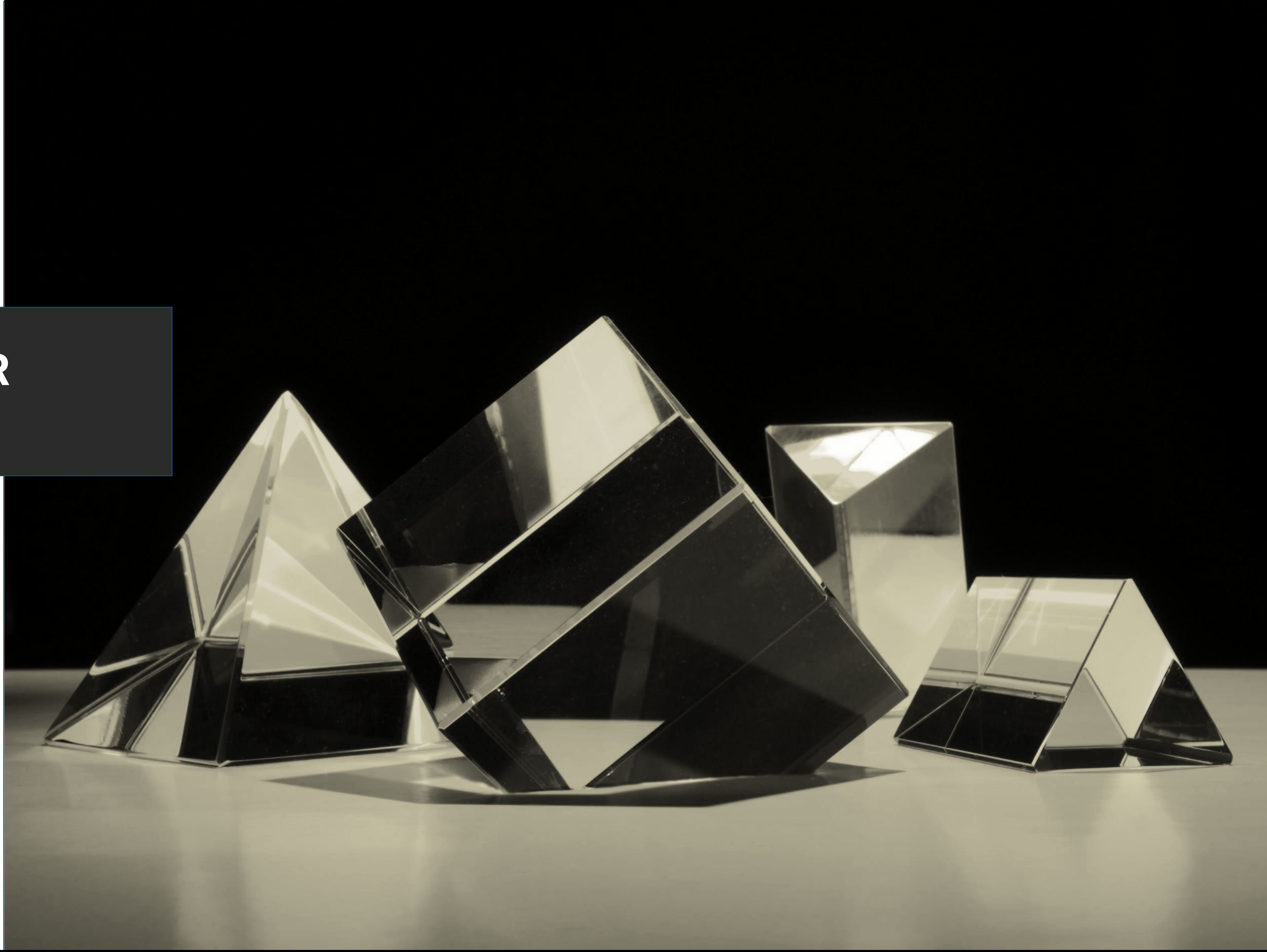
LES FACULTÉS
DE L'UNIVERSITÉ
CATHOLIQUE DE LILLE

Linear Models examples

MULTIPLE LINEAR REGRESSION



Baptiste Mokas

baptiste.mokas@gmail.com

weeki.io/dynamical-système

linktr.ee/baptistemokas

+33 7 69 08 54 19

KNOWLEDGE TREE
AMIRI COURS M1

Part 1: Introduction to Multiple Linear Regression

1.1 The studied model

- the multiple regression model
- Vector of dimension
- Note
- The model (4)
- Example with 2 facto

1.2 Matrix notation

- model in matrix notation
- Note
- Example

1.3 Estimation of B using the OLS method

- The vertical distance between the data points

1.4 Matrix differentiation

- Definition
- Properties
- Exercise

1.5 Application to OLS estimation

- Equation
- Exercise
- Theorem
- Exercise

1.6 Properties of OLS estimators

- Theorem

1.7 Study of residuals (fitting errors)

- Definition of model residuals
- Exercise
- Fact
- Decomposition

1.8 Measure of the goodness of fit of a model

- the coefficient of determination
- The mean squared error

1.9 Comparison of models

- Definition
- Exercise
- Comparison of nested models
- GRETL

1.10 Estimation of σ^2

- Theorem
- Proof
- Consequence of Cochran's theorem

1.11 Confidence intervals

- Equation
- Note
- Construction of a confidence interval

1.12 Confidence interval for a linear combination of the coefficients

- Vector
- Equation

1.13 Prediction interval

- Random variable
- Note

1.14 Simultaneous confidence regions

- Theorem
- Reminder
- Proof
- Cochran's theorem
- Equation
- Note
- Examples
- Note
- Proofs
- Exercise
- Example

1.15 Hypothesis testing: significance of a coefficient Bj

- Test of the true value
- Mathematical translation
- Decision rule
- Returning to the choice of s
- Equation
- Summary
- Note

1.16 Notion of p-value

- Definition
- Note

1.17 Test of a linear constraint on the coefficients

- Equation
- Hypotheses
- Decision
- Example

1.18 Test of significance of multiple coefficients

- Testing the joint nullity
- Test of overall significance of the model
- Comparing the models
- Statistics
- Test of comparison of two nested models
- Note

