Tecnológico de Monterrey

# Modeling and Evaluation
# Time and Probabilistic Modelling

## CS5056 Data Analytics

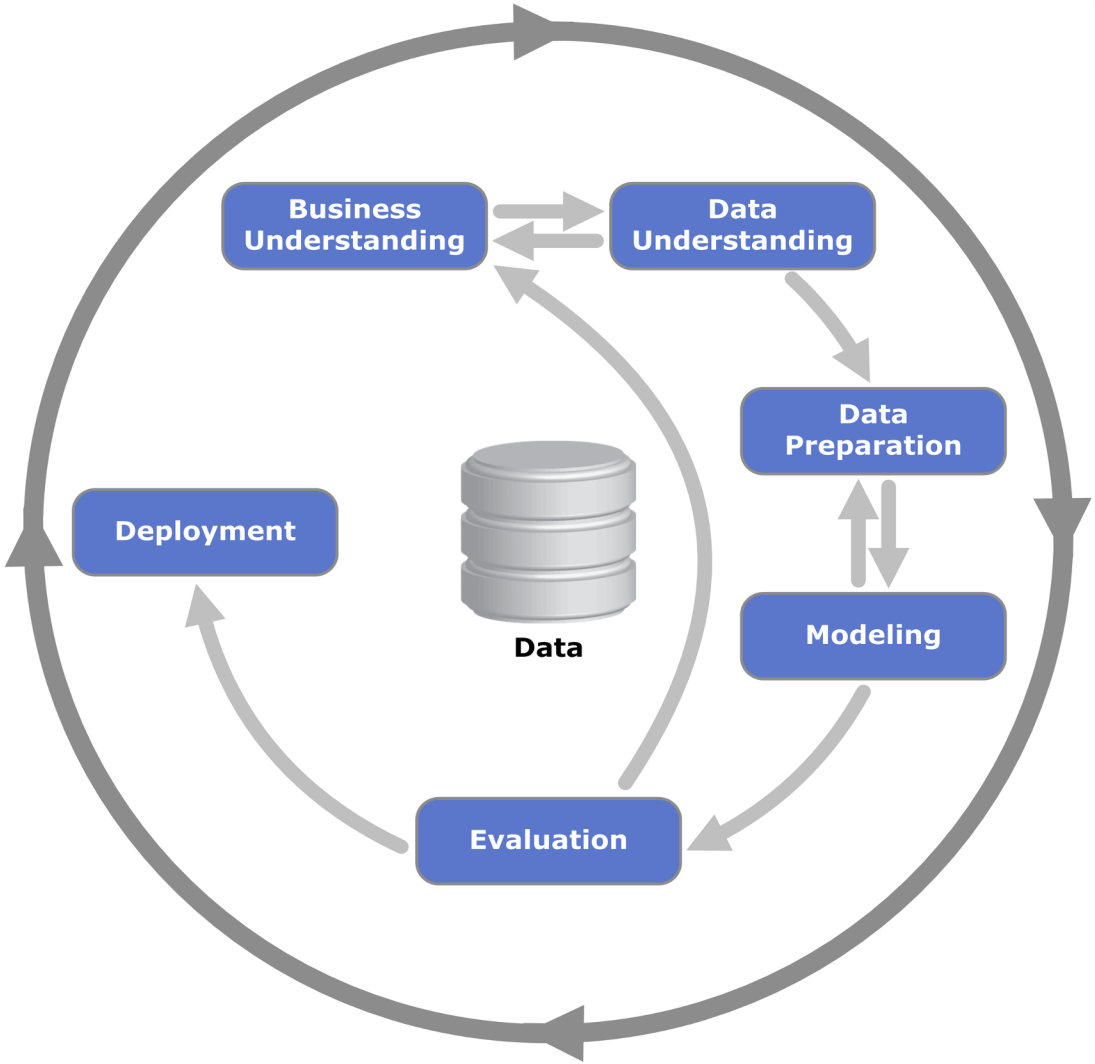
Francisco J. Cantú, Héctor Ceballos
Tecnológico de Monterrey

