R Project

Introduction: Econometrics applies statistical and mathematical models to datasets to create theories or test hypotheses in economics and forecast future developments based on historical evidence. Econometricians are those who participate in this research [1].We use three R packages for our analysis that is plm() for data, readr() for import and read dataset from the csv file, and then, after all, lmtest() for the linear model.

```
library(plm)
library(readr)
library(lmtest)
```

Panel Data: Panel data, also referred to as longitudinal data, is a set of cross-sections collected over time. Countries, corporations, individuals, and demographic groups are examples of groups included in a panel data set [2].

Dataset Description: Grunfeld's Data, which can be found in the R package plm, is used in the following examples (). As a result, we must first install plm packages in RStudio before using the library () function to call packages [3]. The dataset has totaled 200 observations of a production unit of the **USA**, and this data is perfect for the Econometrics model, Panel data analysis using RStudio. The dataset also can be found in R packages that are plm().

Variables Description

A data frame containing :

firm

observation

year

date

inv

gross Investment

value

value of the firm

capital

stock of plant and equipment

Objectives of the study: The study of economic interactions is referred to as econometrics. It combines economics, panel dataset, and statistics to assign numerical values to financial interaction parameters. The relationship between variables is explored in this analysis using different econometrics models and a panel dataset.

Literature review: Grunfeld, who died in a drowning accident at the age of 30 in the second half of the 1950s, was a very promising applied econometrician, as his signature on the page of his Ph.D. thesis (Grunfeld, 1958) pertaining to reproduction rights indicates. Patinkin, 1961; Goodman and Grunfeld, 1961). He provides panel information on a select number of large US firms from 1935 to 1954 in an appendix to his research at the University of Chicago, titled "The Determinants of Corporate Investment." Since his untimely death, these figures have been used to demonstrate multiple-equation and panel data methodology in research and teaching. Panel details on a select number of large US firms from 1935 to 1954 are included in an appendix to his thesis at the University of Chicago, titled "The Determinants of Corporate Investment." Since his untimely death, his numbers have been used in science and teaching to demonstrate multiple-equation and panel evidence methodology.

Read dataset: Import required data from working directory.

```
dataset <- read_csv("data.csv")
```

Summary statistics for the variables used in the analysis:

```
##        X1              firm           year           inv
## Min.:  1.00   Min.: 1.0   Min.   :1935   Min.:   0.93
## 1st Qu.: 50.75   1st Qu.: 3.0   1st Qu.:1940   1st Qu.:  33.56
## Median :100.50   Median: 5.5   Median :1944   Median:  57.48
## Mean   :100.50   Mean: 5.5   Mean   :1944   Mean: 145.96
## 3rd Qu.:150.25   3rd Qu.: 8.0   3rd Qu.:1949   3rd Qu.: 138.04
## Max.   :200.00   Max.   :10.0   Max.   :1954   Max.   :1486.70
##      value            capital
## Min.:  58.12   Min.:   0.80
## 1st Qu.: 199.97   1st Qu.:  79.17
## Median: 517.95   Median: 205.60
## Mean   :1081.68   Mean: 276.02
## 3rd Qu.:1679.85   3rd Qu.: 358.10
## Max.   :6241.70   Max.   :2226.30
```

Trending variables graphs analysis: The following graph shows the trend of gross Investment with respect to the year.

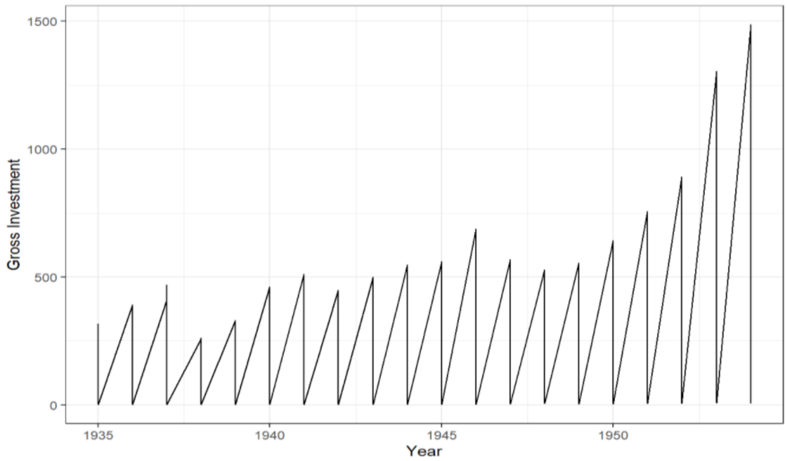

Fig.: Visualization of gross Investment concerning year.