1.19 Test of multiple linear constraints on the coefficients

- Testing multiple linear restrictions
- Hypotheses
- Test of multiple linear constraints on the coefficients
- Proof

1.20 Analysis of results and validation of model sumptions

- Two related questions

1.21 Analysis of residuals

- Study of residuals
- Curvature in the form of residuals

KNOWLEDGE TREE
AMIRI COURS M1

Part 1: Introduction to Multiple Linear Regression

1.22 Notion of leverage

- Definition + Note
- Properties
- Note

1.23 Standardized residuals

- Definition

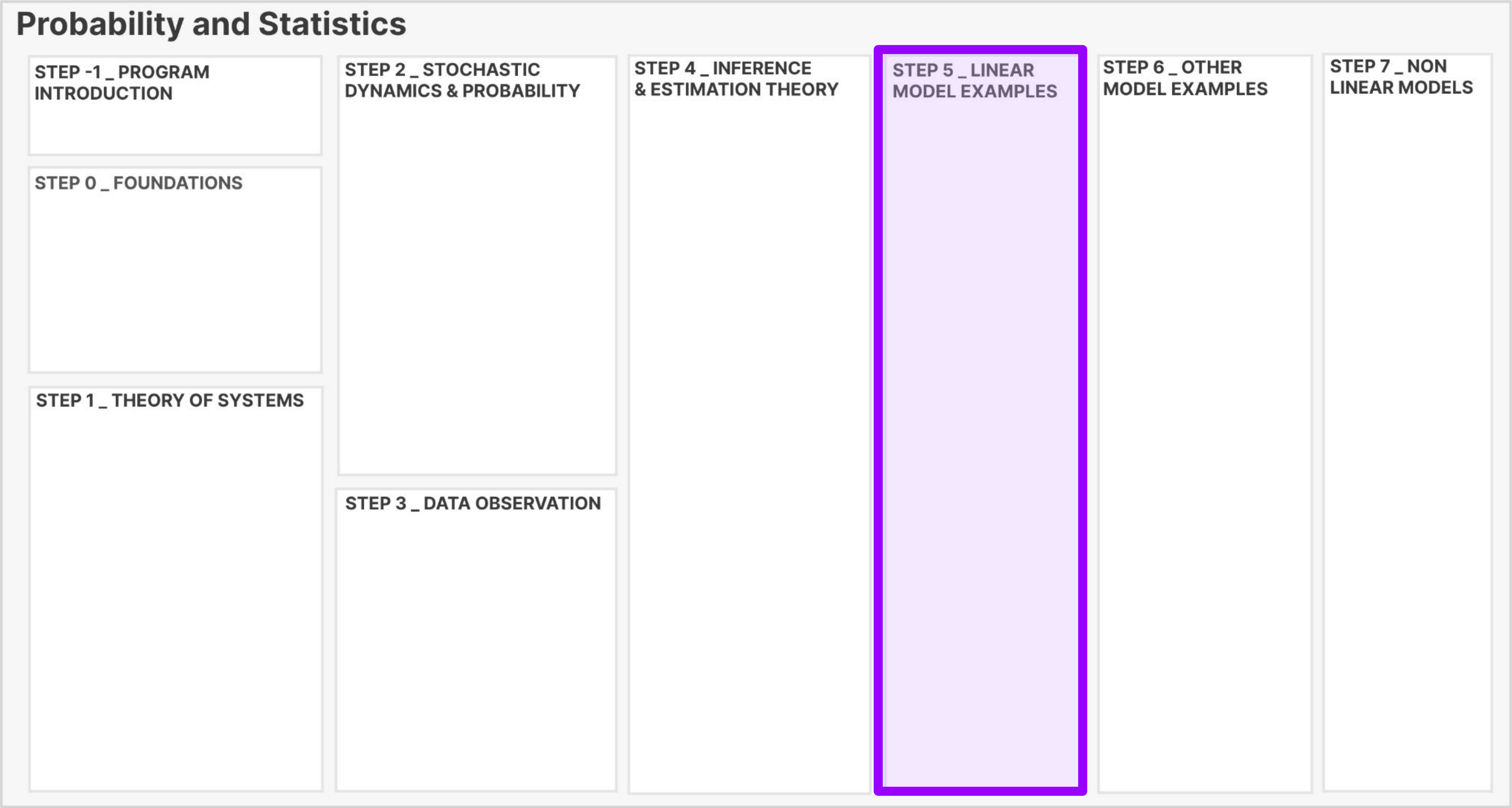
1.24 Studentized residuals

- Definition
- Properties

1.25 Cook's distance

- Definition
- DFFITS

OUR STORY LINE



Probability and Statistics

STEP -1 _ PROGRAM INTRODUCTION

STEP 0 _ FOUNDATIONS

STEP 1 _ THEORY OF SYSTEMS

STEP 2 _ STOCHASTIC DYNAMICS & PROBABILITY

STEP 3 _ DATA OBSERVATION

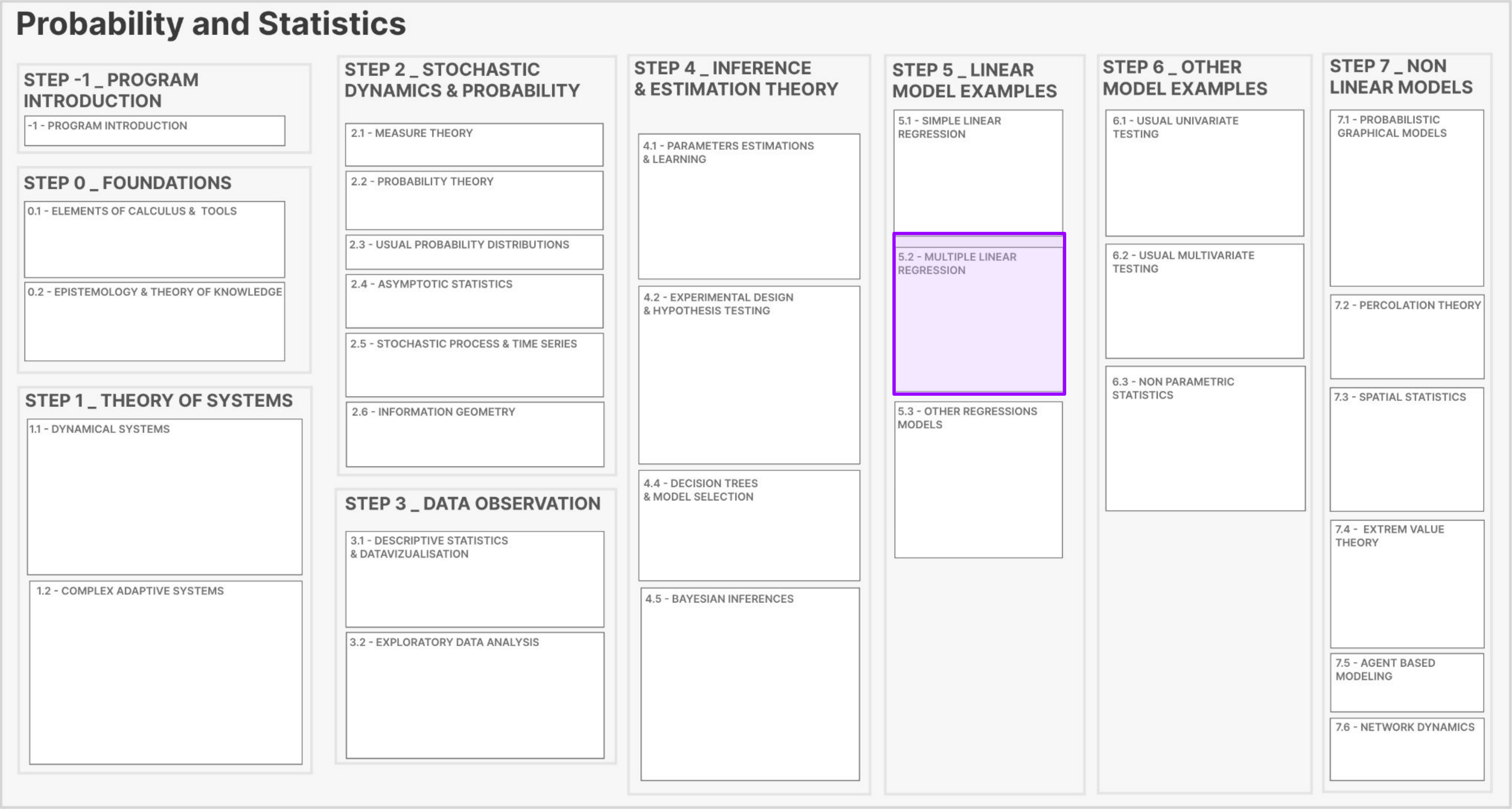
STEP 4 _ INFERENCE & ESTIMATION THEORY

STEP 5 _ LINEAR MODEL EXAMPLES

STEP 6 _ OTHER MODEL EXAMPLES

STEP 7 _ NON LINEAR MODELS

OUR STORY LINE



Probability and Statistics

STEP -1 _ PROGRAM INTRODUCTION

- -1 - PROGRAM INTRODUCTION

STEP 0 _ FOUNDATIONS

- 0.1 - ELEMENTS OF CALCULUS & TOOLS
- 0.2 - EPISTEMOLOGY & THEORY OF KNOWLEDGE

STEP 1 _ THEORY OF SYSTEMS

- 1.1 - DYNAMICAL SYSTEMS
- 1.2 - COMPLEX ADAPTIVE SYSTEMS