March 17, 2021

Februrary-June, 2021

# How do we Handle Time and Uncertainty in Fitting a Model to Dataset

# Data Mining Cycle: CRISP-DM



# Panel Data
# Multiple Regression Analysis

# Panel Data

- **Panel data** are multidimensional data involving measurements over time. It is also called longitudinal data

- Panel data contain observations of multiple phenomena obtained over periods of time for the same cross-sections: individuals, firms, or countries

- The same cross-sectional unit (individual, firm, or country) is surveyed over time, so we have data which is pooled over **space** as well as **time**.

https://towardsdatascience.com/understanding-panel-data-regression-c24cd6c5151e

# Reasons for using Panel Data

- Panel data can take explicit account of individual-specific heterogeneity

- By combining data in two dimensions, panel data gives more data variation, less collinearity and more degrees of freedom.

- Panel data is better suited than cross-sectional data for studying the dynamics of change. For example, it is well suited to understanding transition behavior, for example company bankruptcy or merger.

# Reasons for using Panel Data

- It is better in detecting and measuring the effects which cannot be observed in either cross-section or time-series data.

- Panel data enables the study of more complex behavioral models –for example, the effects of technological change, or economic cycles.

- Panel data can minimize the effects of aggregation bias, from aggregating firms into broad groups.

# Reasons for using Panel Data

- The data are usually collected over time and over the same individuals and then a regression is run over these two dimensions
- Allows control over variables that we cannot observe or measure
- Observe variables that change over time but not across groups
- Accounts for individual heterogeneity

# Panel Data combines Cross-sectional and Time series

| City | Population | Avg. Income | % Educated | Rain Fall |
|---|---|---|---|---|
| Florida | 10 | 25000 | 74 | 20 |
| California | 15 | 29000 | 87 | 30 |
| New York | 24 | 55000 | 90 | 15 |

| Date | Stock Price |
|---|---|
| 1-Dec | 120 |
| 2-Dec | 122 |
| 3-Dec | 143 |
| 4-Dec | 134 |


Analytics University
Published on Dec 11, 2016

This video is on Panel Data Analysis. Panel data has features of both Time series data and Cross section data. You can use panel data regression to analyse such data, We will use Fixed Effect Panel data regression and Random Effect panel data regression to analyse panel data. We will also compare with Pooled OLS , Between effect & first difference estimation

https://www.youtube.com/watch?v=f01WjeCdgEA

# Panel Data combines Cross-sectional and Time series

| City | Population | Avg. Income | % Educated | Rain Fall | Time |
|---|---|---|---|---|---|
| Florida | 10 | 25000 | 74 | 20 | 2014 |
| California | 15 | 29000 | 87 | 30 | 2014 |
| New York | 24 | 55000 | 90 | 15 | 2014 |
| Florida | 11 | 25500 | 74 | 12 | 2015 |
| California | 16 | 29600 | 88 | 34 | 2015 |
| New York | 25 | 56700 | 89 | 14 | 2015 |

# Balanced VS Unbalanced Panel Data

| City | Population | Avg. Income | % Educated | Rain Fall | Time |
|---|---|---|---|---|---|
| Florida | 10 | 25000 | 74 | 20 | 2014 |
| California | 15 | 29000 | 87 | 30 | 2014 |
| New York | 24 | 55000 | 90 | 15 | 2014 |
| Florida | 11 | 25500 | 74 | 12 | 2015 |
| California | 16 | 29600 | 88 | 34 | 2015 |
| New York | 25 | 56700 | 89 | 14 | 2015 |

| City | Population | Avg. Income | % Educated | Rain Fall | Time |
|---|---|---|---|---|---|
| Florida | 10 | 25000 | 74 | 20 | 2014 |
| California | 15 | 29000 | 87 | 30 | 2014 |
| | | | | | |
| Florida | 11 | 25500 | 74 | 12 | 2015 |
| California | 16 | 29600 | 88 | 34 | 2015 |
| New York | 25 | 56700 | 89 | 14 | 2015 |