Differences (DID): The difference-in-differences method is a quasi-experimental technique that compares the changes in results over time between a treatment group and a control group (the comparison group). It's a valuable data analysis tool.

DID is the performance of our dataset model, and it contains the minimum, maximum, quartile, and median residual values. The p-value for the linear model is 2e-16***, which is less than.05, indicating a significant relationship between the two variables, gross investment and money, which are correlated with 198 degrees of freedom. Also, we can see the minimum value, first quartile, median,3rd quartile, and maximum value of residuals as we know **residual** is the distance between original data and the **straight-line difference.**

```
## Call:
## lm(formula = inv ~ capital, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -316.92  -96.45  -14.43   17.07  481.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.23620   15.63927    0.91    0.364
## capital      0.47722    0.03834   12.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 162.9 on 198 degrees of freedom
## Multiple R-squared:  0.439, Adjusted R-squared:  0.4362
## F-statistic: 154.9 on 1 and 198 DF, p-value: < 2.2e-16
```

Deal with Heteroskedasticity: Checking for the presence of heteroskedasticity is also possible in panel settings. The null hypothesis in the Breusch-Pagan test for heteroskedasticity is that homoskedasticity occurs. The test is written in the bptest() feature of the lmtest bundle.

First-difference Estimator: Specifying model = "fd" in function plm is another way of estimating a FE model (). In comparison to the LSDV method and within-groups estimator, the capital coefficient is now different. This is because the first-differenced model's coefficients and standard errors are similar to the previously obtained results when there are two-time intervals. Both the coefficients and the common mistakes would be different for more extended time series.

```
## One way (individual) effect First-Difference Model
##
## Call:
## plm(formula = inv ~ capital - 1, data = dataset, effect = "individual",
##     model = "fd", index = c("firm", "year"))
##
## Balanced Panel: n = 10, T = 20, N = 200
## Observations used in estimation: 190
##
## Residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -240.4   -11.7     0.1     3.5    12.6   333.2
##
## Coefficients:
##         Estimate Std. Error t-value Pr(>|t|)
## capital 0.230780   0.059639 3.8696 0.00015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    584410
## Residual Sum of Squares: 561210
## R-Squared:      0.04476
## Adj. R-Squared: 0.04476
## F-statistic: 14.9739 on 1 and 189 DF, p-value: 0.00014998
```

Comment: We can see the residuals value and, most importantly, the p-value that is .00014998, which indicates a positive relationship between the two variables.

Fixed Effects (FE): A Fixed effect model can be calculated using the linear model function and using dummy variables for all firms. I've previously shown that the factor variable firm recognizes each firm in the dataset. I'm regressing inv on capital in the same way as the pooled OLS model. When there are a large number of people, the LSDV approach is computationally costly.

```
## Call:
## lm(formula = inv ~ capital, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -316.92  -96.45  -14.43   17.07  481.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.23620   15.63927    0.91    0.364
## capital      0.47722    0.03834   12.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 162.9 on 198 degrees of freedom
## Multiple R-squared:  0.439, Adjusted R-squared:  0.4362
## F-statistic: 154.9 on 1 and 198 DF, p-value: < 2.2e-16
```

Comment: We can see ten factors and their estimate variation from output also see the p-value that is 2.2e-16, which indicates that all the factors are significantly related.