STEP 2 _ STOCHASTIC DYNAMICS & PROBABILITY

- 2.1 - MEASURE THEORY
- 2.2 - PROBABILITY THEORY
- 2.3 - USUAL PROBABILITY DISTRIBUTIONS
- 2.4 - ASYMPTOTIC STATISTICS
- 2.5 - STOCHASTIC PROCESS & TIME SERIES
- 2.6 - INFORMATION GEOMETRY

STEP 3 _ DATA OBSERVATION

- 3.1 - DESCRIPTIVE STATISTICS & DATAVIZUALISATION
- 3.2 - EXPLORATORY DATA ANALYSIS

STEP 4 _ INFERENCE & ESTIMATION THEORY

- 4.1 - PARAMETERS ESTIMATIONS & LEARNING
- 4.2 - EXPERIMENTAL DESIGN & HYPOTHESIS TESTING
- 4.4 - DECISION TREES & MODEL SELECTION
- 4.5 - BAYESIAN INFERENCES

STEP 5 _ LINEAR MODEL EXAMPLES

- 5.1 - SIMPLE LINEAR REGRESSION
- 5.2 - MULTIPLE LINEAR REGRESSION
- 5.3 - OTHER REGRESSIONS MODELS

STEP 6 _ OTHER MODEL EXAMPLES

- 6.1 - USUAL UNIVARIATE TESTING
- 6.2 - USUAL MULTIVARIATE TESTING
- 6.3 - NON PARAMETRIC STATISTICS

STEP 7 _ NON LINEAR MODELS

- 7.1 - PROBABILISTIC GRAPHICAL MODELS
- 7.2 - PERCOLATION THEORY
- 7.3 - SPATIAL STATISTICS
- 7.4 - EXTREM VALUE THEORY
- 7.5 - AGENT BASED MODELING
- 7.6 - NETWORK DYNAMICS

KNOWLEDGE TREE

Part 1: Introduction to Multiple Linear Regression

1.1 Overview of Multiple Linear Regression

- Definition and Purpose of Multiple Linear Regression
- Distinction from Simple Linear Regression

1.2 Multivariate Data and Variables

- Identifying Dependent and Independent Variables in Multivariate Data
- Role of Multiple Independent Variables

