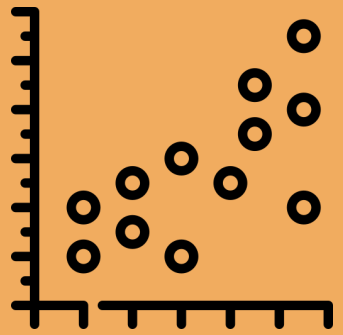


LINEAR REGRESSION, HYPOTHESIS TESTING & PAPERS

Student Learning Outcomes (SLOS)

The student should be able to:

- [ ] **a**. Describe what a variable is and type of variables
- [ ] **b**. Describe what econometrics is
- [ ] **c**. Describe the purpose of econometrics
- [ ] **d**. Describe the three stage method for econometrics
- [ ] **e**. Understand the core of a linear regression
- [ ] **f**. Analyze the difference between Population Regression Function and Sample Regression Function

Student Learning Outcomes (SLOS)

- [ ] **g**. Analyze the mathematical approach behind Simple Linear Regression and Multiple Linear Regression
- [ ] **h**. Understand when to reject or not to reject a null hypothesis
- [ ] **i**. Understand how to state a hypotesis testing
- [ ] **j**. Identify the components of a scientific paper
- [ ] **k**. Analyze the basic of reading a scientific paper

What is a **variable?**

A variable is a characteristic that can be measured and that can assume different values.

Height, age, income, province or country of birth, grades obtained at school and type of housing are all examples of variables.

Variables

Variables

Variables may be classified into **two** main **categories**

Categorical

A categorical variable (also called qualitative variable) refers to a characteristic that cannot be quantifiable.

Nominal

A nominal variable is one that **describes** a name, label or category without natural order.

Ordinal

An ordinal variable is a variable whose values are **defined by an order** relation between the different categories.

Numeric

A numeric variable (also called quantitative variable) is a quantifiable characteristic whose values are numbers

Continuous

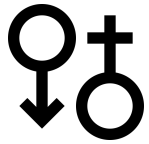
A variable is said to be continuous if it can assume an **infinite number** of real values within a given interval.

Discrete

A discrete variable can assume only a **finite number** of real values within a given interval.

Variables

Categorical



Nominal

Men, women
Green, Red, Blue

Ordinal

S, M, L
A, B, C (Grades)

Numeric

Continuous

1,2 ,3
26 students

Discrete

Age
Weight

But really…what is **econometrics**?

What is **econometrics**?

"[Econometrics]… is the unification of statistics, economic theory and mathematics."

-Ragnar Frisch

"Econometrics is based upon the development of statistical methods for estimating economic relationships, testing economic theories, and evaluating and implementing government and business policy."

-Jeffrey Wooldridge

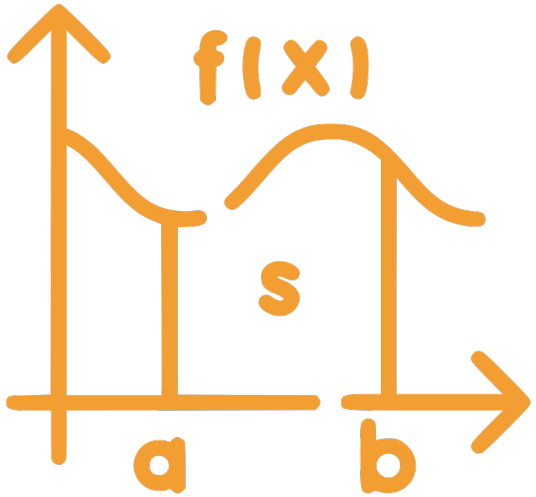
"[Econometrics is the]…social science in which the tools of economic theory mathematics, and statistical inference are applied to the analysis of economic phenomena"

-Arthur Golberger

What is **econometrics** about?

It is about the application of…

Economics theory

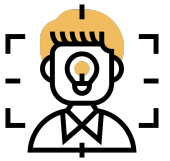


Maths

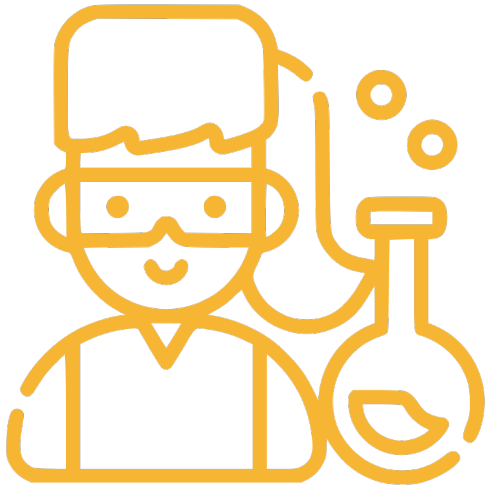
Statistical techniques

What is **econometrics** for?

Three main **functions**

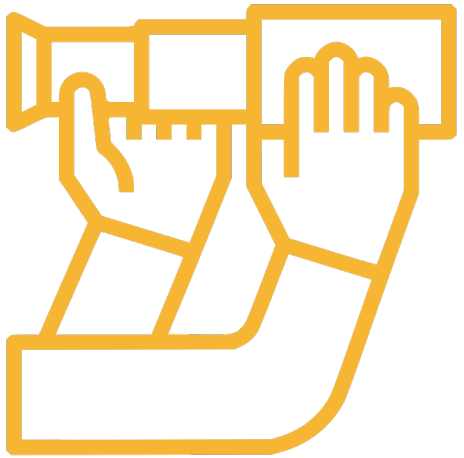


What is the difference between a theory and a hypothesis?



