Multiple Linear Regression

Panel Data Econometrics

Prof. Alexandre Gori Maia

State University of Campinas

Ementa

Linear Regression – Definition

Ordinary Least Squares

Analysis of Variance

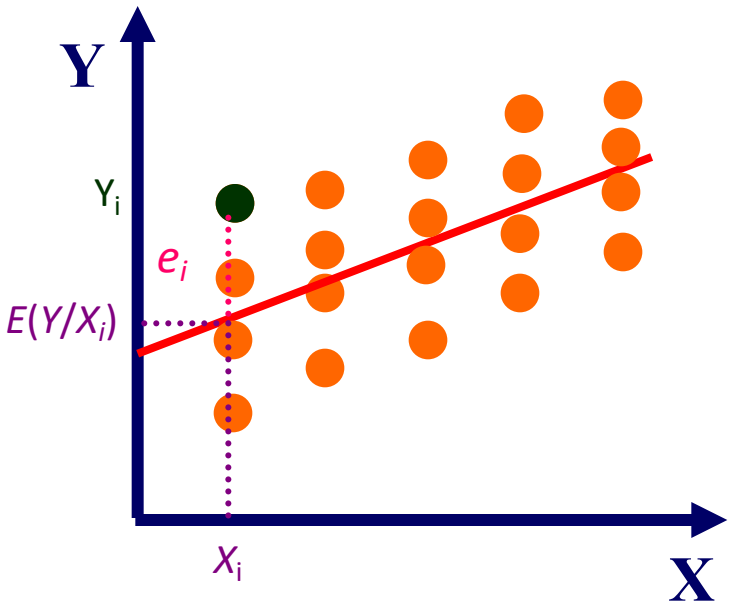
Test *t* and *F*

Reference

Maia, A. G. 2014. Econometria: conceitos e aplicações. Insitituto de Economica. Caps. 1 a 8.

Population Regression Function

Suppose the relation between *Y* and *X* in the population:



$$Y_i = \alpha + \beta X_i + e_i$$

or

$$E(Y/X_i) = \alpha + \beta X_i$$

Model of Simple Linear Regression for *Y* in the population.

Where:

Y is the dependent variable (regressand)

X is the independent variable (regressor)

α is the intercept

β is the slope coefficient

Prediction error:

Given that X_i is the i-th observation of X, then:

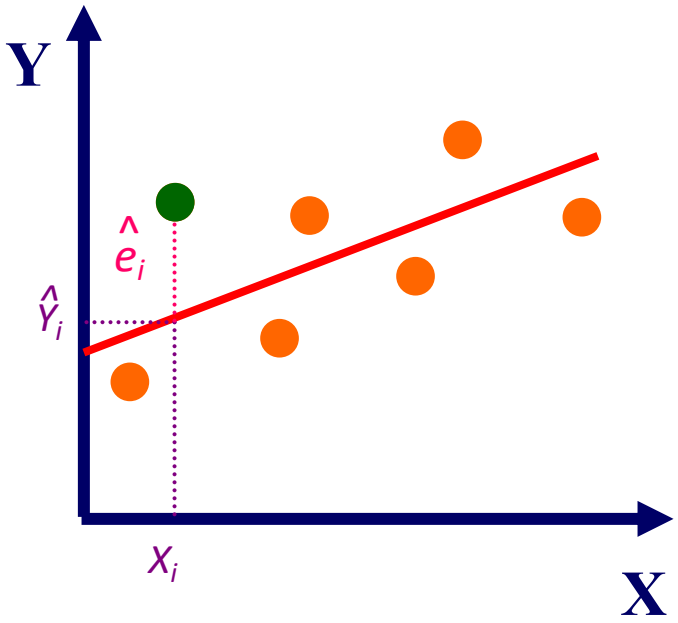
Y_i is the observed value of Y for the i-th observation

$E(Y/X_i)$ is the conditional expectation of Y for the i-th observation, given X

e_i is the error, or the variation of Y_i that is not explained by the model

Ordinary Least Squares (OLS)

The relation between *Y* and *X* in the sample is given by:



Sample Regression Function:

$$Y_i = \hat{\alpha} + \hat{\beta}X_i + \hat{e}_i$$

***P*redicted *Y* (fitted value):**

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$$

Residual (sample error):

$$\hat{e}_i = Y_i - \hat{Y}_i$$

Error Sum of Squares (ESS):

$$EQT = \hat{e}_1^{\,2} + \hat{e}_2^{\,2} + \ldots + \hat{e}_n^{\,2}$$

$$EQT = (Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + \ldots + (Y_n - \hat{Y}_n)^2$$

$$EQT = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}[Y_i - (\hat{\alpha} + \hat{\beta}X_i)]^2$$

OLS – Matrix Form

- The OLS estimates minimize the error sum of squares;

Given the function: $Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e_i$

In matrix form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$

Which means:

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{pmatrix}}_{\mathbf{y}_{n\times 1}} = \underbrace{\begin{pmatrix} 1 & X_{1_1} & X_{2_1} & \ldots & X_{k_1} \\ 1 & X_{1_2} & X_{2_2} & \ldots & X_{k_2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_{1_n} & X_{2_n} & \ldots & X_{k_n} \end{pmatrix}}_{\mathbf{X}_{n\times p}} \underbrace{\begin{pmatrix} \alpha \\ \beta_1 \\ \ldots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}_{p\times 1}} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{pmatrix}}_{\mathbf{e}_{n\times 1}}$$

Minimizing the ESS: $\dfrac{\partial EQT}{\partial \hat{\boldsymbol{\beta}}} = 0 \;\Rightarrow\; \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y})$

Example

- The dataset Data_CO2.csv contains information on CO^2 (tons per capita), GDP (US$ per capita) and the % of manufacturing in the GDP;
- OLS estimates in Stata:

```
* import dataset with information for CO2, GDP and % manufacturing
insheet using "Z:\HO-235\Data\Data_CO2.csv", clear

* linear regression by ols
regress co2 gdp ind
```

- OLS estimates in R:

```
# import dataset with information for CO2, GDP and % manufacturing
countries <- read.csv("Z:/HO-235/Data/Data_CO2.csv")

# linear regression by ols
ols <- lm(co2 ~ gdp + ind, data=countries)
summary(ols)
```

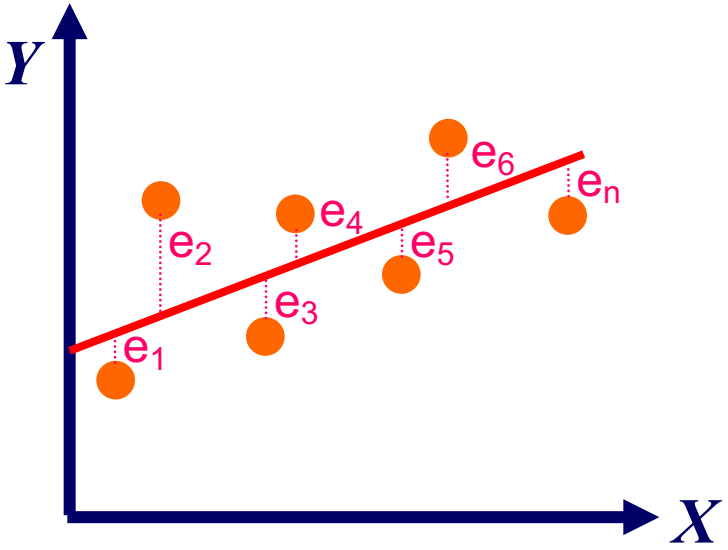
Example

- OLS estimates in Python:

```
# package for data analysis
import pandas as pd

# import dataset with information for CO2, GDP and % manufacturing
countries = pd.read_csv("Z:/HO-235/Data/Data_CO2.csv")

# package for linear regression
import statsmodels.api as sm

# define matrix y and x
x = countries[['gdp','ind']]
y = countries['co2']

# add a constant (1) to matrix x
x = sm.add_constant(x)

# linear regrression by ols
ols = sm.OLS(y, x).fit()
print(ols.summary())
```