1.3 The Multiple Linear Regression Model

- Formulating the Multiple Linear Regression Equation
- Interpretation of Coefficients

1.4 Assumptions and Diagnostic Checks

- Assumptions of Multiple Linear Regression
- Diagnostic Techniques for Assumption Violations

Part 2: Estimation and Inference

2.1 Estimation Methods

- Least Squares Estimation in Multiple Linear Regression
- Calculation of Model Coefficients

2.2 Hypothesis Testing

- Testing the Overall Significance of the Model
- Individual Coefficient Tests and Interpretation

2.3 Confidence Intervals

- Constructing Confidence Intervals for Coefficients
- Practical Interpretation of Confidence Intervals

Part 3: Model Evaluation and Selection

3.1 Model Fit Measures

- R-squared, Adjusted R-squared, and Their Interpretations
- Comparing Models for Model Selection

3.2 Residual Analysis

- Diagnosing Residual Patterns and Model Fit
- Residual Plots and Their Interpretations

3.3 Multicollinearity

- Understanding Multicollinearity and Its Effects
- Detection and Remedies

Part 4: Interaction Effects and Nonlinearity

4.1 Interaction Terms

- Incorporating Interaction Effects in Multiple Linear Regression
- Interpretation of Interaction Coefficients

4.2 Polynomial Regression

- Introduction to Polynomial Terms
- Handling Nonlinear Relationships with Polynomial Regression

Part 5: Model Validation and Assumptions

5.1 Cross-Validation Techniques

- K-Fold Cross-Validation for Model Validation
- Preventing Overfitting in Multiple Regression

5.2 Outlier Detection and Handling

- Identifying and Addressing Outliers in Multiple Regression
- Robust Regression Methods

The multiple regression model

⇒ The multiple regression model is a generalization with multiple factors (p) of the simple model. It is written as follows:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, i = 1 \ldots, n.$$

The terminology remains the same, and we assume that $n > p + 1$.
- Y: Dependent variable
- $x_1, \ldots, x_p$: Independent variables
- $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1}$: Coefficient vector or parameter vector.

Vector of dimension

Part 1: Introduction to Multiple Linear Regression

1.1 The studied model

- Each individual is described by p variables, forming a p-dimensional vector (a $p \times 1$ matrix), called an individual vector or observation vector.

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{ip} \end{pmatrix} \in \mathbb{R}^p.$$

The ε_i's are i.i.d; for all $i = 1, \ldots, n$ and $\mathbb{E}(\varepsilon_i) = 0$.
for all $i = 1, \ldots, n$, $\mathbb{V}(\varepsilon_i) = \sigma^2$; $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.
for all $i = 1, \ldots, n$, $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$.

Note

Part 1: Introduction to Multiple Linear Regression

1.1 The studied model

Note

Linearity and Gaussianity of the model are assumptions that need to be validated.

To verify them, one can either use the prior knowledge of the model or perform a statistical test.

The model (4)

⇒ The model (4) can be written in the form:

$$Y_i = \left[1, x_i^\top\right] \beta + \varepsilon_i, i = 1 \ldots, n$$

Example

(1) We want to estimate the price of an apartment based on its location, size, luxury level, location, and age.

(2) We have data on the age, mileage in thousands of kilometers, and price in thousands of euros for a sample of used cars of the same type.

Age	5	4	6	5	5	5	6	6	2	7	7
Km	92	64	124	97	79	76	93	63	13	111	143
Price	7.8	9.5	6.4	7.5	8.1	9	6.1	8.7	15.4	6.4	4.4

Example with 2 factor

Part 1: Introduction to Multiple Linear Regression

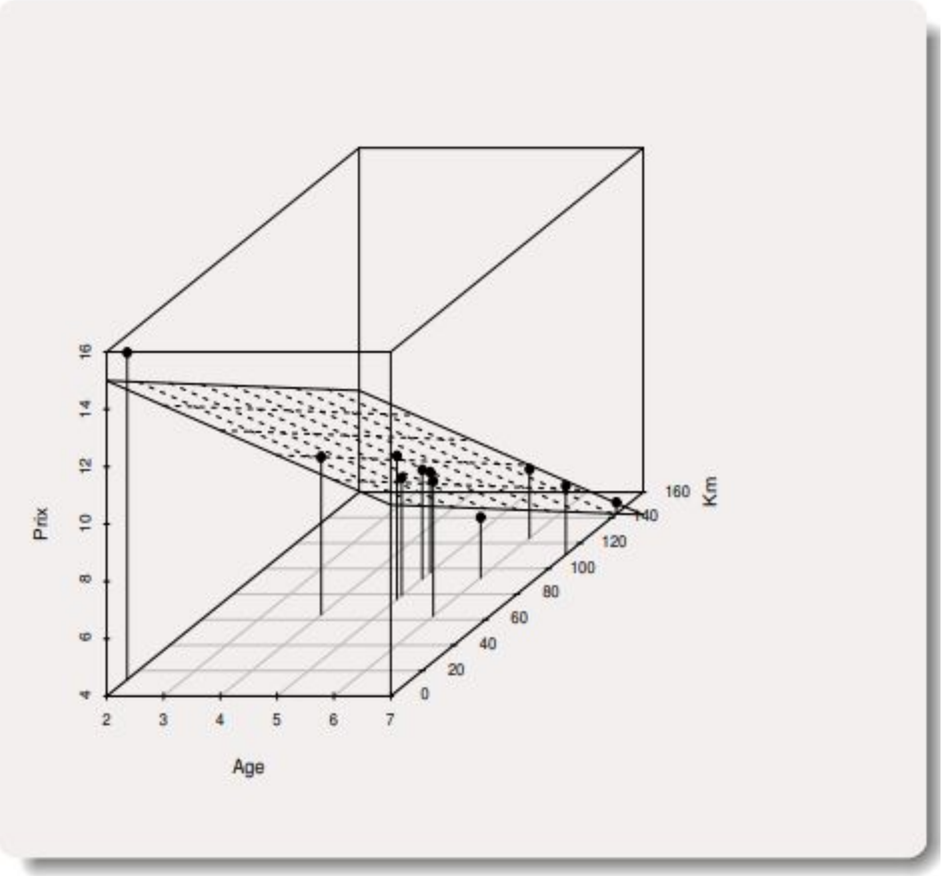
1.1 The studied model

- We have an example with 2 factors to which we associate the model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

⇒ Alternatively:

$$Y_i = [1 \quad x_{i1} \quad x_{i2}] \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon_i$$



Age

- The visual verification indeed appears close to a plane.