# Panel Data Analysis

- A common panel data regression model looks like: $y_{it} = a + bx_{it} + e_{it}$
- where y is the dependent variable, x is the independent variable, a and b are coefficients, i and t are indices for individuals and time.
- The term $e_{it}$ represents the error

# Panel Data Assumptions

- Assumptions about the error term determine whether we speak of fixed effects or random effects.
- In a fixed effects model, $e_{it}$ is assumed to vary non-stochastically over i or t making the fixed effects model analogous to a dummy variable model in one dimension.
- In a random effects model, $e_{it}$ is assumed to vary stochastically over i or t requiring special treatment of the error variance matrix

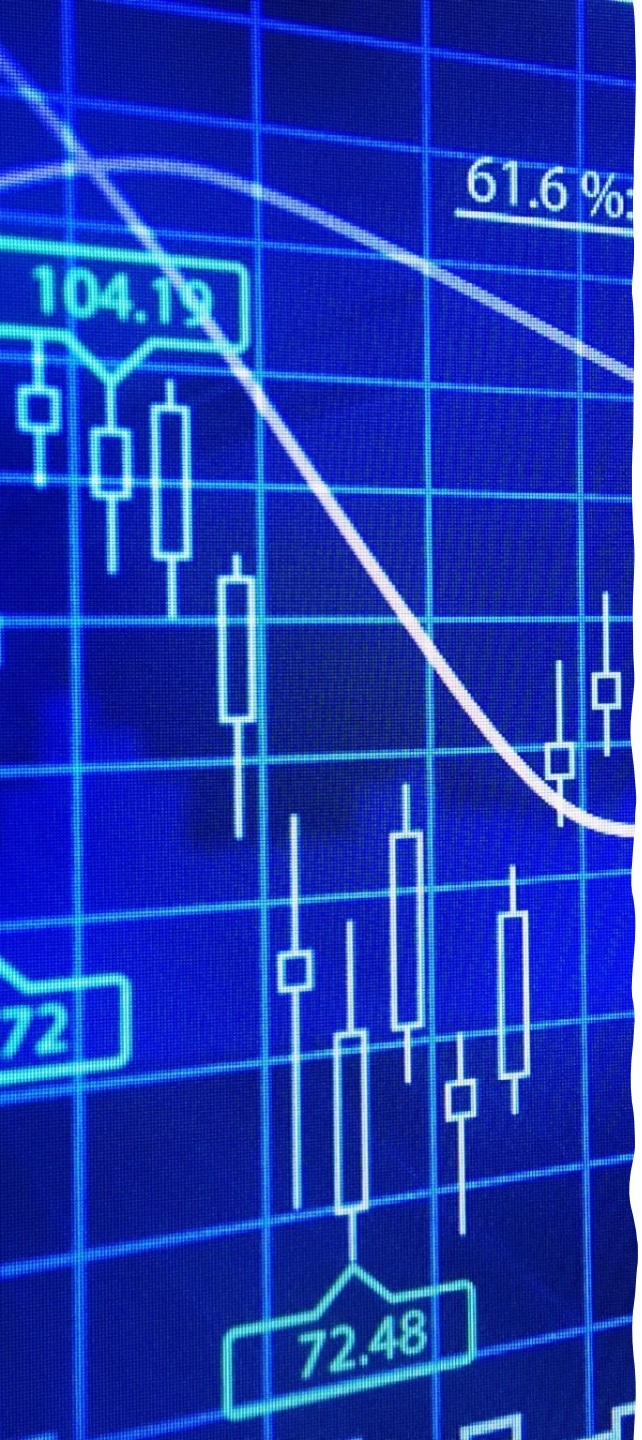
# Panel Data Methods

- Pooled OLS
- Fixed effects models or first differenced models
- Random effects models

# Panel Data Methods

- Go to PDM explanation in:

https://towardsdatascience.com/a-guide-to-panel-data-regression-theoretics-and-implementation-with-python-4c84c5055cf8




