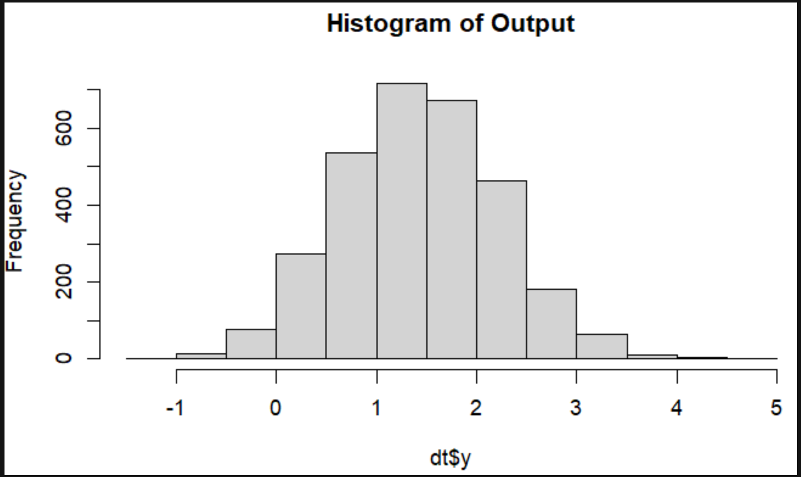
Production function estimation

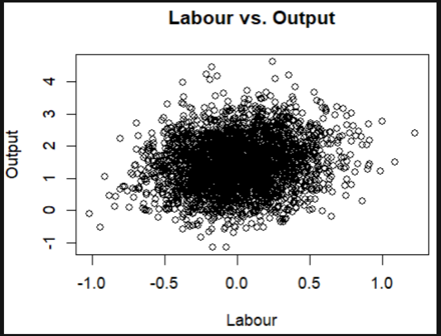
Costanza D’Ercole- 3159923
Francesco Iaccarino – 3170051
Davide Romano - 3164081

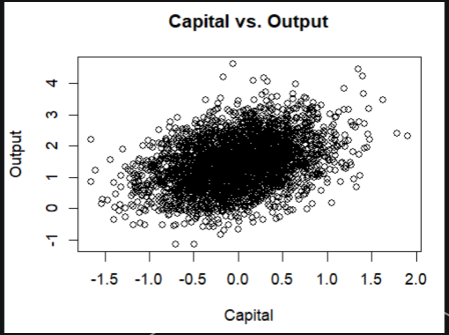
Initial visualization of the variables

Our initial panel data is composed of 604 observations, each one recorded in 5 different time periods.
The independent variables we have are: k (In of capital) and l (In of labour).
We wish to estimate y (In of output).

By plotting an histogram of the output, we notice that is seems to be normally distributed, this is due to the fact that the output is transformed in logarithmic scale:

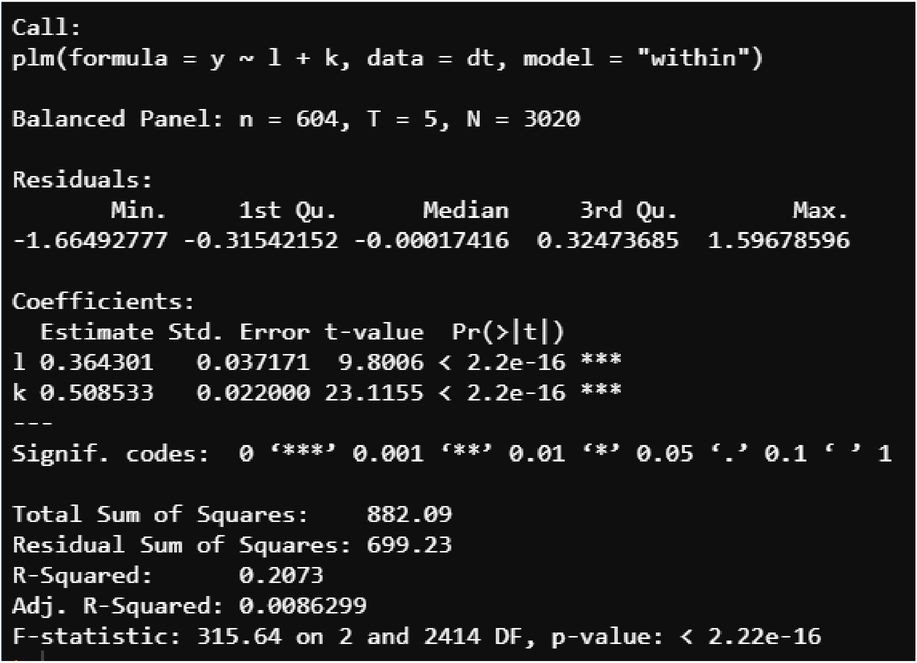






FE model

The first model we will try is a Fixed Effect model, which assumes the presence of individual LH components, correlated with X but not with the error term.



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

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

Residuals:
       Min.      1st Qu.       Median      3rd Qu.         Max. 
-1.66492777 -0.31542152 -0.00017416  0.32473685  1.59678596 