⇒ However, interpretation becomes challenging when dealing with 3 or more factors.

The vertical distance between the data points

Part 1: Introduction to Multiple Linear Regression

1.3 Estimation of B using the OLS method

- The vertical distance between the point cloud in $\mathbb{R}^{p+1}$ and the regression hyperplane is given by:

$$F(\beta) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = \|Y - X\beta\|^2.$$

Goal:
To find $\hat{\beta}$ such that $F(\hat{\beta})$ is the minimum of $F(\beta)$, i.e.,

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \|Y - X\beta\|^2.$$

Definition of model residuals

Part 1: Introduction to Multiple Linear Regression

1.7 Study of residuals (fitting errors)

⇒ The residuals of the model are defined as:

$$\hat{e}_i = Y_i - \widehat{Y}_i = Y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij}, i = 1, \ldots, n,$$

which can be denoted as:

$$\hat{e} = \begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \vdots \\ \hat{e}_i \\ \vdots \\ \hat{e}_n \end{pmatrix} = Y - X\hat{\beta} = Y - \widehat{Y}$$

Exercise

Part 1: Introduction to Multiple Linear Regression

1.7 Study of residuals (fitting errors)

Exercise. We consider the matrices:

$$\mathbf{1}_n = \underbrace{(11\dots1)}_{n \text{ times}}; \quad H_X = X\left(X^\top X\right)^{-1} X^\top \text{ and } J_n = \frac{1}{n}\underbrace{[\mathbf{1}_n\mathbf{1}_n\dots\mathbf{1}_n]}_{n \text{ times}}.$$

(1) Show that H_X is a projector. What is the rank of H_X?
(2) Show that

$$H_X X = X; \quad H_X Y = \widehat{Y} \text{ and } \hat{e} = (I_n - H_X)Y = (I_n - H_X)\varepsilon.$$

Interpret these results.
(3) Deduce that $H_X J_n = J_n$.
(4) Verify that $J_n Y = \bar{Y}\mathbf{1}_n$. Show that

$$(Y - \widehat{Y})^\top \left(\widehat{Y} - \bar{Y}\mathbf{1}_n\right) = Y^\top (I_n - H_X)(H_X - J_n)Y.$$

(5) Conclude that:

$$(Y - \widehat{Y})^\top \left(\widehat{Y} - \bar{Y}\mathbf{1}_n\right) = 0$$

Fact

Part 1: Introduction to Multiple Linear Regression

1.7 Study of residuals (fitting errors)

Using the fact that

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2=\left\|Y-\bar{Y}\mathbf{1}_n\right\|^2,$$

we have:

$$\begin{aligned}\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2 =&\left(Y-\bar{Y}\mathbf{1}_n\right)^\top\left(Y-\bar{Y}\mathbf{1}_n\right)\\ =&(Y-\widehat{Y})^\top(Y-\widehat{Y})+2(Y-\widehat{Y})^\top\left(\widehat{Y}-\bar{Y}\mathbf{1}_n\right)\\ &+\left(\widehat{Y}-\bar{Y}\mathbf{1}_n\right)^\top\left(\widehat{Y}-\bar{Y}\mathbf{1}_n\right)\\ =&\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2+\sum_{i=1}^{n}\left(\widehat{Y}_i-\bar{Y}\right)^2.\end{aligned}$$

Decomposition

Part 1: Introduction to Multiple Linear Regression

1.7 Study of residuals (fitting errors)

⇒ We thus obtain the decomposition seen in the context of simple regression:

$$SC_{\text{tot}} = SC_{\text{res}} + SC\text{reg.}$$

The coefficient of determination

Part 1: Introduction to Multiple Linear Regression

1.8 Measure of the goodness of fit of a model

⇒ The coefficient of determination is defined as:

$$R^2 = \frac{SC_{\text{reg}}}{SC_{\text{tot}}} = 1 - \frac{SC_{\text{res}}}{SC_{\text{tot}}}$$

⇒ The adjusted R^2 is a version of the coefficient of determination that takes into account the number of observations and the number of explanatory variables:

$$R^2_{\text{adj}} = 1 - \frac{SC_{\text{res}}/(n-p-1)}{SC_{\text{tot}}/(n-1)} = 1 - \frac{n-1}{n-p-1}\left(1 - R^2\right)$$

The mean squared error

Part 1: Introduction to Multiple Linear Regression

1.8 Measure of the goodness of fit of a model

⇒ The Mean Squared Error (MSE) is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2 = \frac{SS_{\text{res}}}{n}.$$

- The Mean Absolute Error (MAE) is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{e}_i|.$$

⇒ The Mean Absolute Percentage Error (MAPE) is defined as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{e}_i}{Y_i}\right|.$$

Definition

Part 1: Introduction to Multiple Linear Regression

1.9 Comparison of models

Definition

Two models are said to be nested if one of them can be obtained as a special case of the other.

They explain the same variable Y, and the explanatory variables of the "small" model are all included in the "large" model.

Exercise

Exercise. Show that for two nested models:

$$R_G^2 \geq R_P^2$$

An increase in the number of explanatory variables mechanically increases the coefficient of determination.

Hint: SS_{res} decreases when adding an explanatory variable to the model.

Comparison of nested models

⇒ The comparison of two nested models can be done using the R^2_{adj}.