To **test** theories or hypothesis

To **estimate**



To **forecast**

Method of **econometrics**

Researching stages

3 MAIN STAGES

Method of econometrics

Researching stages

1
Theoretical foundation

Establish economic theory

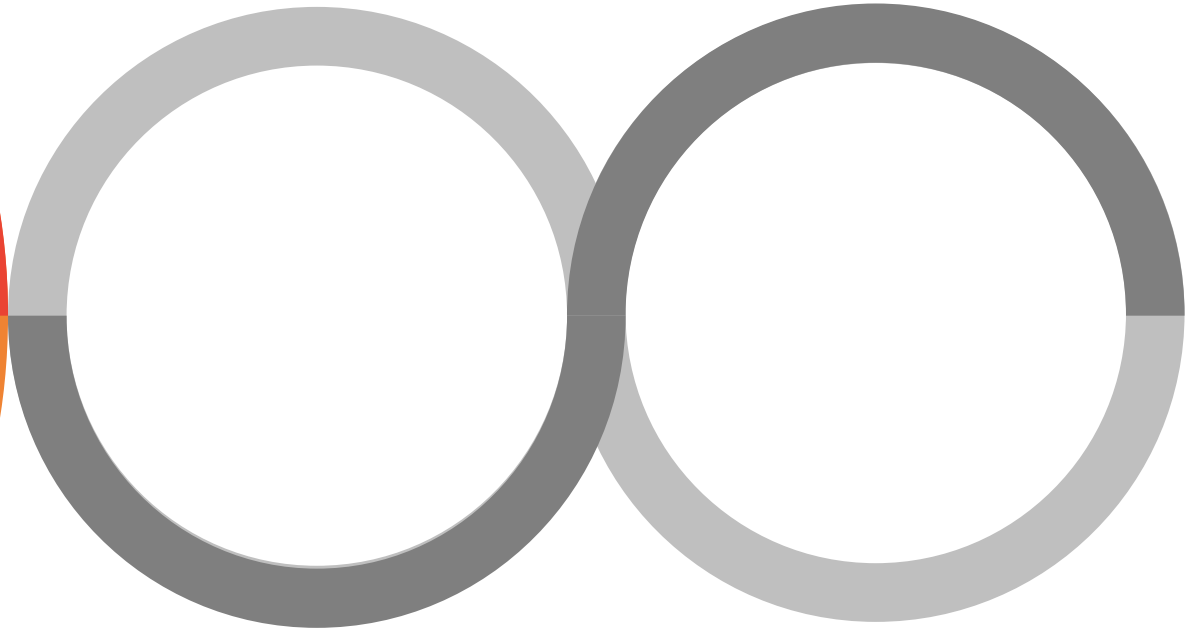
Demand theory

The quantity demanded of a product Y
its a function that depends on

$$P_i = Price\ per\ unit\ of\ good\ X$$

$$I_i = Consumers'\ Income$$

$$S_i = Price\ of\ another\ good$$



1

Theoretical foundation

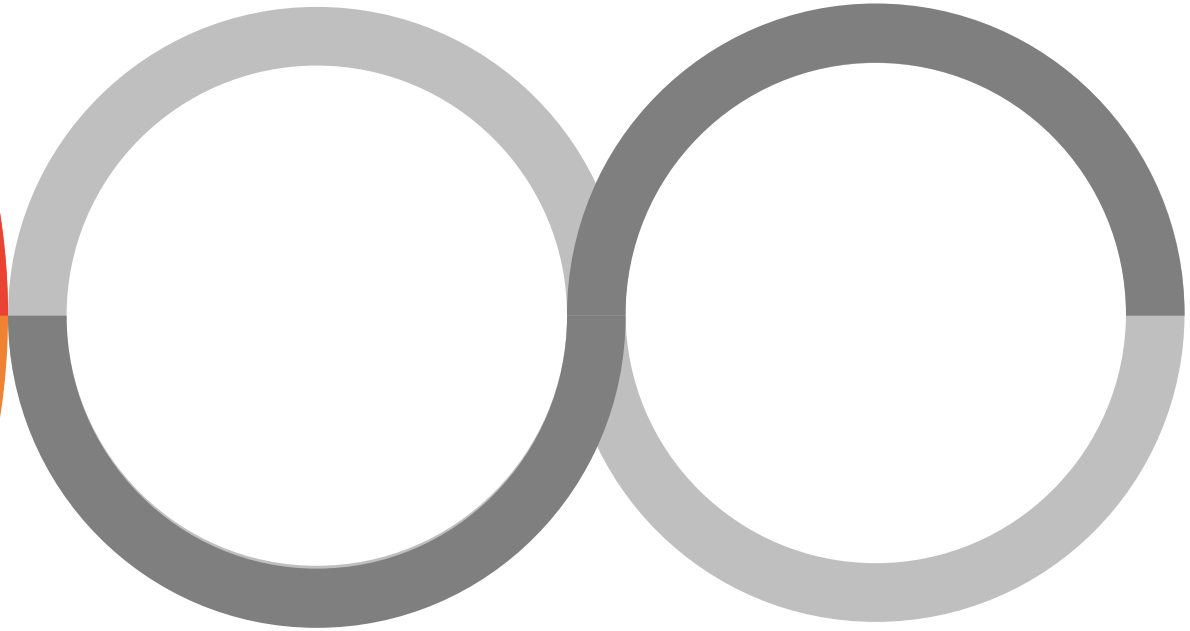
State equation

Explicit or linear

$$y = b_0 + b_1 P_i + b_1 I_i + b_1 S_i$$

Stochastic

$$y = b_0 + b_1 P_i + b_1 I_i + b_1 S_i + u$$



2
Data collection - modelling

Data for dependent and independents variables

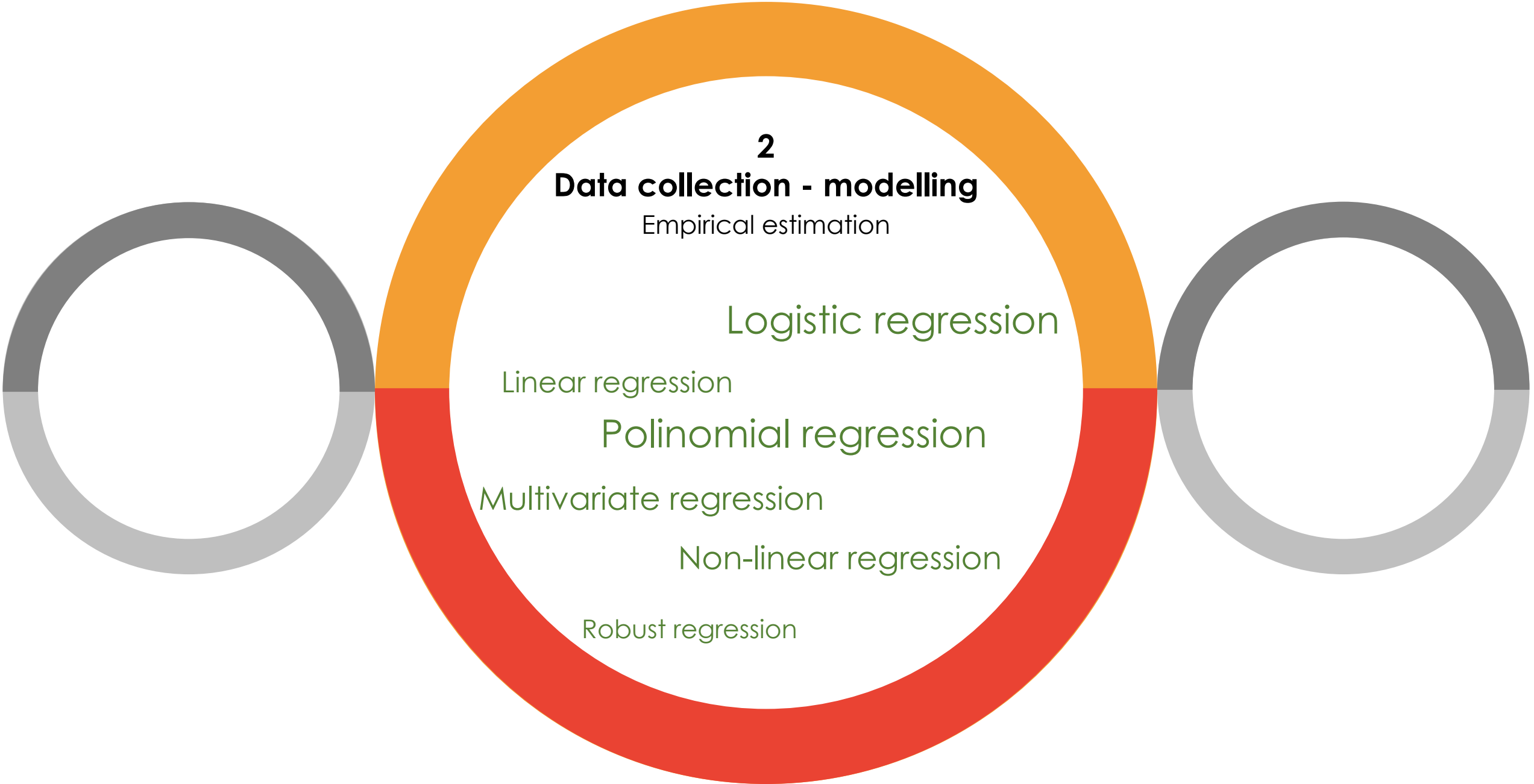
Bloomberg

INEGI

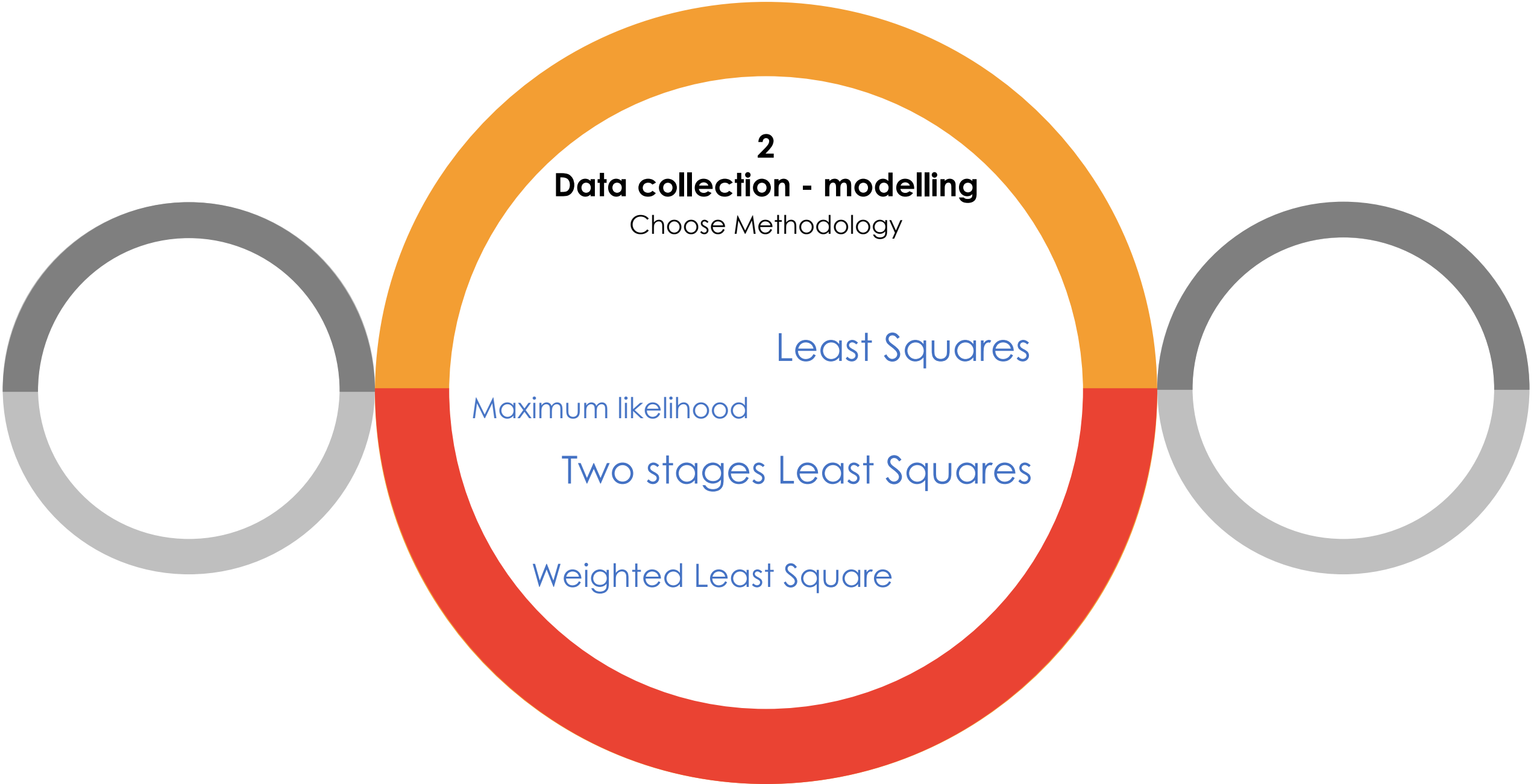
THE WORLD BANK