Coefficients:
  Estimate Std. Error t-value  Pr(>|t|)    
l 0.364301   0.037171  9.8006 < 2.2e-16 ***
k 0.508533   0.022000 23.1155 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    882.09
Residual Sum of Squares: 699.23
R-Squared:      0.2073
Adj. R-Squared: 0.0086299
F-statistic: 315.64 on 2 and 2414 DF, p-value: < 2.22e-16
```

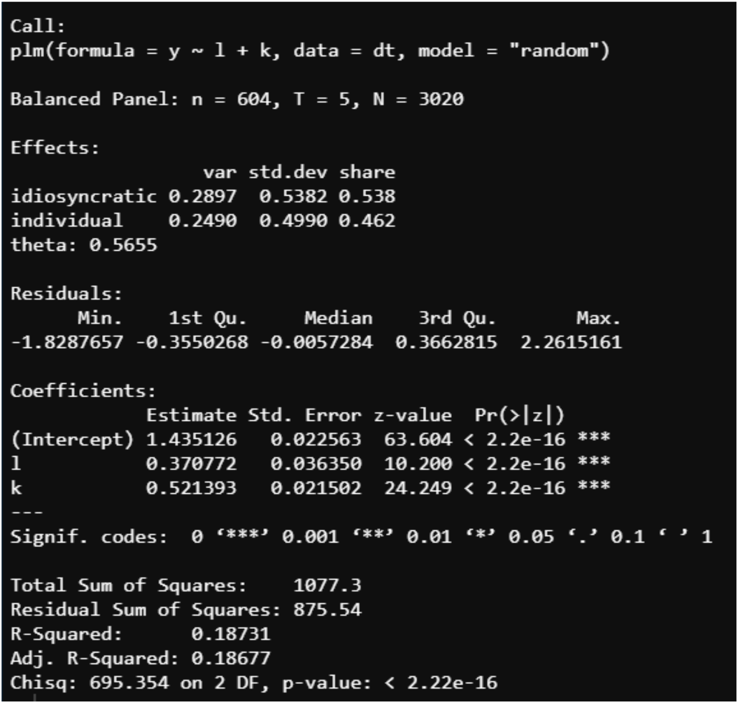
From the output of the regression we can notice:

- Both the variables l and k are highly significant in the model, due to the small p-value.
- Since the Adjusted R-squared is much lower than the R-squared, we may think that the model is overfitting, or we've included unnecessary variables.
- The p-value of the F-statistic is really small: the model is statistically significant.

RE model

Now, we will try to use a Random Effect model: we now assume that the infividual-specific effects are correlated with both the explanatory variables and the error term.

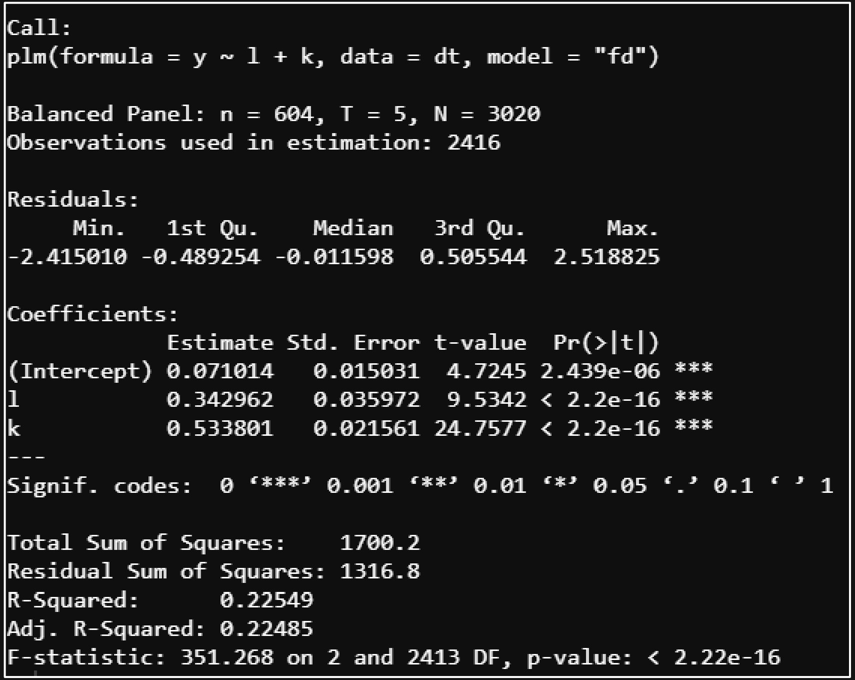
- The value of Theta, the estimated variance component in the model, is positive, implying the presence of heterogeneity across individuals, which affect the dependent variable.
- The variables l and k are still highly relevant to describe y.
- The R-squared is almost equal to the Adjusted R-squared, meaning that the variables included in the model explain well y, without overfitting or the need of adding additional variables.
- The result of the Chi-Squared test shows a high significance of the model, that strongly rejects the null hypothesis of the model having no relationship with the dependent variable.



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

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

Effects:
                 var std.dev share
idiosyncratic 0.2897  0.5382 0.538
individual    0.2490  0.4990 0.462
theta: 0.5655

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-1.8287657 -0.3550268 -0.0057284  0.3662815  2.2615161

Coefficients:
            Estimate Std. Error z-value  Pr(>|z|)
(Intercept) 1.435126   0.022563  63.604 < 2.2e-16 ***
l           0.370772   0.036350  10.200 < 2.2e-16 ***
k           0.521393   0.021502  24.249 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    1077.3
Residual Sum of Squares: 875.54
R-Squared:      0.18731
Adj. R-Squared: 0.18677
Chisq: 695.354 on 2 DF, p-value: < 2.22e-16
```

First-Differenced model

By using this model, we are removing from the regression individual-specific effects, leaving only within individual changes.