# Panel Data Analysis using R

# A Panel Data Case

| A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| log_sal | emp_id | time | experience | projects | educ | |
| 5.56198 | 1 | 1 | 3 | 32 | 9 | |
| 5.72413 | 1 | 2 | 4 | 43 | 9 | |
| 4.76142 | 1 | 3 | 5 | 40 | 9 | |
| 5.96414 | 1 | 4 | 6 | 39 | 9 | |
| 6.09112 | 1 | 5 | 7 | 42 | 9 | |
| 6.98123 | 1 | 6 | 8 | 35 | 9 | |
| 6.54286 | 1 | 7 | 9 | 32 | 9 | |
| 6.94276 | 2 | 1 | 30 | 34 | 11 | |
| 6.18599 | 2 | 2 | 31 | 27 | 11 | |
| 6.71945 | 2 | 3 | 32 | 33 | 11 | |
| 6.2456 | 2 | 4 | 33 | 30 | 11 | |

# Modeling using Panel Data

- How to fit the best model that shows the relationship between employee salary and his experience, projects undertaken and number of years of education?

- Panel Data modeling in R is done by the plm (Panel Linear Model) package

  - Let us try first Ordinary Linear Square (OLS) Pooled Panel Data

```
> pool<-plm(dep~indep,data=d,model="pooling")
> summary(pool)
Oneway (individual) effect Pooling Model

Call:
plm(formula = dep ~ indep, data = d, model = "pooling")

Balanced Panel: n=120, T=7, N=840

Residuals :
     Min.  1st Qu.   Median  3rd Qu.     Max.
-1.38000 -0.25500  0.00574  0.27900  1.41000

Coefficients :
                  Estimate Std. Error t-value Pr(>|t|)
(Intercept)      4.9648729  0.1521943 32.6219   <2e-16 ***
indepexperience 0.0156509  0.0013366 11.7092   <2e-16 ***
indepprojects   0.0036308  0.0027941  1.2994   0.1941
indepeduc       0.0962621  0.0047533 20.2517   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    204.01
Residual Sum of Squares: 131.14
R-Squared      :  0.3572
      Adj. R-Squared :  0.3555
F-statistic: 154.851 on 3 and 836 DF, p-value: < 2.22e-16
```

# Analysis of OLS Pooled Panel Data

- Bad fit to the data
- Ignore the fact that it is a panel data
- Correlation of the error terms

# Between Estimation Panel Data

- It calculates the average of the dependent and the independent variables over time and does the OLS regression of the former on the latter.
- Uses only cross-sectional information and discards time variation in the data.
- The between estimator is consistent only if the average of regressors over time are independent of the error term.

```
> between<-plm(dep~indep,data=d,model="between")
> summary(between)
Oneway (individual) effect Between Model

Call:
plm(formula = dep ~ indep, data = d, model = "between")

Balanced Panel: n=120, T=7, N=840

Residuals :
    Min. 1st Qu.  Median 3rd Qu.     Max.
-0.8780 -0.1770  0.0324  0.2200  0.7500

Coefficients :
                 Estimate Std. Error t-value  Pr(>|t|)
(Intercept)     4.8413875  0.4710134 10.2787 < 2.2e-16 ***
indepexperience 0.0124702  0.0030095  4.1436 6.526e-05 ***
indepprojects   0.0085089  0.0093326  0.9117    0.3638
indepeduc       0.0934853  0.0104833  8.9176 8.091e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    21.844
Residual Sum of Squares: 12.617
R-Squared      :  0.42241
     Adj. R-Squared :  0.40833
F-statistic: 28.278 on 3 and 116 DF, p-value: 8.4878e-14
```

# First Difference Estimation Panel Data

- Exploits the features of a panel data.
- Finds the association between the individual-specific changes in the repressors and the individual-specific changes in the dependent variable.

# First Difference Estimation Panel Data

- It lags the individual-specific variables by one period and takes the difference between the two equations.
- Thus, individual heterogeneity is eliminated from the model.