UNITED STATES FEDERAL RESERVE SYSTEM

Method of econometrics



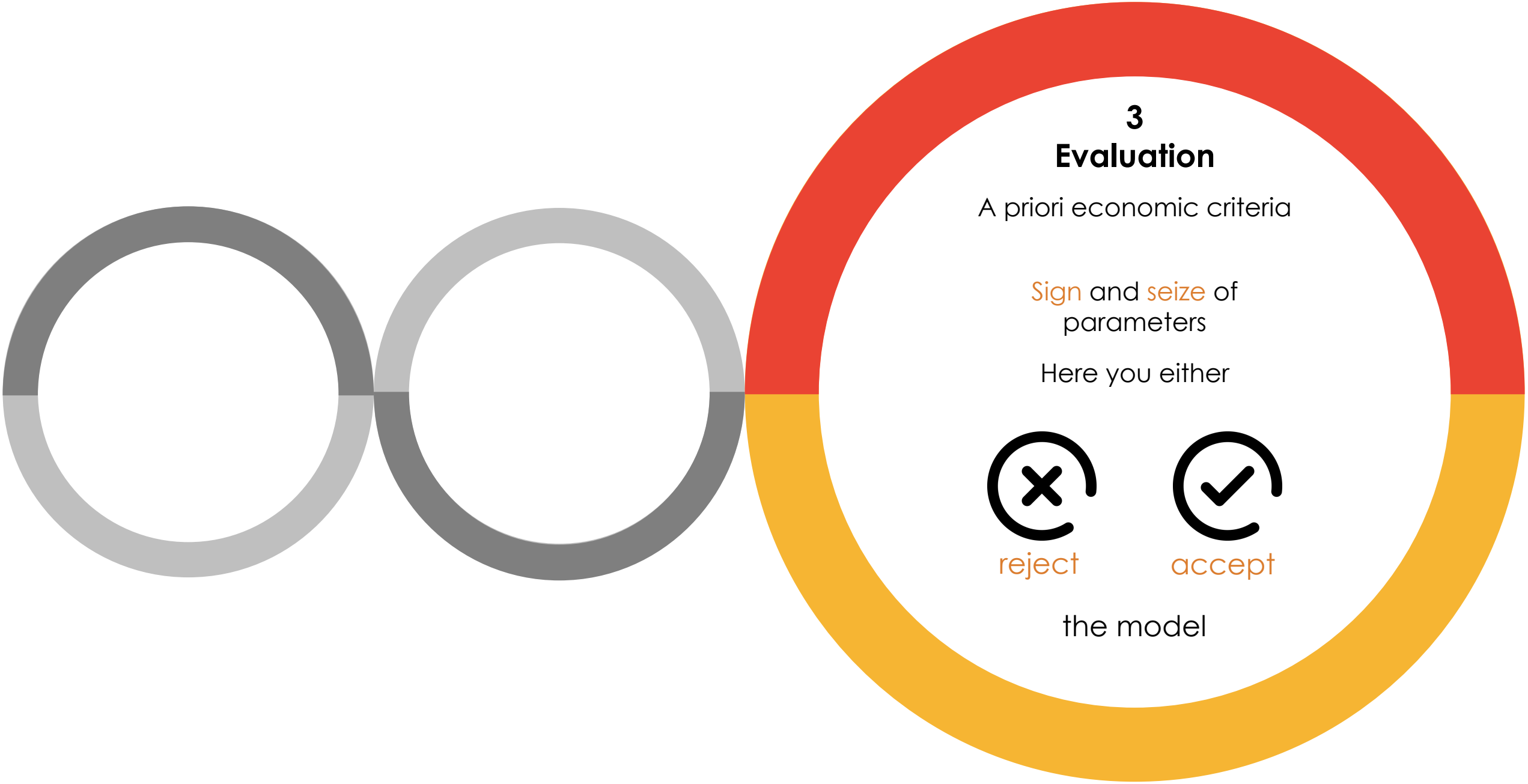
Method of econometrics



2

Data collection - modelling

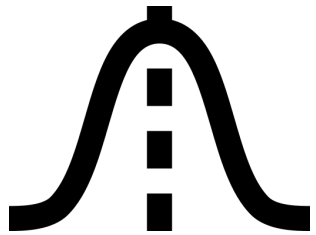
Choose
Language/Software


3
Evaluation
A priori economic criteria
Sign and seize of parameters
Here you either
reject
accept
the model

Method of econometrics

3
Evaluation

Statistical criteria



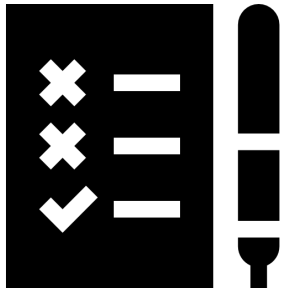
The dispersion of every estimated coefficient around the parameter must be close enough to create confidence on estimation

Method of econometrics

3
Evaluation

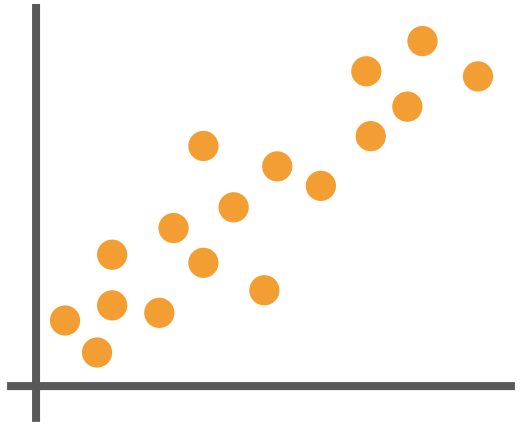
Econometrical criteria

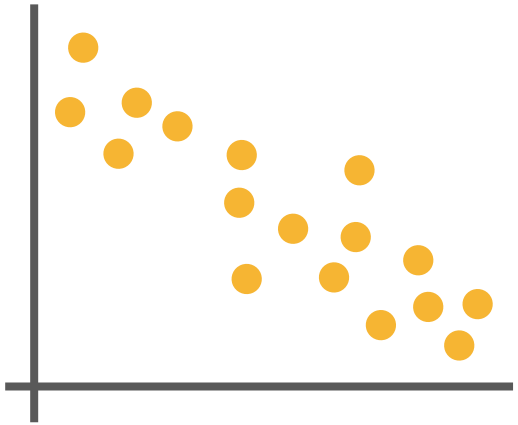
We as econometricians love tests, they indicate us if the assumptions under the model are satisfied

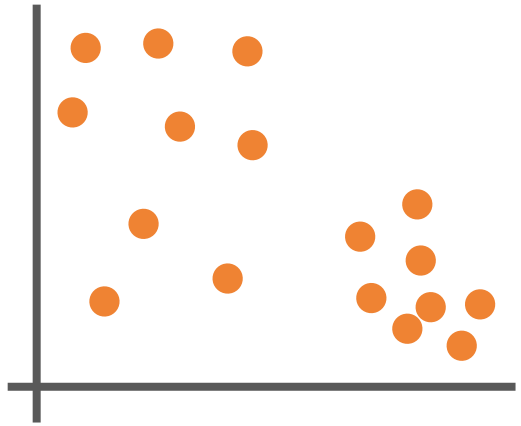


What is a **linear regression**?