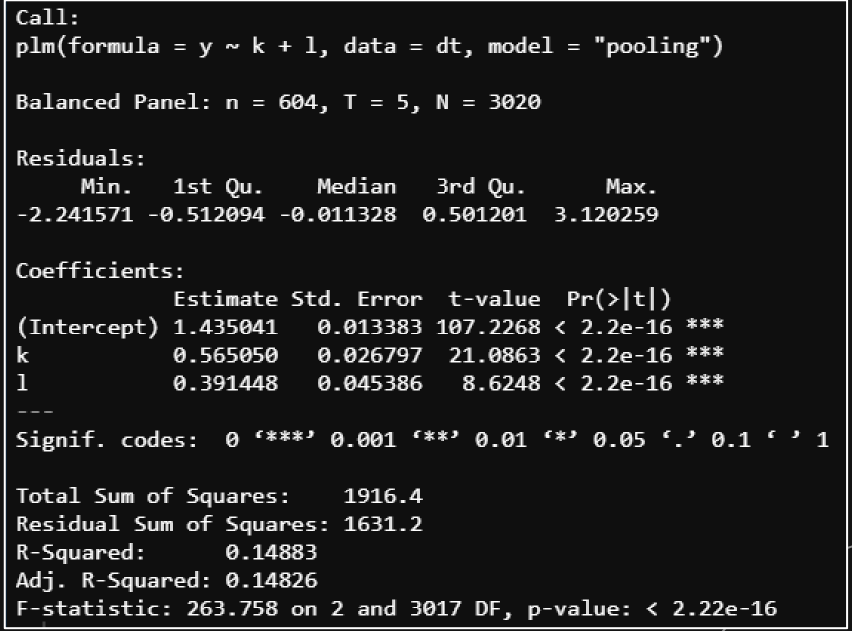
```
Call:
plm(formula = y ~ l + k, data = dt, model = "fd")

Balanced Panel: n = 604, T = 5, N = 3020
Observations used in estimation: 2416

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-2.415010 -0.489254 -0.011598  0.505544  2.518825

Coefficients:
            Estimate Std. Error t-value  Pr(>|t|)
(Intercept) 0.071014   0.015031  4.7245 2.439e-06 ***
l           0.342962   0.035972  9.5342 < 2.2e-16 ***
k           0.533801   0.021561 24.7577 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    1700.2
Residual Sum of Squares: 1316.8
R-Squared:      0.22549
Adj. R-Squared: 0.22485
F-statistic: 351.268 on 2 and 2413 DF, p-value: < 2.22e-16
```

- Also in this case, all the independent variables are highly significant in the model, with extremely small p-values.
- Also, similar to the FE model, the Adjusted R-squared is close to the R-squared, meaning that in general this is a good fit for our data.
- The p-value of the F-statistic is small: the model is statistically significant, and we reject the null hypothesis of all coefficients of the explanatory variables set to 0.

POLS model

Next is the POLS regression. It assumes that the individual-specific and time-specific effects are constant across individuals and time periods.

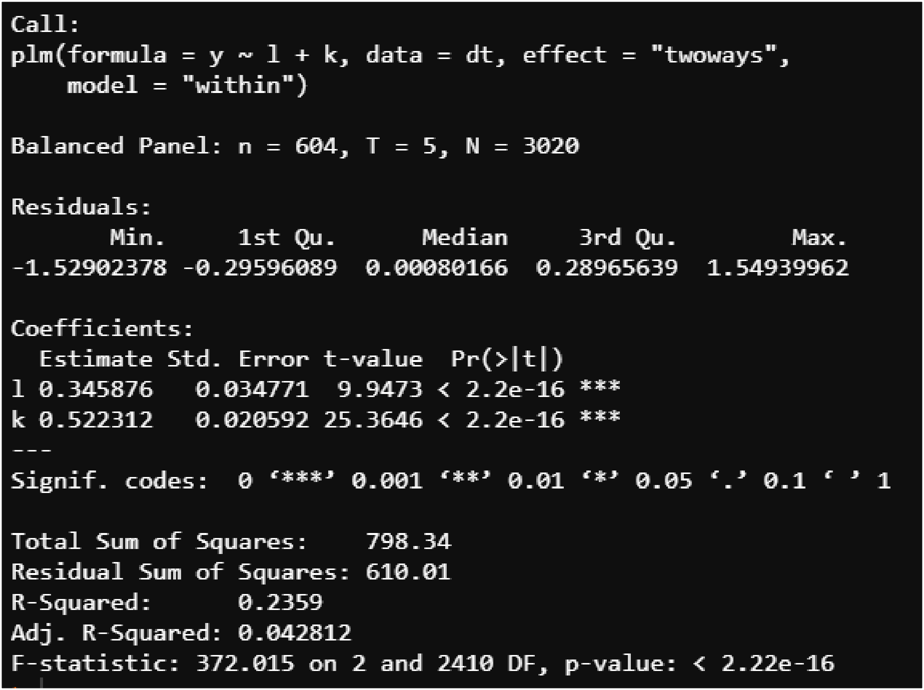
- The explanatory variables are all relevant in regression.
- The R-squared is almost equal to the Adjusted R-squared, so the variables included in the model explain well y, without overfitting or the need of adding additional variables.
- The result of the F-test shows a high significance of the model.



```
Call:
plm(formula = y ~ k + l, data = dt, model = "pooling")

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

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-2.241571 -0.512094 -0.011328  0.501201  3.120259

Coefficients:
            Estimate Std. Error  t-value  Pr(>|t|)
(Intercept) 1.435041   0.013383 107.2268 < 2.2e-16 ***
k           0.565050   0.026797  21.0863 < 2.2e-16 ***
l           0.391448   0.045386   8.6248 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    1916.4
Residual Sum of Squares: 1631.2
R-Squared:      0.14883
Adj. R-Squared: 0.14826
F-statistic: 263.758 on 2 and 3017 DF, p-value: < 2.22e-16
```