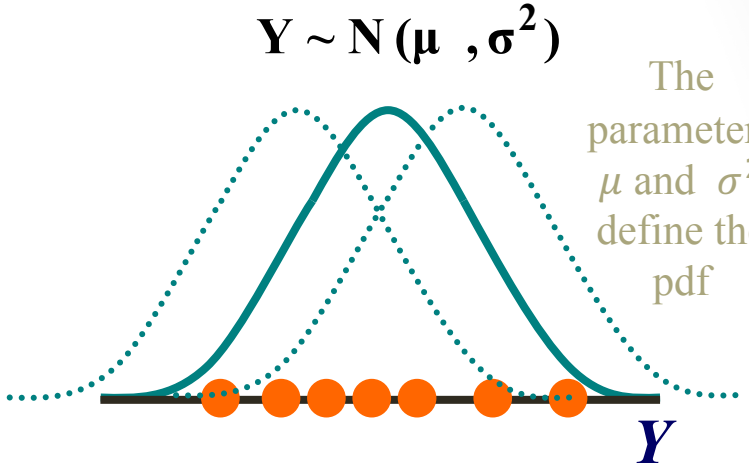
Maximum Likelihood (ML)

OLS



We must define the function $Y=f(X)+e$. Once we know $f(X)$, we can estimate its parameters by minimizing the the sum of $[Y-f(X)]^2$.

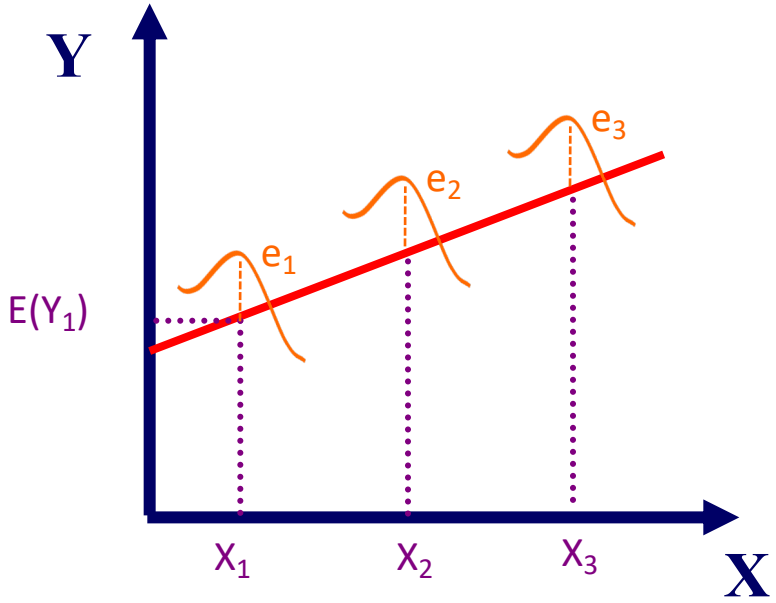
ML



We must know the probability distribution (or density) function (pdf) of Y. Once we know the pdf, we can estimate its parameter by maximizing the likelihood of the observed values of Y belonging to the pdf.

The Distribution of Errors

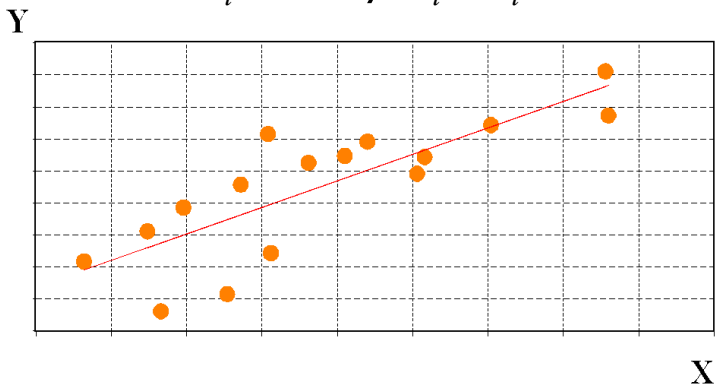
- Given the model $Y = \alpha + \beta X + e$
- The errors are assumed to be independent of each other and identically distribution according to a normal distribution (independent and identically distributed, or i.i.d.);
- If we know the dpf of the errors, we can apply the ML method to estimate the regression line that better fit to the observed data;



