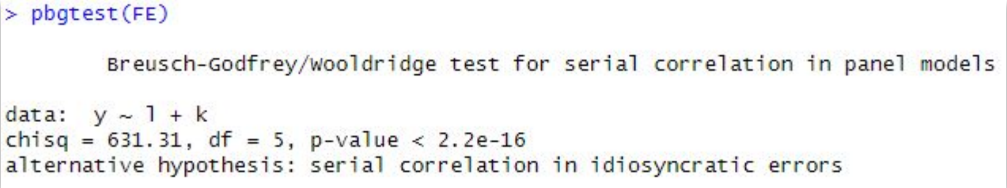
In order to choose between RE and FE, a Hausman-Test is used. However, it is appropriate to **check** beforehand the validity of assumption **FE-3** and **RE-3**, that is homoskedasticity and uncorrelation of the error term.



```
> # Heteroskedasticity
> bptest(FE)

	studentized Breusch-Pagan test

data:  FE
BP = 0.6867, df = 2, p-value = 0.7094
```

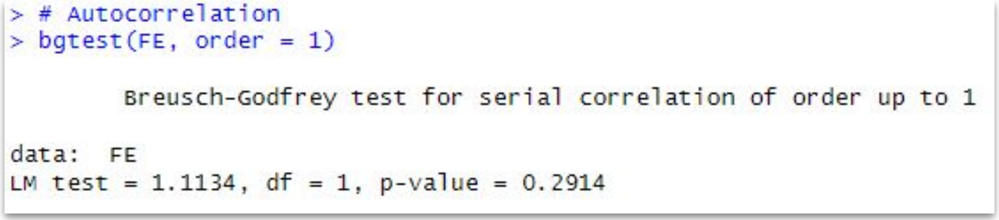
The **Breusch-Pagan** test fails to reject the null hypothesis of **homoskedasticity** due to the high p-value.

The **Wooldridge** Test rejects the null of no **serial correlation** and suggests the presence of serial correlation in idiosyncratic errors. However, this could be caused by the low number of observation per time period, which being only 5 might happen to be randomly serially correlated.



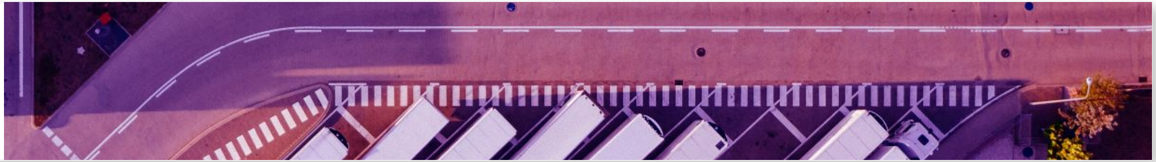
```
> pbgtest(FE)

	Breusch-Godfrey/Wooldridge test for serial correlation in panel models

data:  y ~ l + k
chisq = 631.31, df = 5, p-value < 2.2e-16
alternative hypothesis: serial correlation in idiosyncratic errors
```

To confirm our hypothesis, we decided to ran a test for serial correlation in the **overall** model, and as we thought the null of serial correlation is strongly rejected, suggesting that the previous result might have been biased.

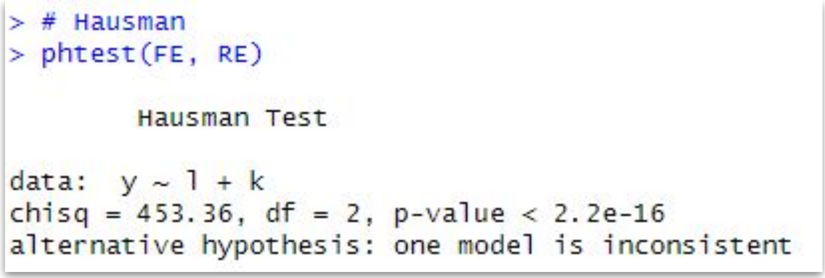


```
> # Autocorrelation
> bgtest(FE, order = 1)

	Breusch-Godfrey test for serial correlation of order up to 1

data:  FE
LM test = 1.1134, df = 1, p-value = 0.2914
```