A model is preferable to another if its R^2_{adj} is higher than that of the other model.
⇒ There are other criteria of the same type that can be used to compare the goodness of fit of two models related to the same dependent variable:

- The Akaike Information Criterion (AIC) is defined as:

$$AIC = n \ln\left(\frac{SS_{\text{res}}}{n}\right) + 2(p+1).$$

- The Schwarz Bayesian Information Criterion (BIC) is defined as:

$$BIC = n \ln\left(\frac{SS_{\text{res}}}{n}\right) + (p+1)\ln n.$$

Comparison of nested models

Part 1: Introduction to Multiple Linear Regression

1.9 Comparison of models

$\Rightarrow$ Unlike R^2_{adj}, the AIC and BIC criteria are increasing functions of SS_{res}.
- A model is therefore preferable to another in terms of AIC (resp. BIC) if its AIC (resp. BIC) criterion is lower than that of the other model.
$\Rightarrow$ Some software packages may use a slightly different formulation of the AIC and BIC criteria, but the interpretations remain the same.
- For example, in GRETL:

$$AIC = -2\ln\mathcal{L} + 2(p+1) \text{ and } BIC = -2\ln\mathcal{L} + 2(p+1)\ln n,$$

where $\mathcal{L}$ is the maximized likelihood of the model provided by the software.

- GRETL also provides the Hannan-Quinn criterion, which has a formulation almost identical to the BIC:

$$HQC = n \ln\left(\frac{SS_{\text{res}}}{n}\right) + (p+1) \ln \ln n.$$

⇒ In GRETL:

$$HQC = -2 \ln \mathcal{L} + 2(p+1) \ln \ln n.$$

- With the software package R, the previous criteria can be obtained using the AIC function, which depends on a parameter k.
⇒ This parameter takes the values $2, 2 \ln n$ and $2 \ln \ln n$ for the AIC, BIC, and HQC criteria, respectively.

Theorem

Part 1: Introduction to Multiple Linear Regression

1.10 Estimation of σ²

Theorem
Consider the estimator

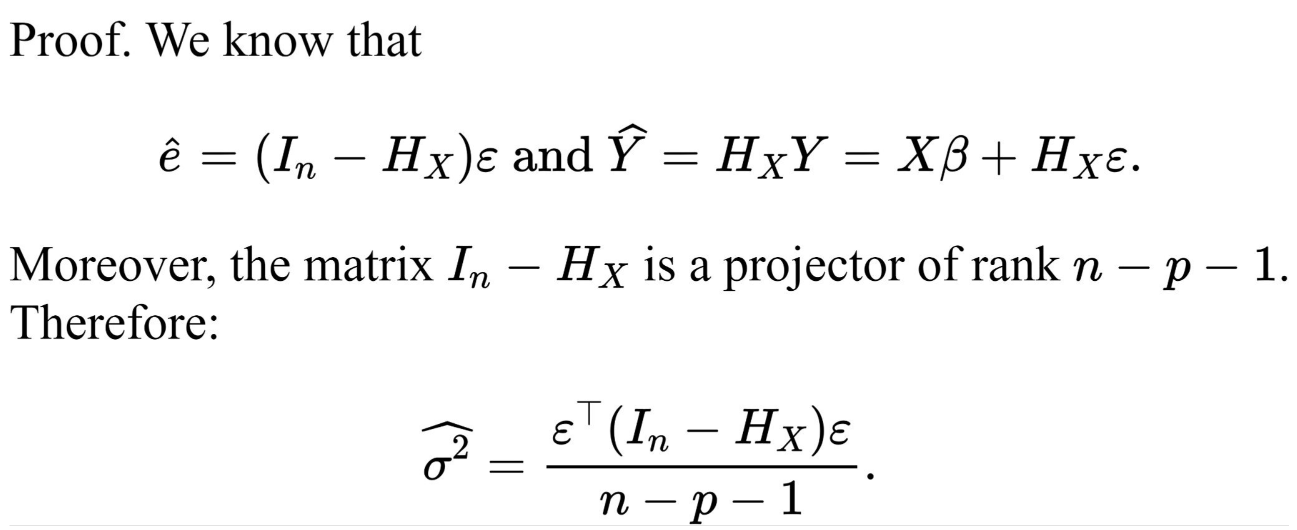
$$\widehat{\sigma^2} = \frac{1}{n-p-1}\sum_{i=1}^{n}\hat{e}_i^2 = \frac{SS_{\text{res}}}{n-p-1}$$

(i) $\widehat{\sigma^2}$ is an unbiased estimator of σ^2 independent of $\widehat{Y}$.
(ii) Moreover, if $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$, then:
- $\frac{(n-p-1)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-p-1}$.
- For any $j = 0, 1, \ldots, p$, we have $\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 \left(X^\top X\right)^{-1}_{j+1,j+1}}} \sim \mathcal{T}_{n-p-1}$, where $\left(X^\top X\right)^{-1}_{j+1,j+1}$ is the $(j+1)^{\text{th}}$ diagonal element of $\left(X^\top X\right)^{-1}$.

Proof

Part 1: Introduction to Multiple Linear Regression

1.10 Estimation of σ²



Proof. We know that

$$\hat{e} = (I_n - H_X)\varepsilon \text{ and } \widehat{Y} = H_X Y = X\beta + H_X\varepsilon.$$

Moreover, the matrix $I_n - H_X$ is a projector of rank $n - p - 1$. Therefore:

$$\widehat{\sigma^2} = \frac{\varepsilon^\top (I_n - H_X)\varepsilon}{n - p - 1}.$$

Consequence of Cochran's theorem

The first part of (ii) is a consequence of Cochran's theorem [†], while the second part follows from the fact that:

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 \left(X^\top X\right)^{-1}\right)$$

and the independence of $\widehat{\sigma^2}$ and $\hat{\beta}$ (since $\widehat{\sigma^2} \perp \widehat{Y} = X\hat{\beta}$).
The result in (i) can be easily obtained from (ii) using the linearity of expectation.

