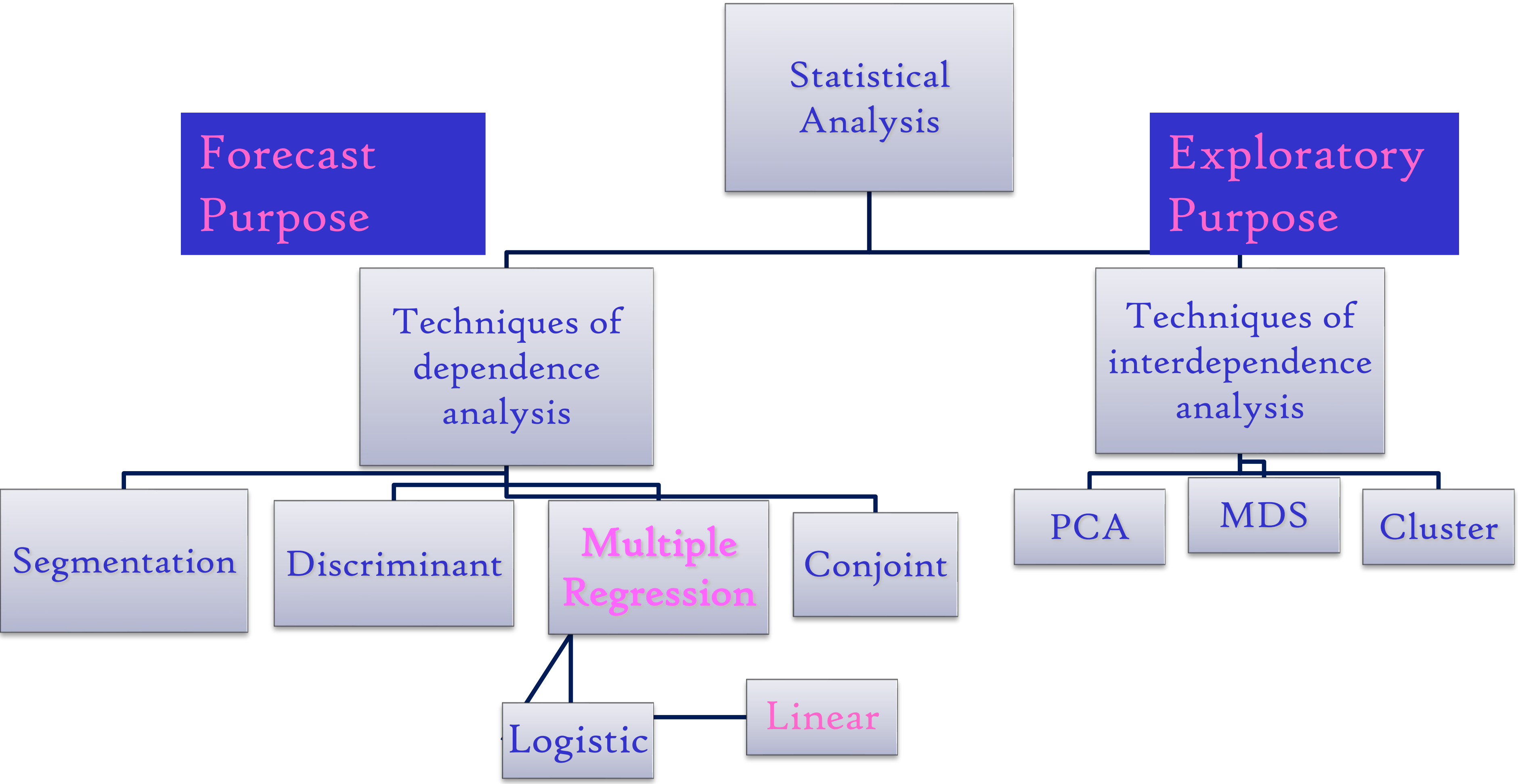
Main quantitative analysis techniques



Multivariate Statistical Analysis

Multiple linear regression

Introduction

The regression is a technique of statistical analysis that has the aim to identify the relationship between a dependent variable and one or a set of explanatory variables

- SIMPLE REGRESSION $Y=f(X)$
- MULTIPLE REGRESSION $Y=f(X_1,X_2,\ldots X_k)$

Multiple linear regression model

It expresses a linear relationship between a dependent variable (Y) and a set of explanatory variables (Xi) or regressors.

k explanatory variables

Intercept

Regression coefficients

error= normal random variable

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

Assumptions of multiple regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

1. Linear relationship between Y and X_i
2. Non-stochastic explanatory variables X_i
3. The expected value of the error is $\longrightarrow E(\varepsilon_i)=0$
4. The error variance is finite and constant (homoskedasticity) $E(\varepsilon_i\ \varepsilon_i)=\sigma^2$ for all i
5. The errors are not related $\longrightarrow COV(\varepsilon_i, \varepsilon_{i-k}) = E(\varepsilon_i, \varepsilon_{i-k})=0$ for all i and K
6. The errors are normally distributed $\longrightarrow$ $N(0, \sigma^2)$
7. The regressors are not related to each other $\longrightarrow$ no multicollinearity