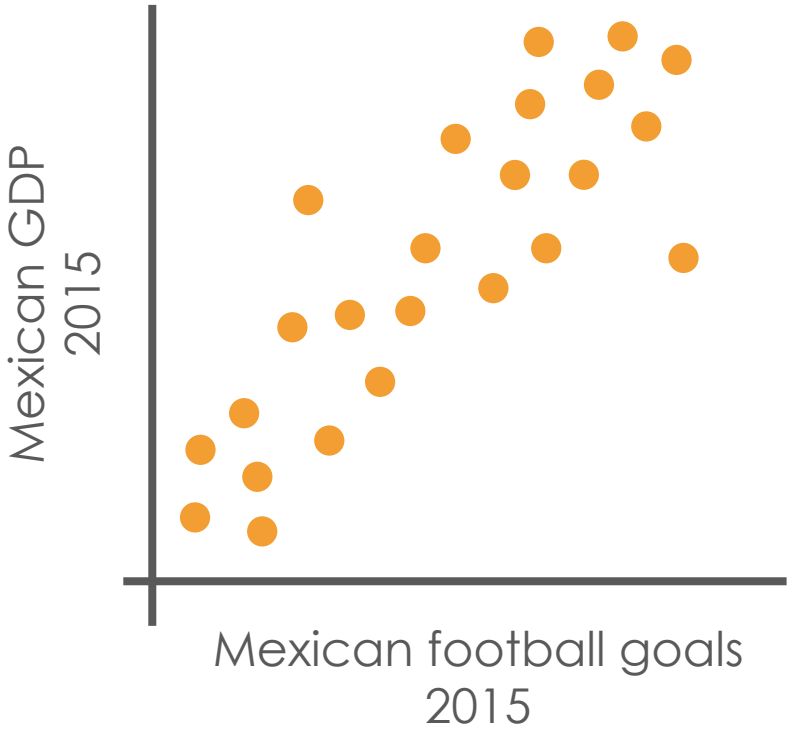
It is the relationship between an independent variable and one or more explanatory variables







What is a **linear regression**?



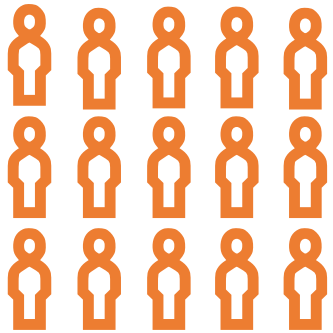
Notice that it does not mean there is a cause-effect implication

This means you can run regressions on your software, it's free, but that does not necessarily indicate it has a theoretical background

Correlation does not imply causation

Cum hoc ergo propter hoc

Population **regression function**

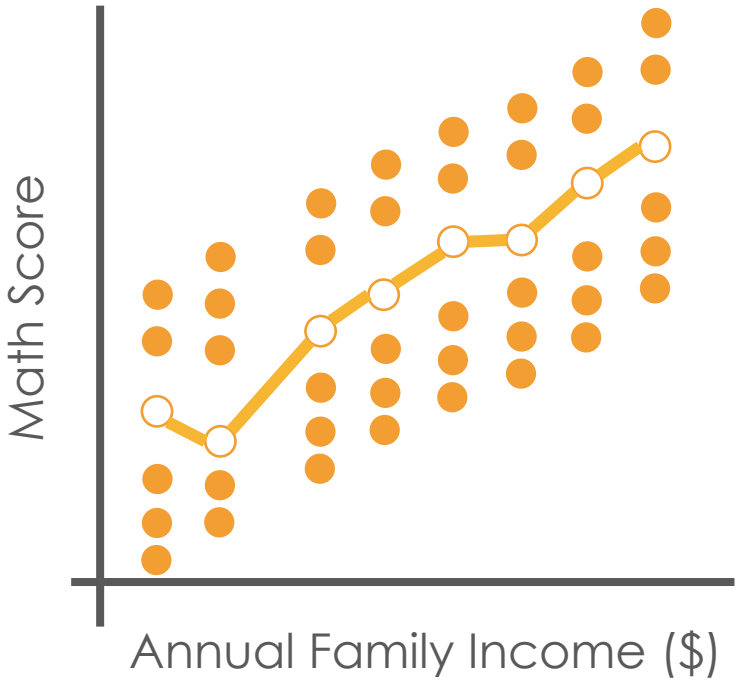


VS

Sample **regression function**

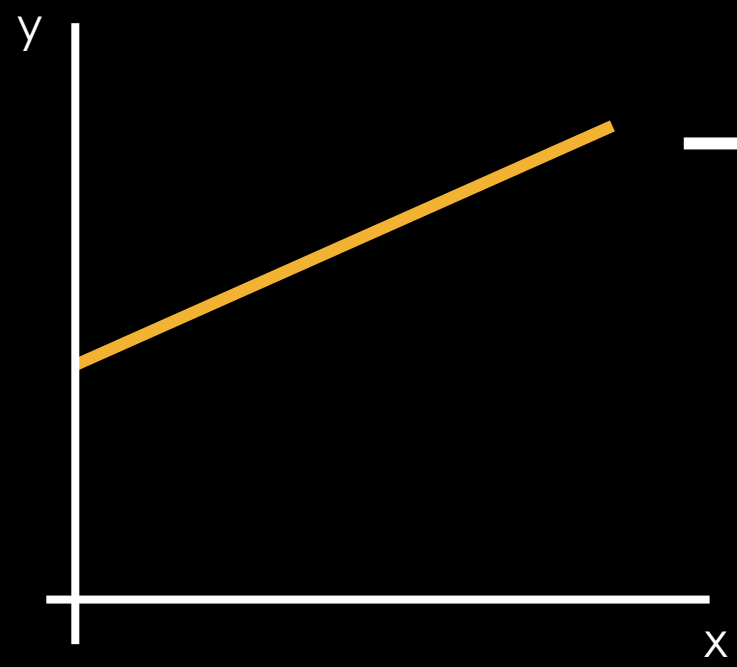
Population **regression function**

Imagine we want to prove there is a relationship between math test score and Annual Family Income



Population **regression function**

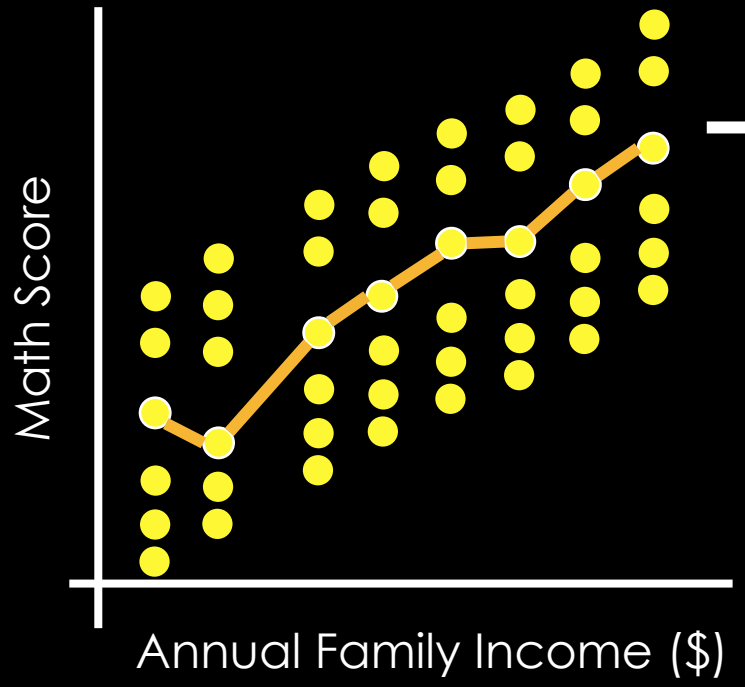
Do you remember the **equation** for a **straight line**?



$$y = mx + b$$

Population **regression function**

That **turns** into this…



$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

The regression is **estimated**, **linear** and **mathematical**

Population **regression function**

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

Regression coefficients

Intercept

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

Slope

Population **regression function**

Expected mean value when all $X = 0$

Intercept

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

Slope

If the **slope** is 2, it means that if the value of X **increases** by 1 then the value of Y **incresaes** by 2

It measures the change in the mean conditional value of Y for a change in X

"How much you can expect Y to change as X increases"

Population **regression function**

This equation that is **conditional**

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

turns into this

$$Y_1 = \beta_1 + \beta_2 X_i + \mu_i$$

Population **regression function**

$$Y_1 = \beta_1 + \beta_2 X_i + \mu_i$$

Stochastic Term

- Random
- Error term

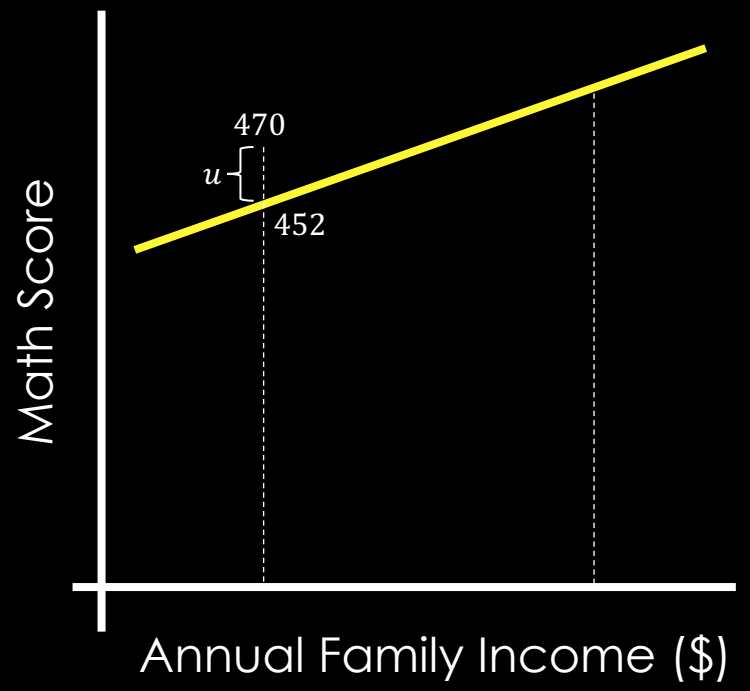
Population **regression function**

$$Y_1 = \beta_1 + \beta_2 \, X_i + \mu_i$$

Two **components**:

Systemic or deterministic

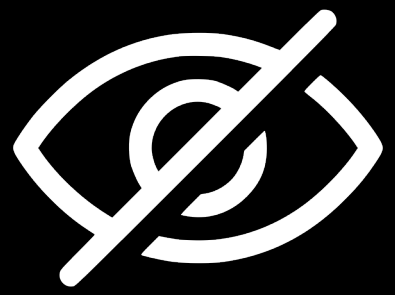
Non-systemic or noise



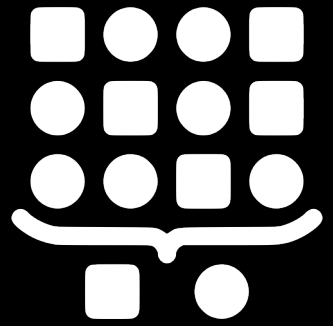
$$Y_1 = \beta_1 + \beta_2 X_i + \mu_i$$

One or **more** of the following:

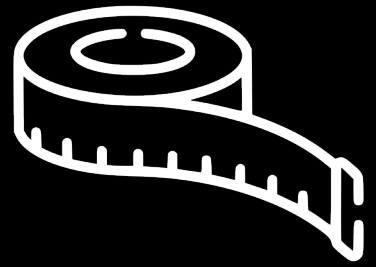
Not specified variables that may impact in the model



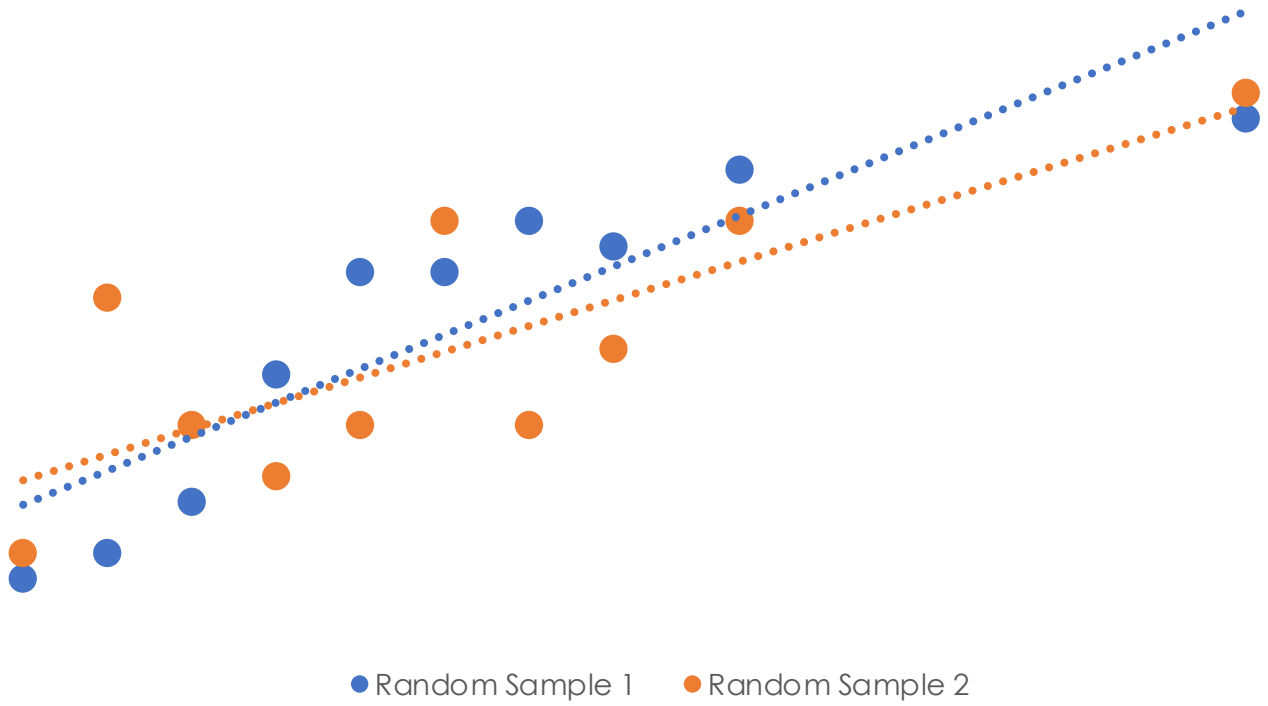
Even when all significant variables are included in the model, there is an **intrinsic random factor** that **cannot be explained**



Measurement errors

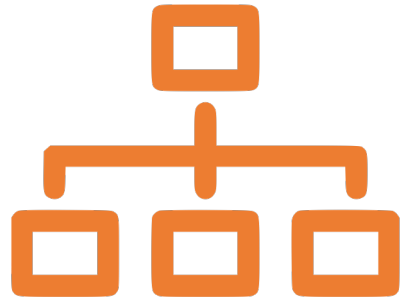


Sample **regression function**



When we cannot access to the whole universe of data, we use the sample regression function in order to estimate the population regression function

Simple linear **regression** vs Mutiple linear **regression**



Simple linear **regression**

It is the relation between a **dependent variable** and **ONE independent variable**

$$Y_t = (a + bX_i) + e_i$$

or

$$Y_t = (b_0 + bX_i) + e$$

where

- a or b_0 are ordinate axis
- b or b_1 slope or gradient, both are regression coefficients
- e is the residual term that represents the difference between the observed value and the predicted one

Simple linear **regression**

Do you remeber the **coefficient term**?

$$E(Y|X_i) = \boxed{\beta_1} + \boxed{\beta_2} X_i$$

Regression coefficients

Simple linear **regression**

We estimate them
from data using OLS

O Ordinary
L Least
S Squares

$$E(Y|X_i) = \boxed{\beta_1} + \boxed{\beta_2} X_i$$

Regression coefficients

Now, we want to **add more variables**, to make our model more complex

What can we do?

We use **Multiple Linear Regression** which is **more suitable** for a ***ceteri paribus* analysis** due to its convenience to control the **rest of factors** that affect simultaneously to the dependent variable

Multiple linear **regression**

We pass from this

$$Y = \beta_0 + \beta_1 X_1 + u_i$$

Multiple linear **regression**

To this

$$Y = \beta_0 + \beta_1 \overline{X_1} + \beta_2 \underline{X_2} + u_i$$

variable

variable

Multiple linear **regression**

Or this

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u_i$$

Where

- β_0 is the intercept
- β_1 is the change in y with respect to x_1, *ceteris paribus*
- β_i is the change in y with respect to x_i, *ceteris paribus*

Do you remember that we said we love testing?

The important task is not to obtain it because a software can easily make it for you but to interpret it. This is the next step right after running your regression.

The smaller the p.value the more evidence we have against H_0

Hypothesis testing

The smaller the p.value the more evidence we have against H_0

or

If p.value is "very" small, then we have more evidence to reject the null hypothesis