Interpreting the Coefficients

Simple Linear Regression

$$Y_i = \alpha + \beta X_i + e_i$$



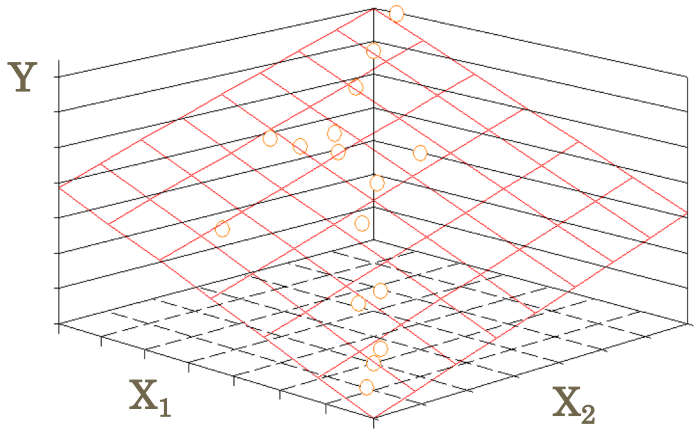
We have:

$E[Y / X = 0] = \alpha$ Expected value of *Y* when *X*=0.

$\frac{dY}{dX} = \beta$ Marginal impact (effect) of *X* on Y.

Multiple Linear Regression

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + e_i$$



We have

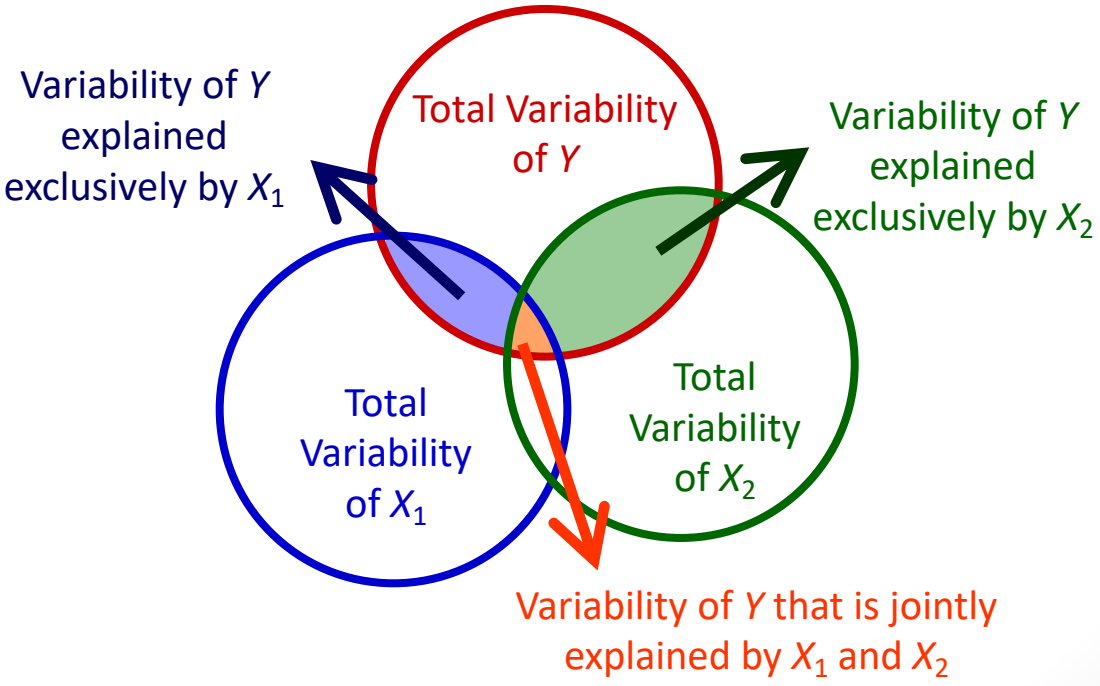
$E[Y / X_1 = 0, X_2 = 0] = \alpha$ Expected value of *Y* when X_1*=0 and* X_2*=0*.