Multiple linear regression equation

Estimated value for predicted dependent variable

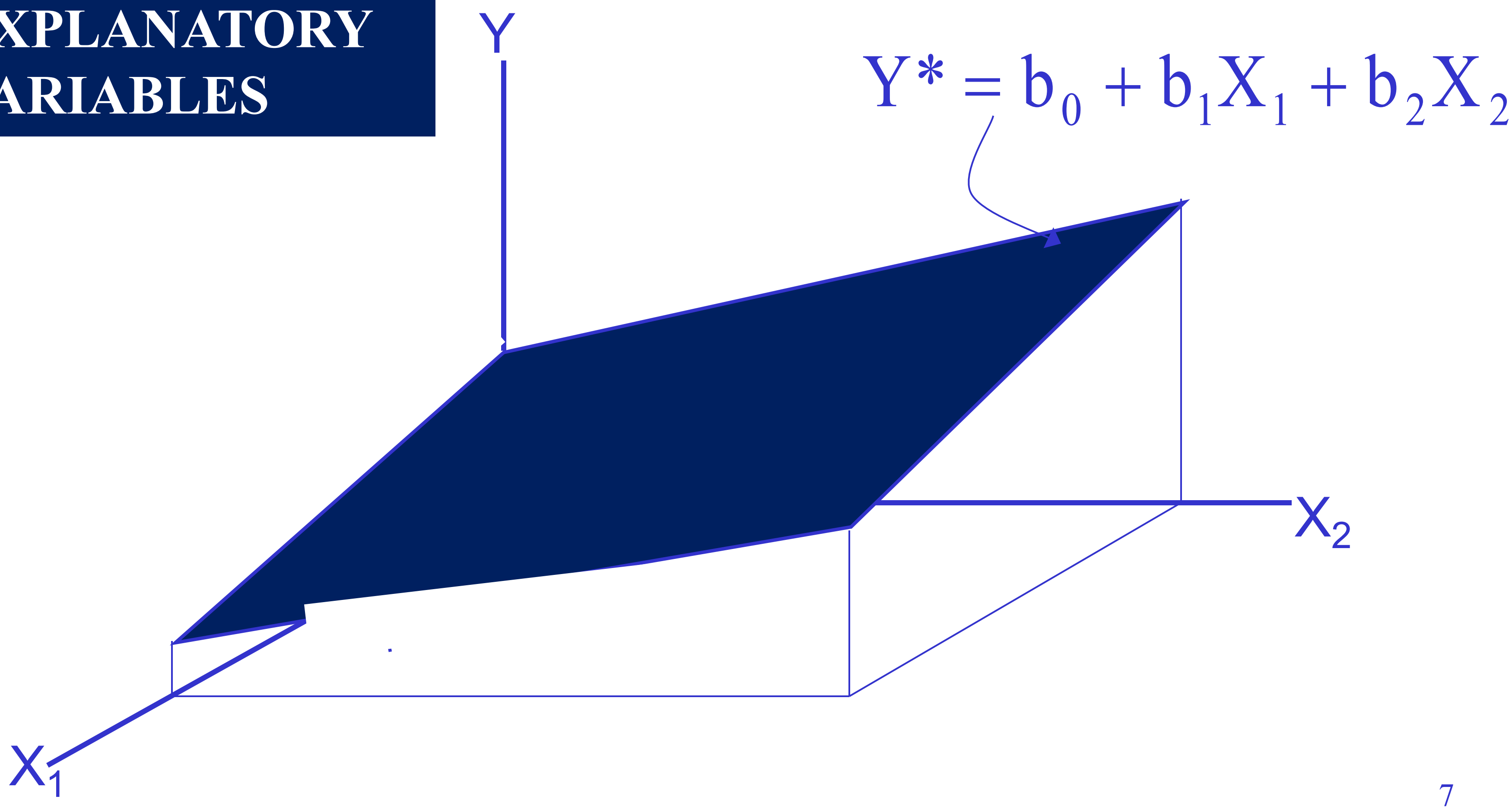
Intercept estimation

Regression coefficient estimations

$$Y_i^* = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik}$$

Graphical representation

EXAMPLE WITH TWO EXPLANATORY VARIABLES



Matrix notation

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_i \\ \cdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{12} & X_{1k} \\ 1 & X_{22} & X_{2k} \\ \cdots & \cdots & \cdots \\ 1 & X_{i2} & X_{ik} \\ \cdots & \cdots & \cdots \\ 1 & X_{n2} & X_{nk} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_j \\ \cdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_i \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

(n x 1) (n x k) (k x 1) (nx1)

Ordinary Least-Squares (OLS)

$$\boxed{y = X\beta + \varepsilon}$$

$$\boxed{y^* = Xb}$$

$$\boxed{e = y - y^* = Xb}$$

$$\boxed{\sum e_i^{\ 2} = e'e = (y - Xb)'(y - Xb)}$$

Ordinary Least-Squares (OLS)