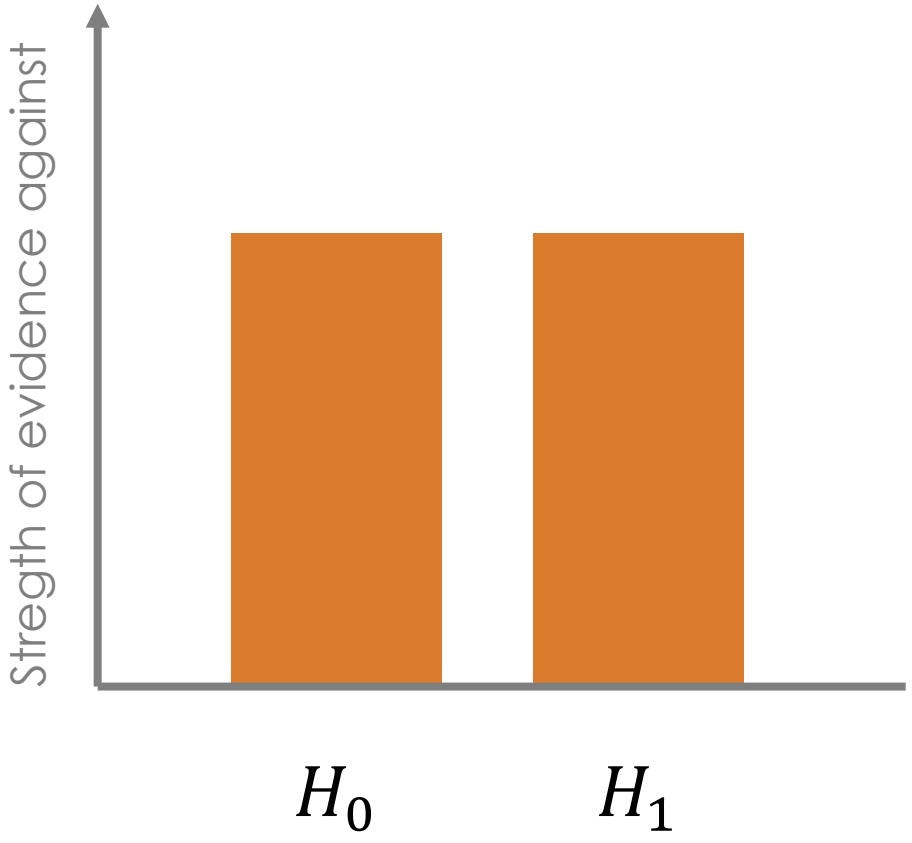
Hypothesis testing

All depends on the significance level, which you may have heard about…

That number you always declare

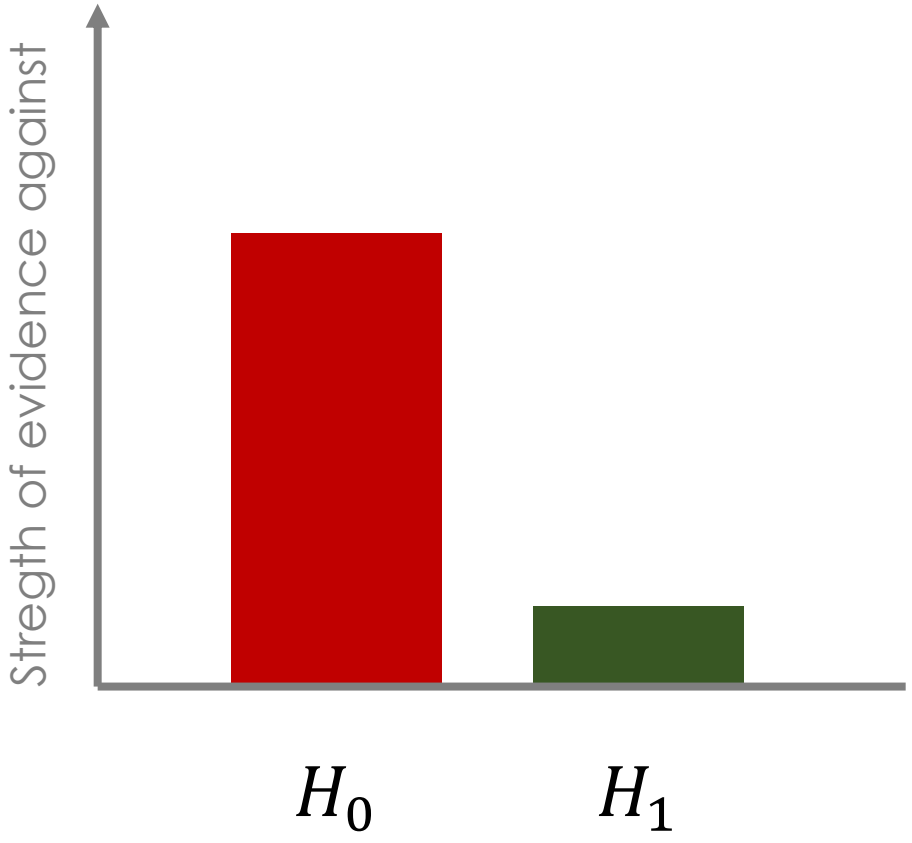
10% *5%* *1%*

Hypothesis testing



Suppose that your software outputs a p.value of 0.0001 and you have a significance level of 5%

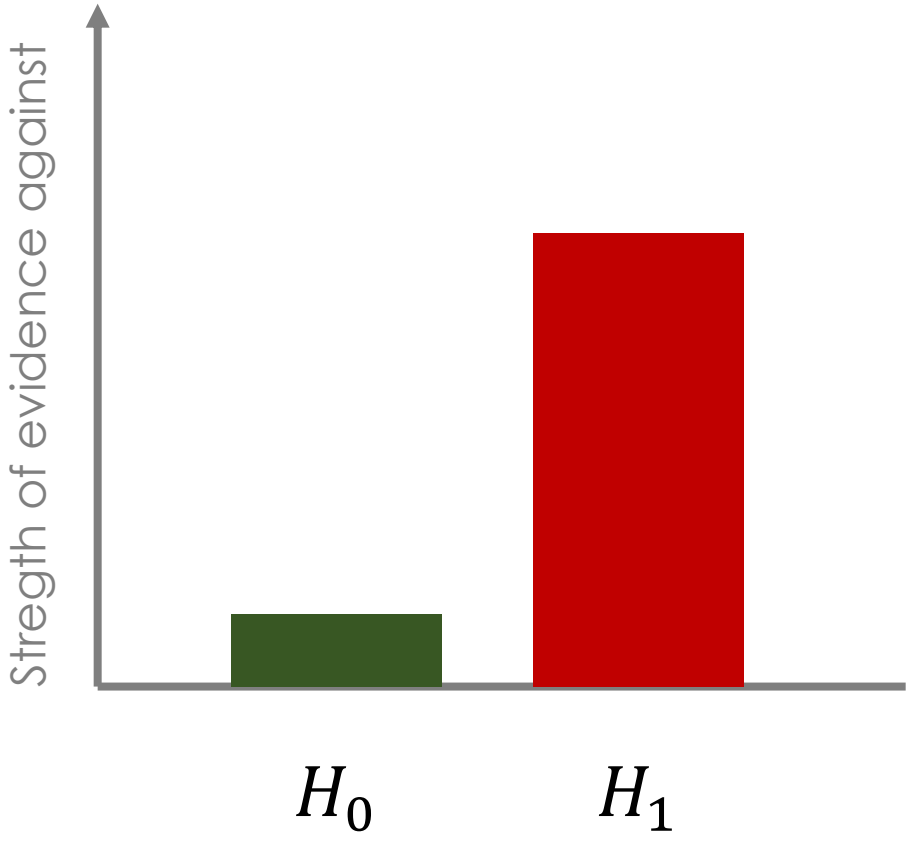
Hypothesis testing



Suppose that your software outputs a p.value of 0.0001 and you have a significance level of 5%

Your p.value is "small" or at least less than the significance level value then your evidence against the null hypothesis increases (you *accept* the alternative one)

Hypothesis testing



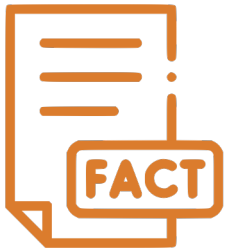
Now suppose that your software outputs a p.value of 0.07 and you have a significance level of 5%

Your p.value is not "small" or at least it is more than the significance level value. Then your evidence against the null hypothesis decreases (you do not *accept* the alternative one)

Hypothesis testing

You can state your own hypothesis testing

Information you already have and you want to reject



H_0

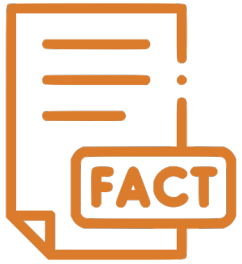
What you desire

H_1

And then you state your significance level

Hypothesis testing

Example



A school report shows that the average age in a class is 12

$$H_0: \mu \geq 12$$

but…

You suspect it is less than 12

$$H_1: \mu < 12$$

p.value outputs 0.06 and we have a significance value of 5%, then we do not have enough evidence to reject the null hypothesis, so it is clear that school is right

Statistical tests

We now that you may have heard about z-test, t-test, R^2 and all that stuff, and that those may be confusing

We want you to memorize this table

Normally Distributed?	Variance known?	Small Sample?	Large Sample?
Yes	Yes	z	z
Yes	No	t	t or z
No	Yes	n/a	z
No	No	n/a	t or z

Statistical tests

We are going to use the t-test because we already have the population mean

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where

- μ population mean
- $\bar{x}$ mean from distribution
- n size of sample
- s standar error from size

constant	slope
$H_0: a = 0$	$H_0: b = 0$
$H_a: a \neq 0$	$H_a: b \neq 0$

Statistical tests

constant

$$H_0: a = 0$$

$$H_a: a \neq 0$$

slope

$$H_0: b = 0$$

$$H_a: b \neq 0$$

What do we want to reject?