$\frac{\partial Y}{\partial X_1} = \beta_1$ Marginal impact of X_1 on *Y,* holding constant X_2 (partial or net effect of X_1)

$\frac{\partial Y}{\partial X_2} = \beta_2$ Marginal impact of X_2 on *Y,* holding constant X_1 (partial or net effect of X_2)

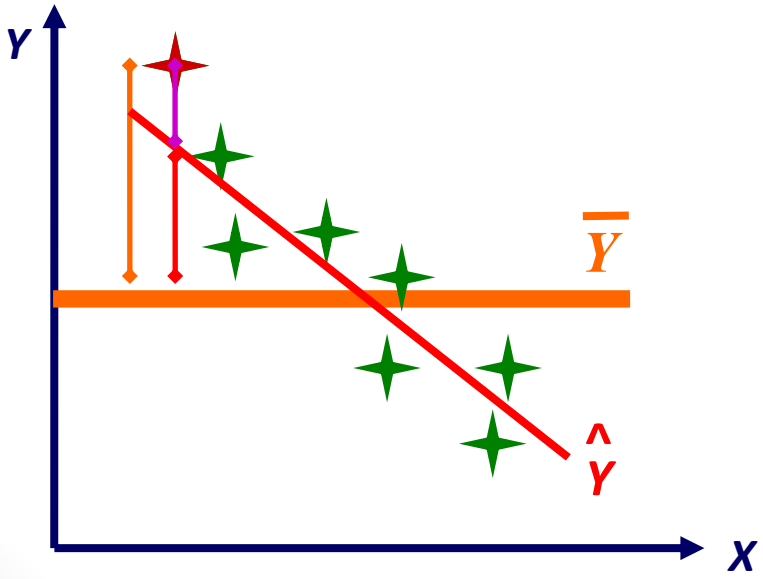
Analysis of Variance

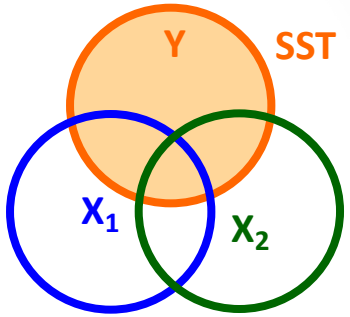
- The total variability of Y represents the diversity of values that Y can assume;
- A share of the total variability of *Y* can be explained exclusively by X_1, another share can be explained exclusively by X_2 and another share can be jointly explained by X_1 and X_2;
- The variability of Y that is not explained by X_1 and X_2 are represented by variability of the errors;

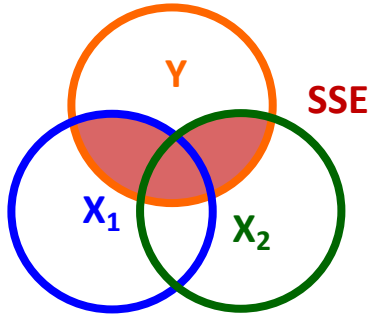


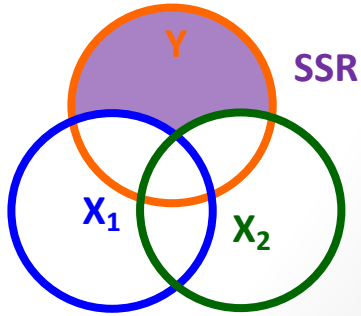
Sum of Squares

- Total Sum of Squares (SST) is a statistical measure for the total variability of *Y*;
- Explained Sum of Squares (SSE) is a statistical measure for the variability of *Y* that is explained by the independent variables;
- Sum of Squared Residuals (SSR) is a statistical measure for the variability of *Y* that is not explained by the independent variables;