Hausman Test

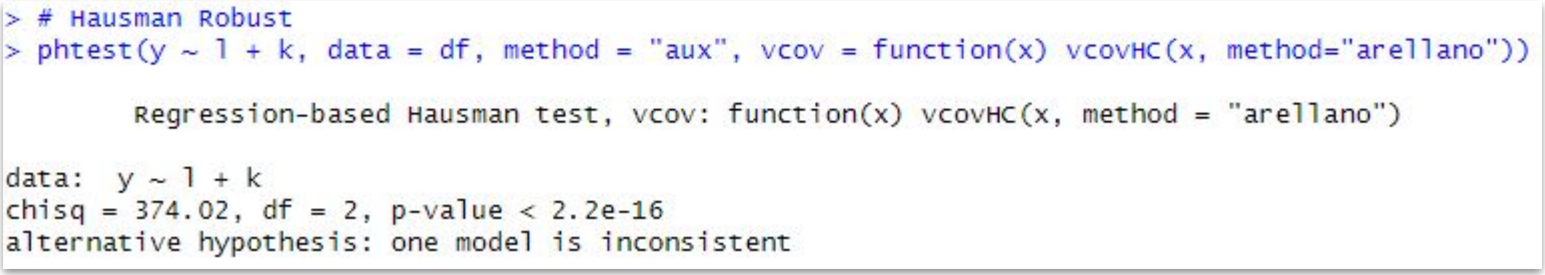


Given the previous ambiguous results, we decided to run both a **traditional** Hausman test and a **robustified** version. The latter implies an auxiliary regression and a robust **VCE**, in this case we chose **Arellano's** which is robust to both heteroskedasticity and correlation.



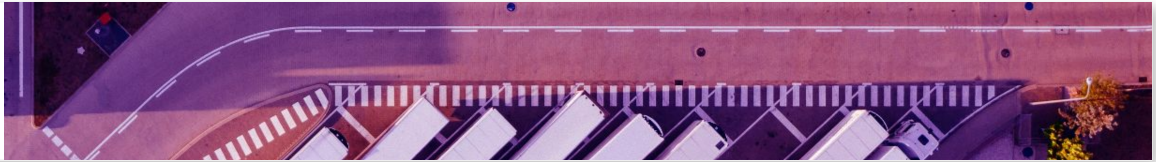
```
> # Hausman
> phtest(FE, RE)

	Hausman Test

data:  y ~ l + k
chisq = 453.36, df = 2, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent
```

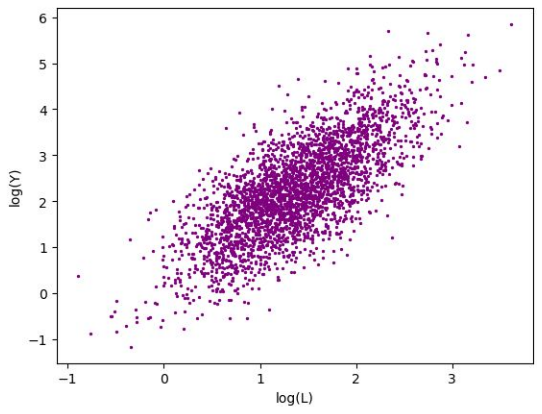
The statistically **zero p-value** in both tests strongly rejects the null hypothesis of uncorrelation between the latent heterogeneity component and the regressors, thus suggesting that the RE model is inconsistent and the **FE** shall be the one of **choice**.



```
> # Hausman Robust
> phtest(y ~ l + k, data = df, method = "aux", vcov = function(x) vcovHC(x, method="arellano"))

	Regression-based Hausman test, vcov: function(x) vcovHC(x, method = "arellano")

data:  y ~ l + k
chisq = 374.02, df = 2, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent
```

Conclusion





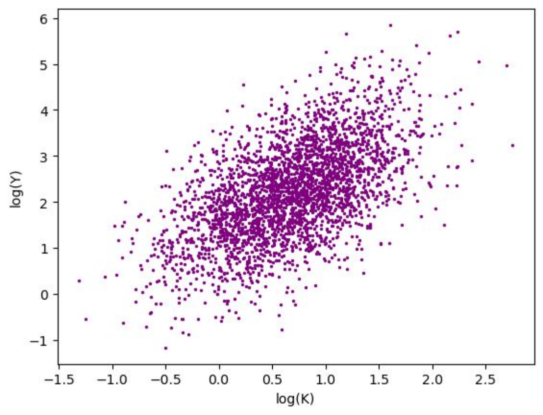
This project aimed at predicting a log-transformed production function, using the available panel data.

To do so, we initially developed the two main panel data models, namely the Fixed Effects and Random Effects.

As suggested by the robust **Hausman** test we decided to discard the RE model as the latent heterogeneity component was statistically correlated to the individual regressors. Furthermore, by comparing the RE results with the estimated POLS ones, it is clear that using the former model does not bring useful improvements, as it is not that much better than just ignoring the LH component.
We conclude that the latent heterogeneity component in the output is due to individual specific effects (e.g. managerial ability, know-how etc. in this specific production function scenario) rather than just to random noise, thus the **FE** model is more **appropriate**.

Additionally, by running a two-way regression, we discovered that there are **time** constant **effects** which are statistically relevant and taking them into accounts helps explaining the variation in the output.
In reality and for this particular scenario, this might be due to, for example, industry-wide shocks or macroeconomic crisis, which are specific to a time period and impact all the firms, rather than just a single individual.

After having observed the different models and drawn the foregoing conclusions, we deemed the **two-way FE** to be the more appropriate model for estimating a production function given our dataset.


log(Y)
log(K)