We want to reject the null hypothesis and to confirm that coefficients are significative (it means that are different from zero)

Statistical tests

What can we do if we want to evaluate the explanatory ability that a group of independent variables have over the variation of dependent variable

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$$

where

SSR Sum of squared regression
k Degrees of Freedom
SSE Sum of squared error
n total number of obs

constant

$$H_0: F = 0$$

$$H_a: F \neq 0$$

constant

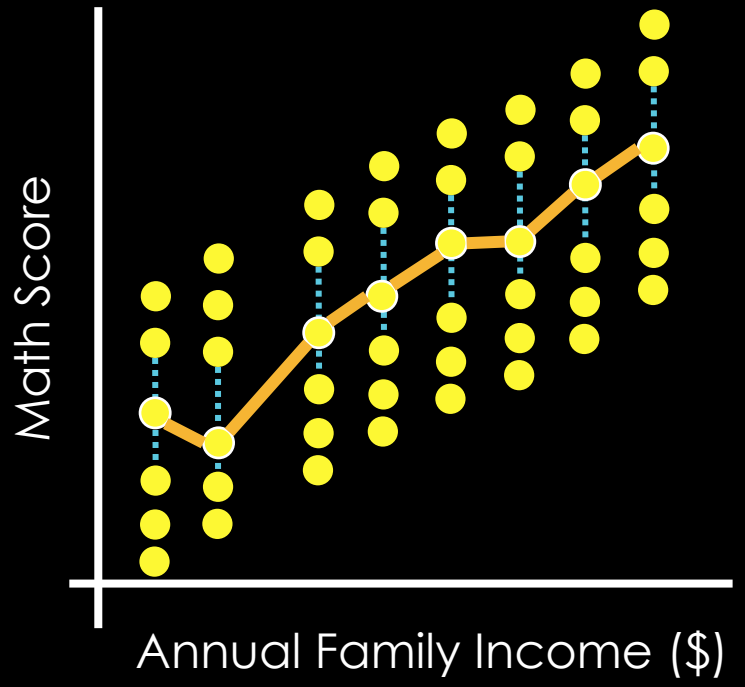
$$H_0: F = 0$$

$$H_a: F \neq 0$$

What do we want to reject?

We want to reject the null hypothesis and to confirm that our model fits the data better than using only the intercept (the null hypothesis states that the model with no independent variables fits the data well)

Statistical tests



Finally, in order to known **how well** our model is when putting it inside reality, we use **R^2** which is the **proportion of variance** in the dependent variable given a bunch of independent variables (or maybe just one)

$$0 \leq R^2 \leq 1$$

Curriculum in Cardiology
Statistics

Linear Regression Analysis Study

Khushbu Kumari, Suniti Yadav
Department of Anthropology, University of Delhi, New Delhi, India

Abstract

Linear regression is a statistical procedure for calculating the value of a dependent variable from an independent variable. Linear regression measures the association between two variables. It is a modeling technique where a dependent variable is predicted based on one or more independent variables. Linear regression analysis is the most widely used of all statistical techniques. This article explains the basic concepts and explains how we can do linear regression calculations in SPSS and excel.

Keywords: Continuous variable test, excel and SPSS analysis, linear regression

Introduction

The concept of linear regression was first proposed by Sir Francis Galton in 1894. Linear regression is a statistical test applied to a data set to define and quantify the relation between the considered variables. Univariate statistical tests such as Chi-square, Fisher's exact test, *t*-test, and analysis of variance (ANOVA) do not allow taking into account the effect of other covariates/confounders during analyses (Chang 2004). However, partial correlation and regression are the tests that allow the researcher to control the effect of confounders in the understanding of the relation between two variables (Chang 2003).

In biomedical or clinical research, the researcher often tries to understand or relate two or more independent (predictor) variables to predict an outcome or dependent variable. This may be understood as how the risk factors or the predictor variables or independent variables account for the prediction of the chance of a disease occurrence, i.e., dependent variable. Risk factors (or dependent variables) associate with biological (such as age and gender), physical (such as body mass index and blood pressure [BP]), or lifestyle (such as smoking and alcohol consumption) variables with the disease. Both correlation and regression provide this opportunity to understand the "risk factors-disease" relationship (Gaddis and Gaddis 1990). While correlation provides a quantitative way of measuring the degree or strength of a relation between two variables, regression analysis mathematically describes this relationship. Regression analysis allows predicting the value of a dependent variable based on the value of at least one independent variable.

In correlation analysis, the correlation coefficient "r" is a dimensionless number whose value ranges from −1 to +1. A value toward −1 indicates inverse or negative relationship, whereas towards +1 indicate a positive relation. When there is a normal distribution, the Pearson's correlation is used, whereas, in nonnormally distributed data, Spearman's rank correlation is used.

The linear regression analysis uses the mathematical equation, i.e., $y = mx + c$, that describes the line of best fit for the relationship between y (dependent variable) and x (independent variable). The regression coefficient, i.e., r^2 implies the degree of variability of y due to x.[1-8]

Significance of Linear Regression

The use of linear regression model is important for the following reasons:

a. Descriptive – It helps in analyzing the strength of the association between the outcome (dependent variable) and predictor variables
b. Adjustment – It adjusts for the effect of covariates or the confounders

Address for correspondence: Khushbu Kumari,
Department of Anthropology, University of Delhi, New Delhi, India.
E-mail: khushukumari38@gmail.com

Access this article online

Quick Response Code:

Website: www.j-pcs.org

DOI: 10.4103/jpcs.jpcs_8_18

How to cite this article: Kumari K, Yadav S. Linear regression analysis study. J Pract Cardiovasc Sci 2018;4:33-6.

How to read a
paper

Khushbu Kumari, Suniti Yadav
Department of Anthropology, University of Delhi, New Delhi, India

Abstract

Linear regression is a statistical procedure for calculating the value of a dependent variable from an independent variable. Linear regression measures the association between two variables. It is a modeling technique where a dependent variable is predicted based on one or more independent variables. Linear regression analysis is the most widely used of all statistical techniques. This article explains the basic concepts and explains how we can do linear regression calculations in SPSS and excel.

Keywords: Continuous variable test, excel and SPSS analysis, linear regression

Introduction

The concept of linear regression was first proposed by Sir Francis Galton in 1894. Linear regression is a statistical test applied to a data set to define and quantify the relation between the considered variables. Univariate statistical tests such as Chi-square, Fisher's exact test, *t*-test, and analysis of variance (ANOVA) do not allow taking into account the effect of other covariates/confounders during analyses (Chang 2004). However, partial correlation and regression are the tests that allow the researcher to control the effect of confounders in the understanding of the relation between two variables (Chang 2003).

In biomedical or clinical research, the researcher often tries to understand or relate two or more independent (predictor) variables to predict an outcome or dependent variable. This may be understood as how the risk factors or the predictor variables or independent variables account for the prediction of the chance of a disease occurrence, i.e., dependent variable. Risk factors (or dependent variables) associate with biological (such as age and gender), physical (such as body mass index and blood pressure [BP]), or lifestyle (such as smoking and alcohol consumption) variables with the disease. Both correlation and regression provide this opportunity to understand the "risk factors-disease" relationship (Gaddis and Gaddis 1990). While correlation provides a quantitative way of measuring the degree or strength of a relation between two variables, regression analysis mathematically describes this relationship. Regression analysis allows predicting the value

Conclusion

The techniques for testing the relationship between two variables are correlation and linear regression. Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation. In this article, we have used simple examples and SPSS and excel to illustrate linear regression analysis and encourage the readers to analyze their data by these techniques.

How to read a **paper**

Begin with:

Abstract

Keywords

Authors

Introduction

Conclusion

Note:
· Reset your expectations, do not worry if you do not understand the first time

f. All the values of "y" are independent from each other, though dependent on "x."

Table 6: Summary output

Regression statistics	Values	Explanation
Multiple *R*	0.96332715	Correlation coefficient: 1 means perfect correlation and 0 means none
R^2	0.927999198	Coefficient of determination: How many points fall on the regression line. Here, 92% points fall within the line
Adjusted R^2	0.891998797	Adjusted R^2: Adjusts for multiple variables, use with multiple variables
SE	516.3490153	
Observations	7	

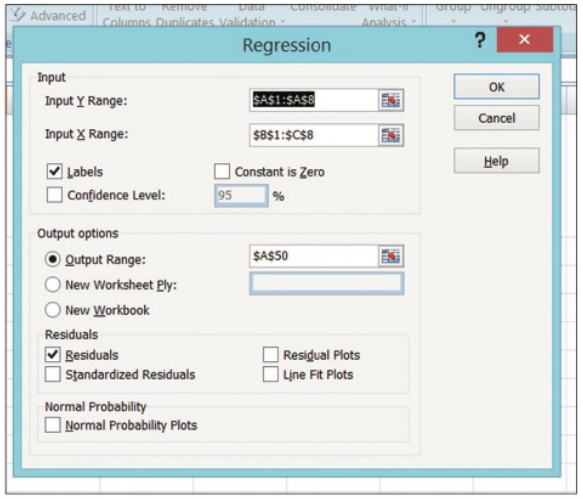


Figure 4: The regression screen. Choose Data > Data Analysis > Regression. Input y Range: A1:A8. Input X Range: B1:C8. Check Labels, Residuals, Output Range as A50.

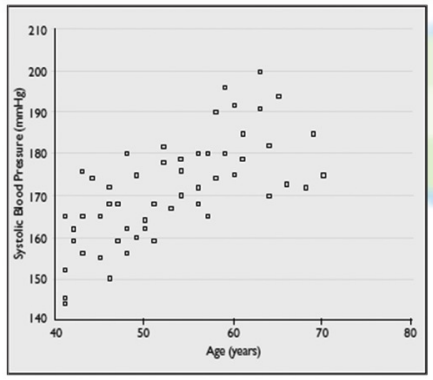


Figure 1: Scatter plot of systolic blood pressure versus age.