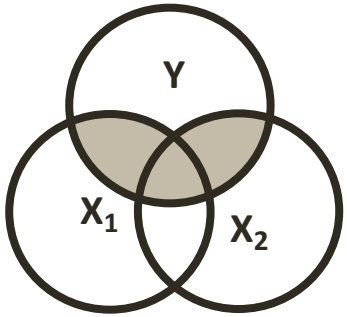




Coefficient of Determination

- The fraction of the sample variation of *Y* that is explained by the independent variables (R-squared);

$$R^2 = \frac{SQ\,\mathrm{Re}\,g}{STQ} = 1 - \frac{SQ\,\mathrm{Re}\,s}{STQ}$$



Scale of measurement:

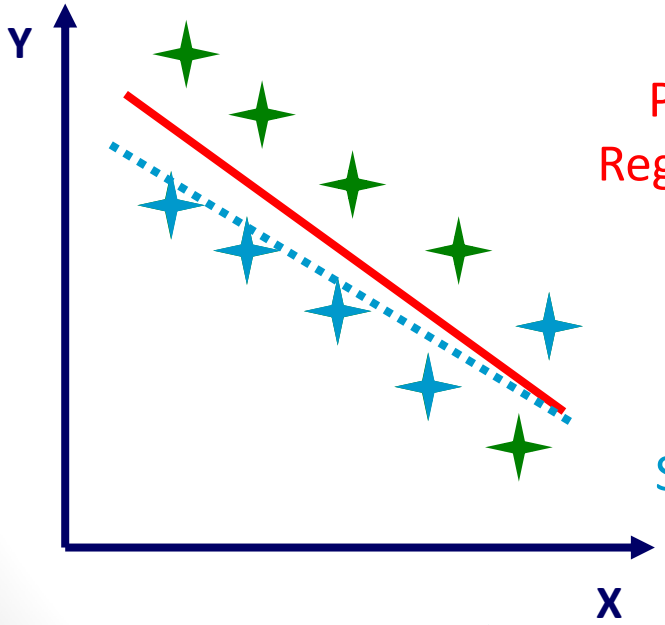
0 Independence

Perfect Linear Relation 1

Sampling Distribution

- While the parameter β is a constant, the estimate $\hat{\beta}$ is a random variable;
- This means that $\hat{\beta}$ can assume any value according to a sampling distribution;

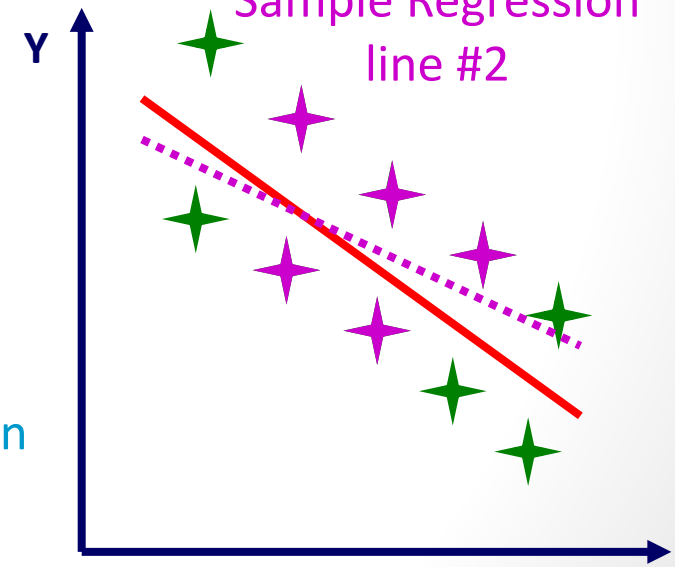
Suppose a population with 10 observations (✦)



$\beta_0+\beta_1X$
Population
Regression Line

$\hat{\beta}_0+\hat{\beta}_1X$
Sample Regression
Line #1

Random sample #1 (✦)



$\hat{\beta}_0+\hat{\beta}_1X$
Sample Regression
line #2

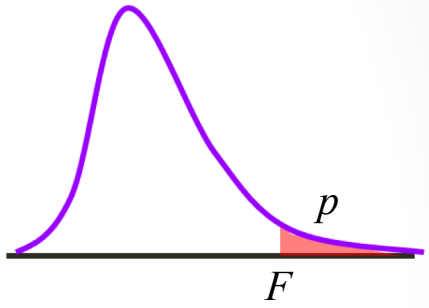
Random sample #2 (✦)