[†]

Theorem
If $Z \sim \mathcal{N}(0, I_n)$ and A is a projection matrix of rank q, then the variables AZ and $(I_n - A)Z$ are independent. Moreover:

$$\|AZ\|^2 = Z^\top AZ \sim \chi_q^2 \text{ and } \|(I_n - A)Z\|^2 = Z^\top (I_n - A)Z \sim \chi_{n-q}^2.$$

Equation

Part 1: Introduction to Multiple Linear Regression

1.11 Confidence intervals

$\Rightarrow$ If the $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$, then the $1-\alpha$ confidence intervals for β_j are given by:

$$CI(\beta_j) = \left[\hat{\beta}_j - t_{n-p-1,1-\alpha/2}\widehat{\sigma_{\hat{\beta}_j}}, \hat{\beta}_j + t_{n-p-1,1-\alpha/2}\widehat{\sigma_{\hat{\beta}_j}}\right],$$

where

$$\widehat{\sigma_{\beta_j}} = \sqrt{\widehat{\sigma^2}\left(X^\top X\right)^{-1}_{j+1,j+1}} \text{ is the standard error of } \hat{\beta}_j.$$

Note

Part 1: Introduction to Multiple Linear Regression

1.11 Confidence intervals

Note
If the ε_i are no longer assumed to be Gaussian,
then the previous confidence interval is an asymptotic confidence interval,
replacing $t_{n-p-1,1-\alpha/2}$ with the $1-\alpha/2$ quantile $z_{1-\alpha/2}$ of the standard normal distribution.

Construction of a confidence interval

Part 1: Introduction to Multiple Linear Regression

1.11 Confidence intervals

We can also easily construct a $1-\alpha$ confidence interval for σ^2:

$$CI(\sigma^2) = \left[\frac{n-p-1}{c_{n-p-1,1-\alpha/2}}\widehat{\sigma^2}, \frac{n-p-1}{c_{n-p-1,\alpha/2}}\widehat{\sigma^2}\right].$$

Vector

$\Rightarrow$ We are given a vector $a = (a_0, \ldots, a_p)^\top$ and we aim to estimate $a^\top \beta$.
$\Rightarrow$ Example: If we take $a^\top = \left(1, x_1^*, \ldots, x_p^*\right)$
where $x_j^*, j = 1, \ldots, p$ represents a new observation of the explanatory variables,

then $a^\top \beta = \mathbb{E}(Y^*)$.
$\Rightarrow$ We have that $a^\top \hat{\beta}$ is an unbiased estimator of $a^\top \beta$,

and furthermore:

$$\frac{a^\top \hat{\beta} - a^\top \beta}{\sqrt{\widehat{\sigma^2} a^\top \left(X^\top X\right)^{-1} a}} \sim \mathcal{T}_{n-p-1},$$

which implies:

$$\frac{a^\top \hat{\beta} - a^\top \beta}{\sqrt{\widehat{\sigma^2} a^\top \left(X^\top X\right)^{-1} a}} \sim \mathcal{T}_{n-p-1}.$$

Equation

Part 1: Introduction to Multiple Linear Regression

1.12 Confidence interval for a linear combination of the coefficients

⇒ Thus, a confidence interval at level $1-\alpha$ for $a^\top\beta$ is given by:

$$CI\left(a^\top\beta\right) = \left[a^\top\hat{\beta} \pm t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma^2}a^\top\left(X^\top X\right)^{-1}a}\right].$$

⇒ In the case where $a^\top = \left(1, x_1^*, \ldots, x_p^*\right)$ where the x_j^* represent a new observation of the explanatory variables, we have:

$$CI(\mathbb{E}(Y^*)) = \left[\widehat{Y^*} \pm t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma^2}a^\top\left(X^\top X\right)^{-1}a}\right].$$

Random variable

Part 1: Introduction to Multiple Linear Regression

1.13 Prediction interval

⇒ Now we are looking for an interval for Y^*, which is a random variable. We have:

$$Y^* = a^\top \beta + \varepsilon^* \text{with } a^\top = \left(1, x_1^*, \ldots, x_p^*\right) \text{ and } \varepsilon^* \sim \mathcal{N}\left(0, \sigma^2\right),$$
$$\varepsilon^* \text{ independent of } \varepsilon_i, i = 1, \ldots, n$$
$$\text{Thus}$$
$$Y^* \sim \mathcal{N}\left(a^\top \beta, \sigma^2\right) \text{ and } \widehat{Y^*} = a^\top \hat{\beta} \sim \mathcal{N}\left(a^\top \beta, \sigma^2 a^\top \left(X^\top . X\right)^{-1} a\right),$$

⇒ Therefore,

$$Y^* \sim \mathcal{N}\left(a^\top \beta, \sigma^2\right) \text{ and } \widehat{Y^*} = a^\top \hat{\beta} \sim \mathcal{N}\left(a^\top \beta, \sigma^2 a^\top \left(X^\top X\right)^{-1} a\right),$$

which implies

$$\widehat{e^*} = Y^* - \widehat{Y^*} \sim \mathcal{N}\left(0, \sigma^2 \left[1 + a^\top \left(X^\top X\right)^{-1} a\right]\right).$$

Note

Part 1: Introduction to Multiple Linear Regression

1.13 Prediction interval

Hence,

$$\frac{Y^* - \widehat{Y^*}}{\sqrt{\widehat{\sigma^2}\left[1 + a^\top \left(X^\top X\right)^{-1} \ldots\ldots\ldots a\right]}} \sim \mathcal{T}_{n-p-1}.$$

$\Rightarrow$ The prediction interval is given by:

$$PI(Y^*) = \left[\widehat{Y^*} \pm \ldots\ldots\ldots t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma^2}\left[1 + a^\top \left(X^\top X\right)^{-1} a\right]}\right]$$

Note:
The quantity

$$a^\top \left(X^\top X\right)^{-1} a$$

represents the leverage of the observation $x^* = \left(x_1^*, \ldots, x_p^*\right)$.