References

1. Schneider A, Hommel G, Blettner M. Linear regression analysis: Part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010;107:776-82.
2. Freedman DA. Statistical Models: Theory and Practice. Cambridge, USA: Cambridge University Press; 2009.
3. Chan YH. Biostatistics 201: Linear regression analysis. Age (years). Singapore Med J 2004;45:55-61.
4. Chan YH. Biostatistics 103: Qualitative data – Tests of independence. Singapore Med J 2003;44:498-503.
5. Gaddis ML, Gaddis GM. Introduction to biostatistics: Part 6, correlation and regression. Ann Emerg Med 1990;19:1462-8.
6. Mendenhall W, Sincich T. Statistics for Engineering and the Sciences. 3rd ed. New York: Dellen Publishing Co.; 1992.
7. Panchenko D. 18.443 Statistics for Applications, Section 14, Simple Linear Regression. Massachusetts Institute of Technology: MIT OpenCourseWare; 2006.
8. Elazar JP. Multiple Regression in Behavioral Research: Explanation and Prediction. 2nd ed. New York: Holt, Rinehart and Winston; 1982.

Table 4: SPSS equation variables

Coefficients[a]							
Model	Unstandardised coefficients		Standardised coefficents	*t*	Sig.	95% Confidence interval for B	
	B	Std. error	Beta			Lower bound	Upper bound
I (Constant)	115.706	7.999	0.696	14.465	0.000	99.662	131.749
Age (years)	1.051	0.149		7.060	0.000	0.752	1.350

[a]Dependent variable: Systolic blood pressure (mmHg)

How to read a **paper**

When you master that keep up with:

Technical language

Diagrams/graphs

Tables (data)

Results

Software

References

Note:
- Highlight key concepts
- Look through references

How to Calculate Linear Regression?

Linear regression can be tested through the SPSS statistical software (IBM Corp. Released 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.) in five steps to analyze data using linear regression. Following is the procedure followed Tables 1-4:

Click Analyze > Regression > Linear > then select Dependent and Independent variable > OK (enter).

Example 1 – Data (n = 55) on the age and the SBP were collected and linear regression model would be tested to predict BP with age. After checking the normality assumptions for both variables, bivariate correlation is tested (Pearson's correlation = 0.696, $P < 0.001$) and a graphical scatter plot is helpful in that case [Figure 2].

Now to check the linear regression, put SBP as the dependent and age as the Independent variable.

This indicates the dependent and independent variables included in the test.

Pearson's correlation between SBP and age is given ($r = 0.696$). R^2 = 0.485 which implies that only 48.5% of the SBP is explained by the age of a person.

The ANOVA table shows the "usefulness" of the linear regression model with $P < 0.05$.

Finally:

Read it more than once

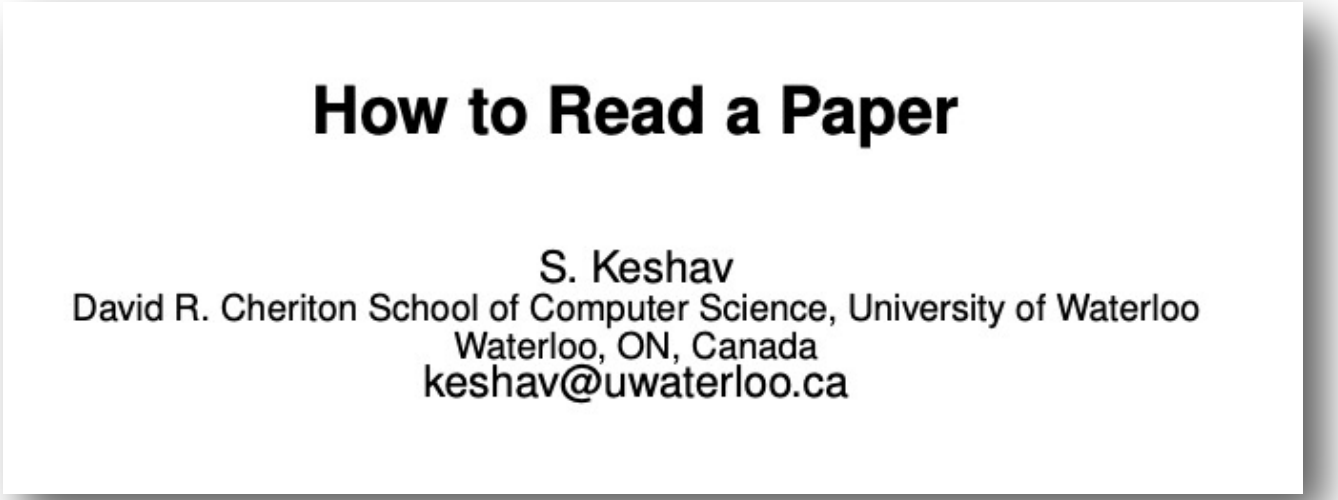
Check details

Replicate it

Note:
· Look for technicalities you don't understand

How to read a **paper**

Read the paper titled "How to Read a Paper"



How to Read a Paper

S. Keshav
David R. Cheriton School of Computer Science, University of Waterloo
Waterloo, ON, Canada
keshav@uwaterloo.ca

How to read a **paper**

You cand search for free papers in several websites,
but you may know the characteristics between them

Scientific papers

- Author (well-known)
- Long process to be accepted and published
- Peer reviewed (peer to peer)

White papers

- Author (Anonymity)
- Anyone can submit a paper
- Community review

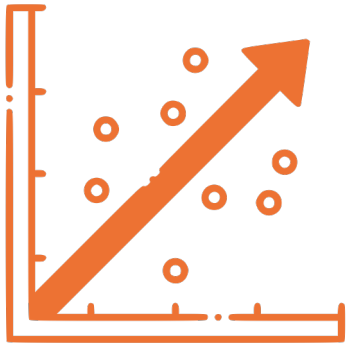
How to read a **paper**

White papers

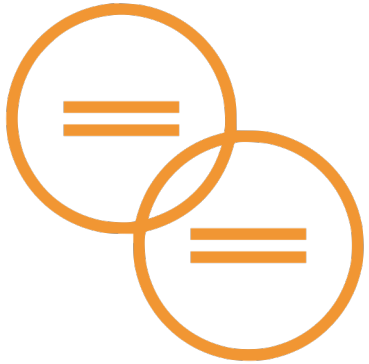
Even when the article or paper is not peer to peer review that does not mean there is no quality in them

Gregory Perelman submitted an article in ArXiv with the resolution of the Poincaré conjecture (one of the millennium problems)

Assumptions associated with **linear regression**



Linearity



Homoscedasticity

Normality

Independence

Terminology

y	x
Dependent variable	Explanatory variable
Explained variable	Independent variable
Predictand	Predictor
Regressand	**Regressor**
Response	Stimulus
Endogenous	Exogenous
Outcome	Convariate
Controlled variable	Control variable

Practice 1

The data set in CEOSAL2 contains information on chief executive officers for U.S. corporations. The variable *salary* is annual compensation, in thousands of dollars, and *ceoten* is prior number of years as company CEO.

(i) Find the average salary and the average tenure in the sample.

(ii) How many CEOs are in their first year as CEO (that is, *ceoten* 5 0)? What is the longest tenure as a CEO?

(iii) Estimate the simple regression model

$$\log(salary) = \beta_0 + \beta_1 ceoten + u$$

and report your results in the usual form. What is the (approximate) predicted percentage increase in salary given one more year as a CEO?

STATA COMMANDS

```
1. mean salary ceoten
2. Tabstat ceoten, satistics (min max)
3. reg salary ceoten
```

Practice 2

Use the data in SLEEP75 from Biddle and Hamermesh (1990) to study whether there is a tradeoff between the time spent sleeping per week and the time spent in paid work. We could use either variable as the dependent variable. For concreteness, estimate the model .

$$sleep = \beta_0 + \beta_1 totwrk + u$$

where *sleep* is minutes spent sleeping at night per week and *totwrk* is total minutes worked during the week.

(i) Report your results in equation form along with the number of observations and R . What does the intercept in this equation mean?

(ii) If *totwrk* increases by 2 hours, by how much is *sleep* estimated to fall? Do you find this to be a large effect?

STATA COMMANDS

```
1. edit
2. .sum
3. reg wage educ
4. reg wage educ exper
```

References

- **Arthur S. Goldberger**, (1946), *Econometric Theory,* John Wiley & Sons, New York, p. 1.
- **Kumari K., J. Pract Cardiovasc, Wooldridge, J.** (2020). *Introductory econometrics : a modern approach*. Boston, MA: Cengage. Gujarati, D. & Porter, D. (2009). *Basic econometrics*. Boston: McGraw-Hill Irwin.
- **Salvatore, D., & Sarmiento, J. C.** (1983). *Econometría* (No. HB141 S39). McGraw-Hill.
- **Keshav, S.** (2007). *How to read a paper*. In ACM SIGCOMM Computer Communication Review (Vol. 37, Issue 3, pp. 83–84). Association for Computing Machinery (ACM). https://doi.org/10.1145/1273445.1273458
- **Salt Solutions** (2021), "salt-cfa-level-1-formulasheet"