Two-way FE

In this model, we include fixed effects for both individuals and time periods. So, we consider the heterogeneity in individual-specific and time-specific effects, allowing for more accurate estimation.

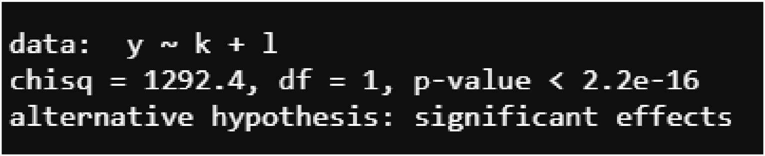


```
Call:
plm(formula = y ~ l + k, data = dt, effect = "twoways",
    model = "within")

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

Residuals:
       Min.     1st Qu.      Median     3rd Qu.        Max.
-1.52902378 -0.29596089  0.00080166  0.28965639  1.54939962

Coefficients:
  Estimate Std. Error t-value  Pr(>|t|)
l 0.345876   0.034771  9.9473 < 2.2e-16 ***
k 0.522312   0.020592 25.3646 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    798.34
Residual Sum of Squares: 610.01
R-Squared:      0.2359
Adj. R-Squared: 0.042812
F-statistic: 372.015 on 2 and 2410 DF, p-value: < 2.22e-16
```

- Both l and k are highly significant in the model, due to their small p-value.
- The Adjusted R-squared is much lower than the R-squared, so the model could be overfitting, or we’ve included unnecessary variables.
- The p-value of the F-statistic is really small: the model is statistically significant.

LM test

The following LM test is used to compare POLS and RE model.
The null hypothesis sets the individual-specific effects to 0. If H0 is rejected, the individual-specific effects are present, and so the RE model is preferred.



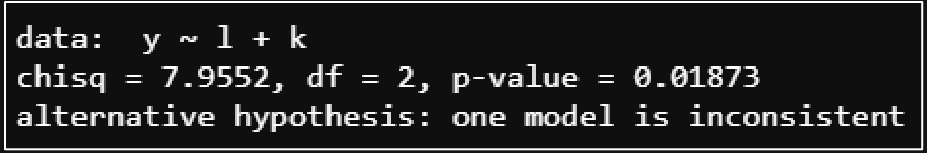
```
data:  y ~ k + l
chisq = 1292.4, df = 1, p-value < 2.2e-16
alternative hypothesis: significant effects
```

The result of the test shows a p-value smaller than 0.05, so, we reject the null hypothesis. This means that the individual-specific effects are present, and the RE model is preferred over a simple OLS on panel data.

→ RE better than POLS

Hasuman test

To compare FE with RE, we use the Hasuman test.
In the null hypothesis we assume that the individual-specific effects are uncorrelated with the regressors. So, if HO is rejected, we prefer the Fixed Effect model over the Random Effect.



```
data:  y ~ l + k
chisq = 7.9552, df = 2, p-value = 0.01873
alternative hypothesis: one model is inconsistent
```

The p-value is lower than 5%. So, we will reject the null hypothesis, implying that the individual effects are correlated with the regressors. So, we prefer FE model over RE model.
This choice is due to the fact that the fixed effect model accounts for this type of correlation, and so it is a better fit for our data.