# First Difference Estimation Panel Data

- Allows Data Scientists to detect and study hidden variables
- IQ
- University
- Parents income

```
> fd<-plm(dep~indep,data=d,model="fd")
> summary(fd)
Oneway (individual) effect First-Difference Model

Call:
plm(formula = dep ~ indep, data = d, model = "fd")

Balanced Panel: n=120, T=7, N=840

Residuals :
     Min.   1st Qu.    Median   3rd Qu.      Max.
-1.05000 -0.07980 -0.00866  0.04950  1.36000

Coefficients :
                Estimate Std. Error t-value Pr(>|t|)
(intercept)    0.0957110  0.0068851  13.901  < 2e-16 ***
indepprojects 0.0027597  0.0012944   2.132  0.03335 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    24.659
Residual Sum of Squares: 24.504
R-Squared      :  0.0062906
      Adj. R-Squared :  0.0062731
F-statistic: 4.54525 on 1 and 718 DF, p-value: 0.033349
>
```

# Analysis of First Difference Estimation Panel Data

- The variable “education” got dropped from the model.
- Education does not vary from one year to the next, thus difference between the education values for two consecutive years is zero.

# Fixed Effect Panel Data

- Correlation between $alfa_i$ and $X_i$ is different from zero

- Treats the unobserved individual heterogeneity ( αi )for each employee to be correlated with the explanatory variables.

- FE estimation involves a transformation to remove the unobserved effect αi  prior to estimation.

# Fixed Effect Panel Data

- FE Transformation
- $y_{it} = \beta_1 x_{it} + \alpha_i + u_{it}$ $\quad t=1,2,\ldots,T$

- On averaging this equation over time,
- $y_i' = \beta_1 x_i' + \alpha_i + u_i'$
- Subtracting these two equations,
  $(y_{it} - y_i') = \beta_1 (x_{it} - x_i') + (u_{it} - u_i')$ $\quad t=1,2,..T$

- Differencing has led to elimination of $\alpha_i$
- The LHS is called time-demeaned y.
- Similarly, time-demeaned x and u.

# Fixed Effect Within Estimation Panel Data

- OLS can be applied on the time-demeaned equation. This is called FIXED EFFECTS / WITHIN ESTIMATION.

- This is because of the use of OLS on the time variation in x and y within each cross-sectional observation.

```
> fxd<-plm(dep~indep, data=d, model="within")
> summary(fxd)
Oneway (individual) effect Within Model

Call:
plm(formula = dep ~ indep, data = d, model = "within")

Balanced Panel: n=120, T=7, N=840

Residuals :
     Min.   1st Qu.    Median   3rd Qu.      Max.
-1.09000 -0.05740  0.00685  0.06610  0.84100

Coefficients :
                  Estimate Std. Error t-value Pr(>|t|)
indepexperience 0.1002007  0.0027065 37.0218  < 2e-16 ***
indepprojects   0.0027540  0.0014705  1.8729  0.06149 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    51.101
Residual Sum of Squares: 17.556
R-Squared      :  0.65644
      Adj. R-Squared :  0.5611
F-statistic: 685.943 on 2 and 718 DF, p-value: < 2.22e-16
>
```

# Analysis of Fixed Effect Within Estimation Panel Data

- The variable “education” is dropped because it is time invariant and gets cancelled.
- “experience” has a positive effect on the log salary.
- If the experience of an employee rises by one year, there is a 10% increase in wages.
- The R-square value 65.644% shows the amount of time variation in y it that is explained by time variation in Xit .

# Random Effect Panel Data

- Correlation between $alfa_i$ and $X_i$ is zero

- Assumes that the individual-specific effects are independent of the regressors.

- This individual-specific effect is included as the error term

```
> ran<-plm(dep~indep, data=d, model="random")
> summary(ran)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = dep ~ indep, data = d, model = "random")

Balanced Panel: n=120, T=7, N=840

Effects:
                   var std.dev share
idiosyncratic 0.02445 0.15637 0.188
individual    0.10527 0.32446 0.812
theta:  0.8208

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-1.00000 -0.10700  0.00924  0.11400  1.04000

Coefficients :
                  Estimate Std. Error t-value Pr(>|t|)
(Intercept)      3.6239454  0.2093277 17.3123   <2e-16 ***
indepexperience  0.0606693  0.0025145 24.1280   <2e-16 ***
indepprojects    0.0015809  0.0018167  0.8702   0.3844
indepeduc        0.1327592  0.0129274 10.2696   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    56.012
Residual Sum of Squares: 32.031
R-Squared      :  0.42814
     Adj. R-Squared :  0.4261
F-statistic: 208.63 on 3 and 836 DF, p-value: < 2.22e-16
```

# Analysis of Random Effect Panel Data

- Theta = 0.8208 shows the percentage of the variation comes from individual variation versus the idiosyncratic one.
- There is a positive effect of experience and education on the log of the salary.