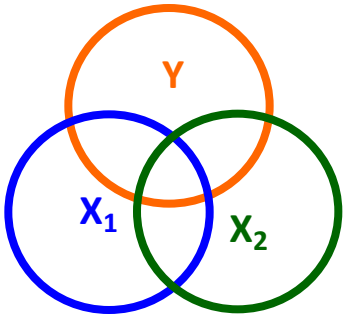
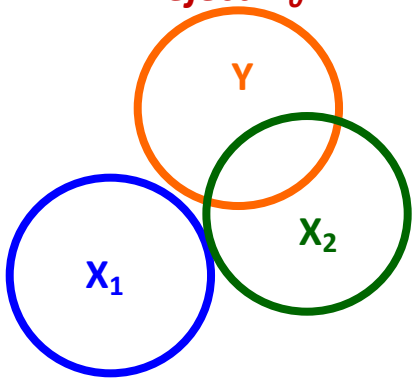
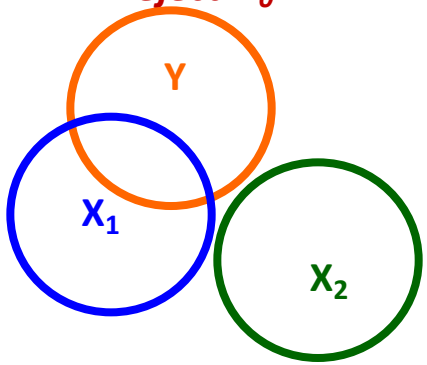
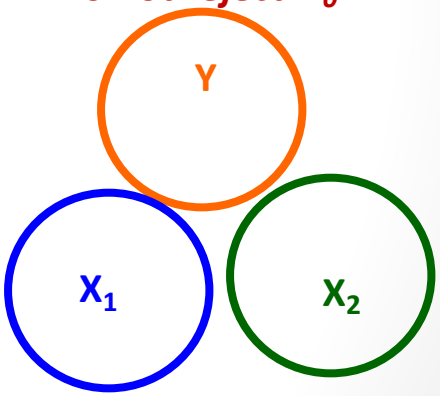
F test

- The *F* statistic is used to estimate the probability of error (p-value) if we assume that the model does explain the variability of *Y*;

Given the model: $Y = \alpha + \beta_1 X_1 + ... + \beta_k X_k + e$

And the hypotheses: $\begin{cases} H_0: \ \beta_1 = ... = \beta_k = 0 \\ H_1: \ \text{Pelo menos um } \beta_k \neq 0 \end{cases}$



Reject H_0	Reject H_0	Reject H_0	Do not reject H_0
			
$\beta_1 \neq 0$ $\beta_2 \neq 0$	$\beta_1 = 0$ $\beta_2 \neq 0$	$\beta_1 \neq 0$ $\beta_2 = 0$	$\beta_1 = 0$ $\beta_2 = 0$
X_1 and X_2 explain Y. H_0 must be rejected.	Only X_2 explains Y. H_0 must be rejected.	Only X_1 explains Y. H_0 must be rejected	Neither X_1 nor X_2 explain Y. H_0 must not be rejected.

ANOVA

- Summarize the results of the analysis of variability;
- Small values of *p* (usually smaller than 0.05) suggest that the model explains significantly the variability of *Y* ($R^2 > 0$);

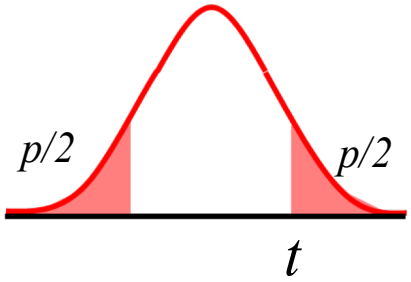
Source	**df**	**SS**	**MS**	***F***	***p***
Regression	k	$\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y} - n\bar{Y}^2$	$\frac{\text{SQReg}}{k}$	$\frac{\text{QMReg}}{\text{QMRes}}$	*p-value*
Residuals	$n-(k+1)$	$\mathbf{y}^T \mathbf{y} - \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y}$	$\frac{\text{SQRes}}{n-(k+1)}$		
Total	$n-1$	$\mathbf{y}^T \mathbf{y} - n\bar{Y}^2$			