Imposing the derivative with respect to the k coefficients equal to zero (b), we obtain the estimation of betas:

$$e'e = (y - Xb)'(y - Xb) = y'y - y'Xb - b'X'y + b'X'Xb =$$
$$= y'y - 2b'X'y + b'X'Xb$$
$$min_b(e'e) = \frac{\partial(e'e)}{\partial b} = -2X'y + 2X'Xb = 0$$
$$then$$
$$b = (X'X)^{-1}X'y$$

T-test on individual regression coefficient

To verify the significance of each parameter included in the model

$$\begin{cases} H_0: \beta_j = 0 \text{ (the variable X has no influence on Y)} \\ H_1: \beta_j \neq 0 \end{cases}$$

Under the hypothesis of normally distributed errors, the statistical test is

$$\frac{b_j}{S_b} \sim t_{n-k} \quad \text{where } S_b\text{=Standard Error of } \beta_j$$

The null hypothesis will be rejected (accepted) if t_{n-k} is outside the range delimitated by the tabulated values of Student't distribution corresponding to +/- $t_{n-k,\, \alpha/2}$

R-square

In order to verify the goodness of fit of the model, we look at R square value, given by the following formulations

$$R^2 = \frac{DEV(R)}{DEV(Y)} = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\sum_i (y_i^* - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Or, equivalentely, by

$$R^2 = 1 - \frac{DEV(E)}{DEV(Y)} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\sum_i (y_i - y_i^*)^2}{\sum_i (y_i - \bar{y})^2}$$

Adjusted R-square

R-square increases with the number of explanatory variables included in the model. To avoid this effect it is corrected in the following way

$$\boxed{\overline{R}^2 = 1 - \frac{SSE / n-k}{SST / n-1} = 1 - \left[(1 - R^2)\left(\frac{n-1}{n-k} \right) \right]}$$

n= sample size

k= number of parameters

Adjusted R^2 is less than R^2

It is used in the comparison between regression models with the same dependent variable and a different number of explanatory variables

F- test on all regression coefficients

Check the overall goodness of fit of a model, simultaneausly, on all regression coefficients

$$\begin{cases} H_0\text{: } \beta_1 = \beta_2 = \ldots = \beta_j = \ldots = \beta_k = 0 \text{ (there is no linear relationship between Y e le Xi)} \\ H_1\text{: otherwise (at least one explanatory variable Xi influence Y)} \end{cases}$$

$$SST \sim \sigma^2 \chi^2_{n-1}$$

$$SSR \sim \sigma^2 \chi^2_{k-1}$$

$$SSE \sim \sigma^2 \chi^2_{n-k}$$

$$F = \frac{\text{DEV(R)}/\text{k}-1}{\text{DEV(E)}/\text{n - k}} = \frac{SSR/k-1}{SSE/n-k} \sim \frac{\dfrac{\chi^2_k}{\sigma^2}\Big/k-1}{\dfrac{\chi^2_{n\ \ k}}{\sigma^2}\Big/n-k} \sim F_{k;n\text{-}k}$$

The null hypothesis will therefore be rejected (accepted) if the sample statistics F is higher (lower) than the Fisher’s distribution quantile corresponding to the significance level imposed by the test ($F_{\alpha, k, n\text{-}k}$)

ANOVA (ANalysis Of Variance)

Based on the decomposition of the total deviance (SST) in deviance of regression (SSR) and of the error (SSE) you can build a statistical test that verifies, through inferential techniques , the overall adjustment of a linear model to original data .

ANOVA	G.L	SS	*MS*	*F*	*P-value F*
Regression	K-1	SSR	MSR=SSR/k-1	MSR/MSE	Probability is lower than 0.05, then we have to reject the null hypothesis
Residual	n-k	SSE	MSE=SSE/n-k		
Total	n-1	SST	MST=SST/n-1		

Multicollinearity

Main measures:

Multiple square correlation coefficient R_j^2 and the Variance Inflation Factors (VIF) obtained by means of auxiliary regressions between each regressor and the other k-2

$$\boxed{VIF = \frac{1}{1 - R_j^2}}$$

High multicollinearity in presence of R^2 values greater than 0,7 where VIF>=3,5

R_j^2 and VIF

R_J^2	VIF
0	1
0,5	2
0,6	2,5
0,7	3,5
0,8	5
0,9	10
0,95	50

Solution methods of multicollinearity

1. Identify the explanatory variable (or the variables) linear combination of the other, and delete it !
2. Increase , if it is possible , the n sample observations
3. Increase , if it is possible , the number of regressors

Testing the assumptions of the model:

- Linearity
  - Linear relationship between Y and each Xi
- Independence among residuals
  - null correlation among residuals
- Normality of residuals
  - normal distribution of residuals
- Homoskedasticity of residuals
  - Finite and constant variance of residuals

LINEARITY

- scatter plot X vs Y
- Scatter plot residuals (studentized) vs predicted values (standardized)
- Correlation coefficient and R^2 between each X and Y

<u>If the relationship is not linear</u>

- Adopt linear transformations (logarithmic) of data