# Linear Model Test for Random Effect VS OLS

- The LM is used to decide between a random effects regression and a simple OLS regression.
- Done using the plmtest function.
- The null hypothesis is that there is no significant difference across cross-sectional units.(i.e. no panel effect) implying that RE model is inappropriate.

```
> plmtest(pool)

        Lagrange Multiplier Test - (Honda)

data:  dep ~ indep
normal = 31.5945, p-value < 2.2e-16
alternative hypothesis: significant effects

>
```

# Linear Model Test for Random Effect VS OLS

- The test statistic is 31.5945 and the p value is less than 0.05
- It is significant and the null hypothesis in favor of OLS is rejected.
- Thus, the Random Effects model is chosen against the OLS.

# Linear Model Test for Fixed Effect VS OLS

- The LM test to choose between the fixed effects model and the OLS is done using the function pFtest on the fixed and pooled estimates.

- The null hypothesis is that there are no time-invariant effects and so OLS should be used.

```
> pFtest(fxd,pool)

        F test for individual effects

data:  dep ~ indep
F = 39.3652, df1 = 118, df2 = 718, p-value < 2.2e-16
alternative hypothesis: significant effects

>
```

# Linear Model Test for Fixed Effect VS OLS

- The test statistic is 39.3652 and the p value is less than 0.05
- The null hypothesis is rejected thus implying that the Fixed effects model should be used instead of the OLS.

# Hausman Test for Fixed Effect VS Random Effect

- The Hausman test is used to decide between the fixed effects model and the random effects model.
- This is done using the phtest (Panel Hausman Test) function.

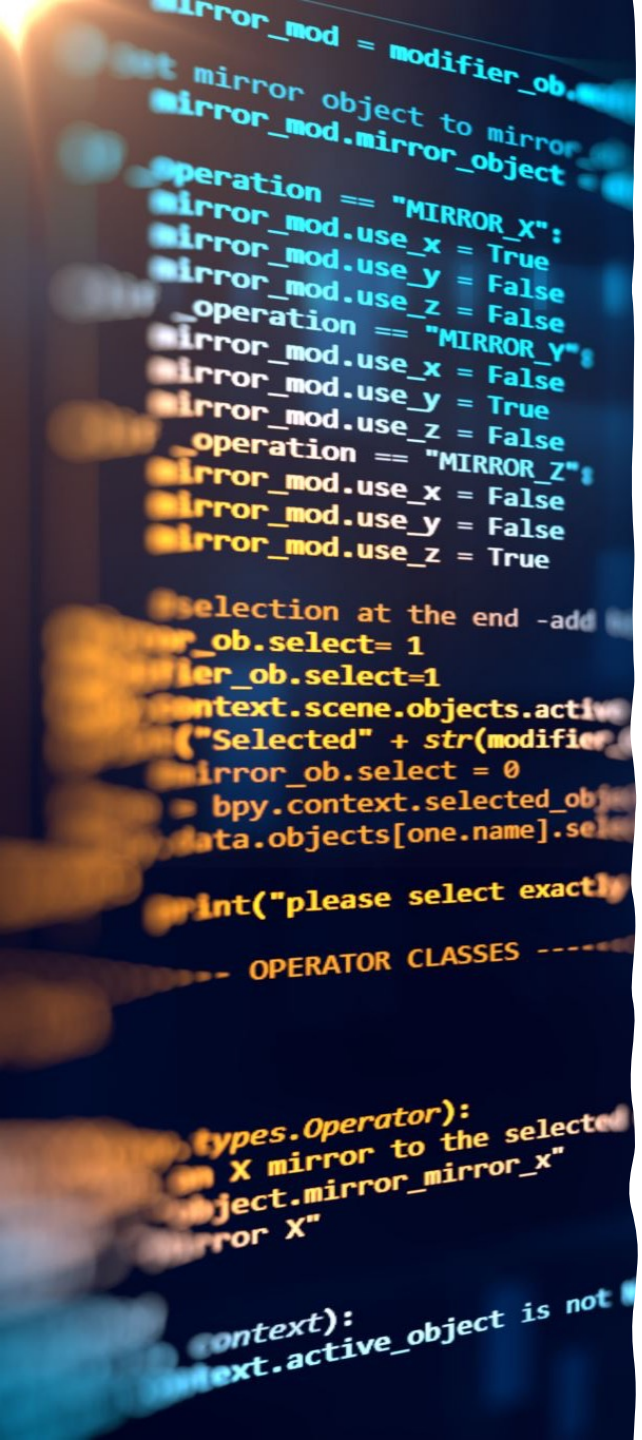
```
> phtest(ran,fxd)

        Hausman Test

data:  dep ~ indep
chisq = 1557.3, df = 2, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent
```