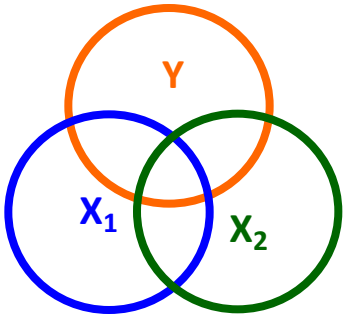
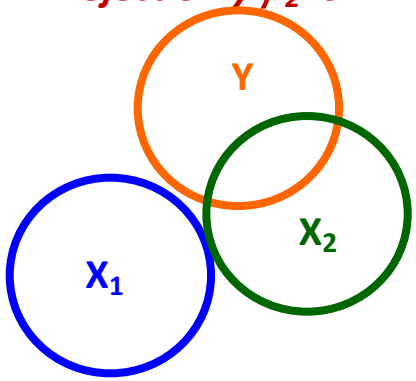
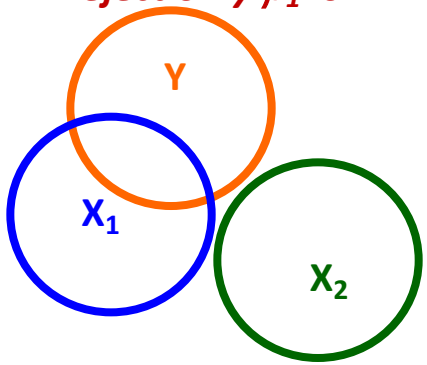
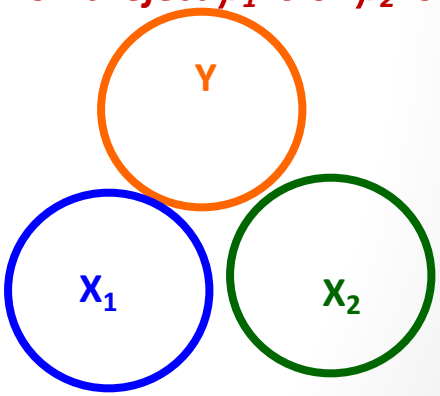
t test

- The Student's *t* statistic is used to estimate the probability of error (*p*-value) if we assume that one specific variable (*Xj*) explains partially the variability of *Y*;

Given the model: $Y = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k + e$

And the hypotheses: $\begin{cases} H_0: \ \beta_j = 0 \\ H_1: \ \beta_j \neq 0 \end{cases}$



Reject β_1=0 and β_2=0	***Reject only β_2=0***	***Reject only β_1=0***	***Don't reject β_1=0 or β_2=0***
			
$\beta_1 \neq 0$ **$\beta_2 \neq 0$**	**$\beta_1 = 0$** **$\beta_2 \neq 0$**	**$\beta_1 \neq 0$** **$\beta_2 = 0$**	**$\beta_1 = 0$** **$\beta_2 = 0$**
X_1 and X_2 explain *Y*. H_0 must be rejected for both $\beta_1 = 0$ and $\beta_2 = 0$	Only X_2 explains *Y*. H_0 :β_2=0 must be rejected	Only X_1 explains *Y*. H_0 :β_1=0 must be rejected	Neither X_1 nor X_2 explain *Y*. None *t* test must be rejected.

Exercise

1) The dataset Data_TravelCosts.csv contains information on travel costs from several municipalities to a national park in Brazil (see *MAIA, A. G. , ROMEIRO, A. . Validade e confiabilidade do método de custo de viagem: um estudo aplicado ao Parque Nacional da Serra Geral. Revista de Economia Aplicada, v. 12, p. 103-123, 2008*):
   a) Analyze the relation between travel costs and the visit rate;
   b) Add control variables in the linear model;
   c) Are the estimates consistent with the microeconomic theory?