Random Effects Model: Inside and between effects are included in the RE model's coefficients. When working with data from different organizations and periods, the coefficient of capital reflects the average impact on inv when money varies by one unit across years and firms.

```
## One way (individual) effect Random Effect Model
## (Swamy-Arora's transformation)
##
## Call:
## plm (formula = inv ~ capital, data = dataset, effect = "individual",
##     model = "random", index = c ("firm", "year"))
##
## Balanced Panel: n = 10, T = 20, N = 200
##
## Effects:
##                     var std.dev share
## idiosyncratic 4040.63    63.57 0.135
## individual    25949.52   161.09 0.865
## theta: 0.9121
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -164.0821 -22.2955   -3.7463   16.9121 319.9564
##
## Coefficients:
##             Estimate Std. Error z-value Pr(>|z|)
## (Intercept) 43.246697 51.411319 0.8412   0.4002
## capital      0.372120   0.019316 19.2652   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    2299300
## Residual Sum of Squares: 799910
## R-Squared:      0.65211
## Adj. R-Squared: 0.65036
## Chisq: 371.149 on 1 DF, p-value: < 2.22e-16
```

Comment: To do so, you compare the log-likelihoods of models with and without the relevant random effect; if eliminating the random product causes a large enough decrease in log-likelihood, then the impact is statistically significant. Chi-square is used, and the p-value is less than here, as it is above.

Pooled OLS Sections: Pooled OLS model is made using the linear model function to see the regress the Investment (inv) on their stock of plant and equipment (capital). The coefficient of money shows the average effect on whether capital is increased, then inv will increase in the same way.

```
## Call:
## lm(formula = inv ~ capital, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -316.92  -96.45  -14.43   17.07  481.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.23620   15.63927    0.91    0.364
## capital      0.47722    0.03834   12.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 162.9 on 198 degrees of freedom
## Multiple R-squared:  0.439, Adjusted R-squared:  0.4362
## F-statistic: 154.9 on 1 and 198 DF, p-value: < 2.2e-16
```

Comment: The coefficient of capital shows the average effect on inv when capital increases by one unit.

Exogeneity assumptions hold: If a regression equation tells us that the independent variables X are not dependent on the dependent variable, this is known as exogeneity (Y). However, in most cases, X in our equation is dependent on Y.

Hausman Test: In a regression model, the Hausman Test (also known as the Hausman definition test) detects endogenous regressors (predictor variables). Using the test () functions, the Hausman test assesses the relationship between the Fixed and Random effect models. At first glance, there appears to be a significant relationship between H0 and Ha in H0. If the p-value is greater than.05, we can conclude no critical connection between the two variables. On the other hand, if the p-value is less than.05, we should end that the two models have a significant relationship.

```
## Hausman Test
## data:  inv ~ capital - 1
## chisq = 6.2748, df = 1, p-value = 0.01225
## alternative hypothesis: one model is inconsistent
```

Arellano–Bond estimator: The Arellano–Bond estimator is a method of moments estimator used in econometrics to estimate panel data models. Based on earlier work by Alok Bhargava and John Denis Sargan in 1983, Manuel Arellano and Stephen Bond proposed it in 1991 as a tool for solving endogeneity problems.

```
##     Estimate        Std. Error        t value         Pr(>|t|)
## Min.   :0.2308   Min.   :0.1487   Min.   :1.552   Min.   :0.1224
## 1st Qu.:0.2308   1st Qu.:0.1487   1st Qu.:1.552   1st Qu.:0.1224
## Median :0.2308   Median :0.1487   Median :1.552   Median :0.1224
## Mean   :0.2308   Mean   :0.1487   Mean   :1.552   Mean   :0.1224
## 3rd Qu.:0.2308   3rd Qu.:0.1487   3rd Qu.:1.552   3rd Qu.:0.1224
## Max.   :0.2308   Max.   :0.1487   Max.   :1.552   Max.   :0.1224
```

Conclusions and Future Work: In this study, some basic econometrics model is used to find the relationship between variables using a default dataset named Grunfeld's Data. Most of the time, there is a significant relationship between variables as the p-value is less than .05. In the future, it can be done for some real-life datasets with more samples to find the connection and predict better estimation.

Related work: The Grunfeld Data at 50, Christian Kleiber, Achim Zeileis (2010)

References

[1] "https://en.wikipedia.org/wiki/Econometrics," [Online].

[2] "https://sites.google.com/site/econometricsacademy/econometrics-models/panel-data-models," [Online].

[3] "https://vincentarelbundock.github.io/Rdatasets/doc/plm/Grunfeld.html," [Online].

[4] "https://rpubs.com/," [Online].

[5] "https://en.wikipedia.org/wiki/Panel_data," [Online].

The End