# Hausman Test for Fixed Effect VS Random Effect

- The Chi Square test statistic is 1557.3 and the p value is less than 0.05
- As the null is rejected, the Fixed effects method is chosen to model the data instead of the random effects model.

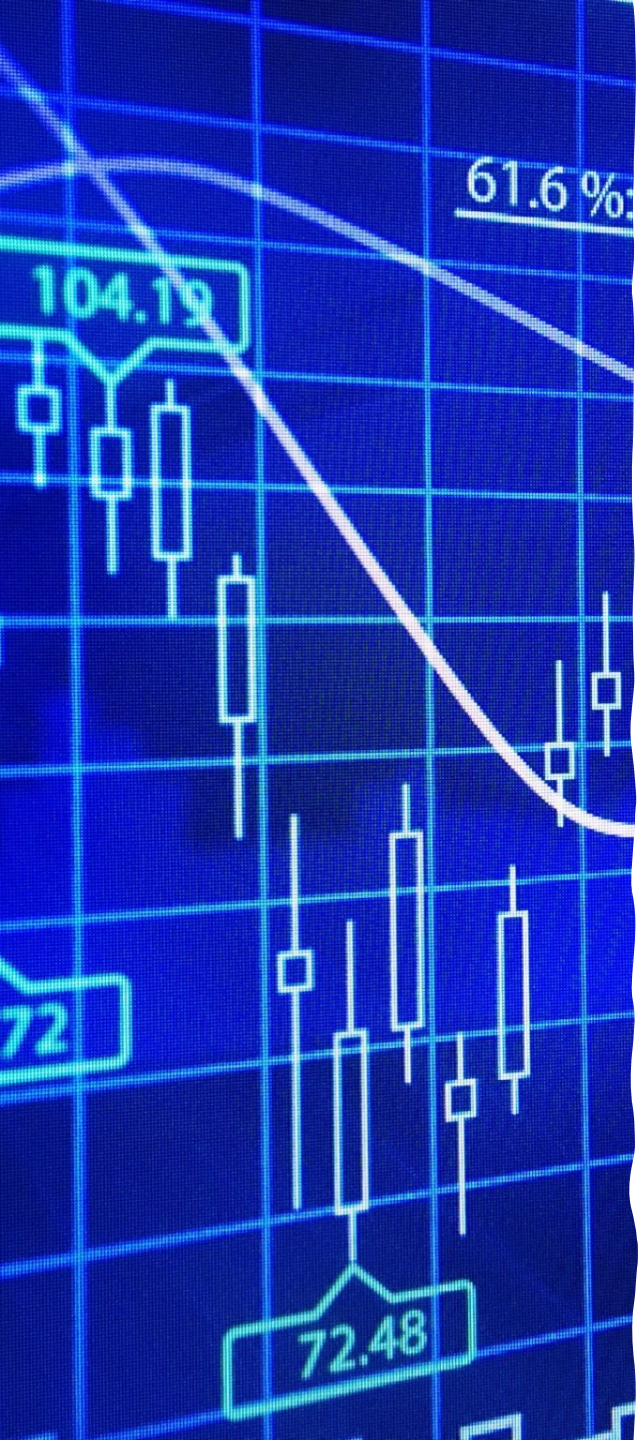



# Hands-on Exercise on Panel Data Analysis using Python

# Panel Data Analysis using Python

Load the Python script panel-data.py

- https://bashtage.github.io/linearmodels/panel/examples/examples.html




# Panel Data Analysis using R

https://rstudio-pubs-static.s3.amazonaws.com/278903_d946943cacf1492ab8d4dee5192c7f93.html

https://towardsdatascience.com/panel-data-regression-a-powerful-time-series-modeling-technique-7509ce043fa8

# Hands-on Exercise

https://cran.r-project.org/web/packages/plm/vignettes/plmPackage.html

```
library("plm")
data("EmplUK", package="plm")
data("Produc", package="plm")
data("Grunfeld", package="plm")
data("Wages", package="plm")
head(Grunfeld)
E <- pdata.frame(EmplUK, index=c("firm","year"), drop.index=TRUE, row.names=TRUE) head(E)
head(attr(E, "index"))
summary(E$emp)
```

# Hands-on Exercise

https://cran.r-project.org/web/packages/plm/vignettes/plmPackage.html

**Modelling:**

```
grun.fe <- plm(inv~value+capital, data = Grunfeld, model = "within")
grun.re <- plm(inv~value+capital, data = Grunfeld, model = "random")
summary(grun.re)
fixef(grun.fe, type = "dmean")
summary(fixef(grun.fe, type = "dmean"))
grun.twfe <- plm(inv~value+capital, data=Grunfeld, model="within",
effect="twoways") fixef(grun.twfe, effect="time")

grun.tways <- plm(inv~value+capital, data = Grunfeld,
effect = "twoways", model = "random", random.method = "amemiya")
summary(grun.tways)