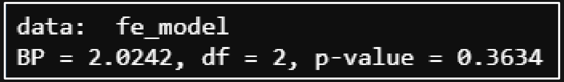
By running the test also between RE and FD, and Two-Way FE, the result is always the same: random effect model is not the right choice for our data.

→ FE, FD, Two-Way FE better than RE

Breush-Pagan test

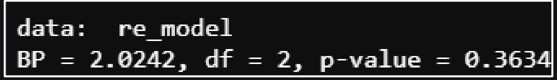
Breush-Pagan tests for heteroskedasticity in the model's residuals.
It tests the null of no heteroskedasticity in the model against the alternative of heteroskedasticity in the residuals.

FE model



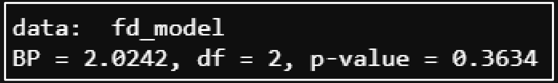
```
data:  fe_model
BP = 2.0242, df = 2, p-value = 0.3634
```

RE model



```
data:  re_model
BP = 2.0242, df = 2, p-value = 0.3634
```

FD model



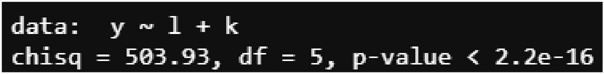
```
data:  fd_model
BP = 2.0242, df = 2, p-value = 0.3634
```

The above are the outputs of the test performed on 3 out of the 5 models. The output of the tests is the same for each one: since the p-values are high, we do not reject the null hypothesis, and we understand that the residuals' variances are the same across observations.
This means that we do not need to adjust the VCE by using robust strategies to account for heteroskedasticity.

Also, the presence of homoskedasticity does not depend on the model used, since the output of the tests is the same.

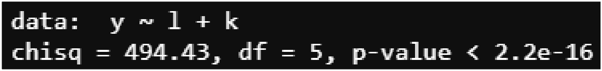
Wooldridge test

The Breush Pagan/ Wooldridge test is used to test for serial correlation in the panel data model.
The null hypothesis is that there is no serial correlation.
By performing the test on the 3 “best models”, the results are:



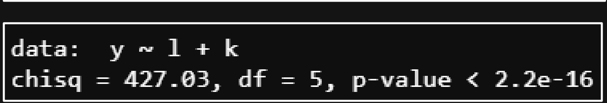
```
data:  y ~ l + k
chisq = 503.93, df = 5, p-value < 2.2e-16
```

→ FE model



```
data:  y ~ l + k
chisq = 494.43, df = 5, p-value < 2.2e-16
```

→ FD model



```
data:  y ~ l + k
chisq = 427.03, df = 5, p-value < 2.2e-16
```

→ Two-Way model

The p-values of all the model are really small: we strongly reject the null hypothesis of no serial correlation. The above result could also be “predicted” because of the result of the Hasuman test: the FE models account for serial correlation by mitigating their effect, while RE models do not perform well in the presence of serial correlation.

Final choice

At this point, we have to choose the overall best model among FE, FD, and Two-Way. There are different considerations we can make:

a. **R Squared / Adjusted R squared**: based on the result of the R squared, the most significant model seems to be the Two-Way FE. However, its adjusted R squared is much lower, which is not a positive sign. The FD model seems more "stable", but it may not be a good choice since it removes individual-specific factors from the regression.

b. Still, the Two-Way FE seems to be a good choice given our panel data, since it keeps into consideration both the individual-specific and the time-specific effects. We can in fact imagine that a firm's production can depend also on the year/period considered, since there may be technological innovations or other time-dependent factors that affect it.

c. However, also the FE model has a similar R squared compared to the Two-Way FE: choosing it when time-specific effect seems not to be important reduces the overall complexity of the model, meaning in a more probable correct regression.

→ Two-Way FE best model.

THANK YOU