```

# Hands-on Exercise

https://cran.r-project.org/web/packages/plm/vignettes/plmPackage.html

**Testing and Comparing Models**

znp <- **pvcm**(inv~value+capital, data=Grunfeld, model="within")

zplm <- **plm**(inv~value+capital, data=Grunfeld, model="within")

**pooltest**(zplm, znp)

**pooltest**(inv~value+capital, data=Grunfeld, model="within")

g <- **plm**(inv ~ value + capital, data=Grunfeld, model="pooling")

**plmtest**(g, effect="twoways", type="ghm")

Hands-on Exercise

https://cran.r-project.org/web/packages/plm/vignettes/plmPackage.html

**Testing and Comparing Models**

**Hausman Test**

```
gw <- plm(inv~value+capital, data=Grunfeld, model="within")
gr <- plm(inv~value+capital, data=Grunfeld, model="random")
phtest(gw, gr)
```

# Conclusion

- Panel data modeling is a method to estimate data which is both time-series and cross-sectional.
- It is crucial to study the same cross-sectional unit (firm, country, industry) over a period of time.
- It accounts for individual-specific heterogeneity.

# Francisco J. Cantú-Ortiz, PhD

Professor of Computer Science and Artificial Intelligence
Tecnológico de Monterrey
Enago-Academy Advisor for Strategic Alliances

E-mail: fcantu@tec.mx, fjcantor@gmail.com
Cel: +52 81 1050 8294, SNI-2 CVU: 9804
Personal Page: http://semtech.mty.itesm.mx/fcantu/
Facebook: fcantu; Twitter: @fjcantor; Skype: fjcantor
Orcid: 0000-0002-2015-0562
Scopus ID:6701563520
Researcher ID: B-8457-2009
https://www.researchgate.net/profile/Francisco_Cantu-Ortiz
https://scholar.google.com.mx/citations?hl=es&user=45-uuK4AAAAJ
https://itesm.academia.edu/FranciscoJavierCantuOrtiz
Ave. Eugenio Garza Sada No. 2501, Monterrey N.L., C.P. 64849, México