Test of the true value

Part 1: Introduction to Multiple Linear Regression

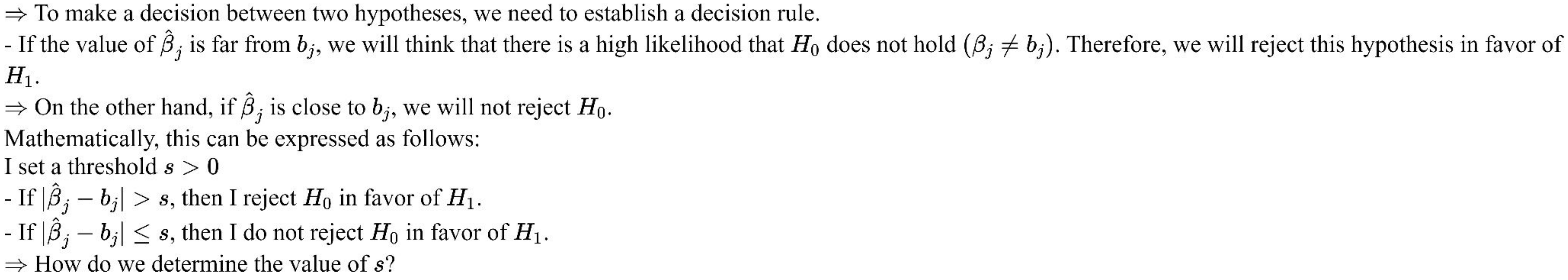
1.15 Hypothesis testing: significance of a coefficient Bj

⇒ For $j = 0, \ldots, p$, we want to perform a test regarding the true value of β_j.
⇒ The hypotheses are:
$H_0 : \beta_j = b_j$ against $H_1 : \beta_j \neq b_j$, where b_j is a fixed value.
- In particular, for $b_j = 0$, we test the effect of the explanatory variable x_j on the variable Y.
⇒ Performing this test involves determining whether or not to exclude the variable x_j from the model (referred to as an exclusion test).
⇒ If H_0 is rejected in favor of H_1, we say that the coefficient β_j is significant.

Mathematical translation

Part 1: Introduction to Multiple Linear Regression

1.15 Hypothesis testing: significance of a coefficient Bj



⇒ To make a decision between two hypotheses, we need to establish a decision rule.

- If the value of $\hat{\beta}_j$ is far from b_j, we will think that there is a high likelihood that H_0 does not hold $(\beta_j \neq b_j)$. Therefore, we will reject this hypothesis in favor of H_1.

⇒ On the other hand, if $\hat{\beta}_j$ is close to b_j, we will not reject H_0.

Mathematically, this can be expressed as follows:

I set a threshold $s > 0$

- If $|\hat{\beta}_j - b_j| > s$, then I reject H_0 in favor of H_1.
- If $|\hat{\beta}_j - b_j| \leq s$, then I do not reject H_0 in favor of H_1.

⇒ How do we determine the value of s?

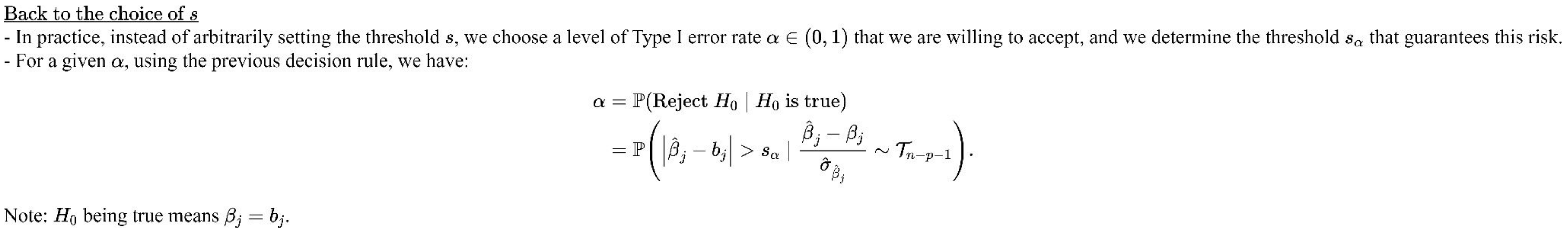
Decision rule

⇒ With the decision rule mentioned above, two types of errors can occur:
(3) Rejecting H_0 in favor of H_1 when H_0 is true (incorrectly concluding that $\beta_j \neq b_j$): Type I error.
(2) Not rejecting H_0 in favor of H_1 when H_0 is false (incorrectly thinking that $\beta_j = b_j$): Type II error.
⇒ These two types of errors are associated with their respective probabilities:
- $\alpha = \mathbb{P}$ (Rejecting H_0 when H_0 is true): Type I error rate.
- $\beta = \mathbb{P}$ (Not rejecting H_0 when H_0 is false): Type II error rate.
⇒ The quantity $1 - \beta$ represents the probability of correctly rejecting H_0. In other words, it is the probability of rejecting the null hypothesis H_0 when it is false.
⇒ Therefore, $1 - \beta$ measures the power of the test.

Returning to the choice of s

Part 1: Introduction to Multiple Linear Regression

1.15 Hypothesis testing: significance of a coefficient Bj



Back to the choice of s

- In practice, instead of arbitrarily setting the threshold s, we choose a level of Type I error rate $\alpha \in (0,1)$ that we are willing to accept, and we determine the threshold s_α that guarantees this risk.
- For a given α, using the previous decision rule, we have:

$$\begin{aligned}\alpha &= \mathbb{P}(\text{Reject } H_0 \mid H_0 \text{ is true}) \\ &= \mathbb{P}\left(\left|\hat{\beta}_j - b_j\right| > s_\alpha \mid \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathcal{T}_{n-p-1}\right).\end{aligned}$$

Note: H_0 being true means $\beta_j = b_j$.

Equation

Part 1: Introduction to Multiple Linear Regression

1.15 Hypothesis testing: significance of a coefficient Bj

$\Rightarrow$ Therefore:

$$1-\alpha=\mathbb{P}\left(-\frac{s_\alpha}{\hat{\sigma}_{\hat{\beta}_j}}\leq\mathcal{T}_{n-p-1}\leq\ldots\ldots\ldots\frac{s_\alpha}{\hat{\sigma}_{\hat{\beta}_j}}\right).$$

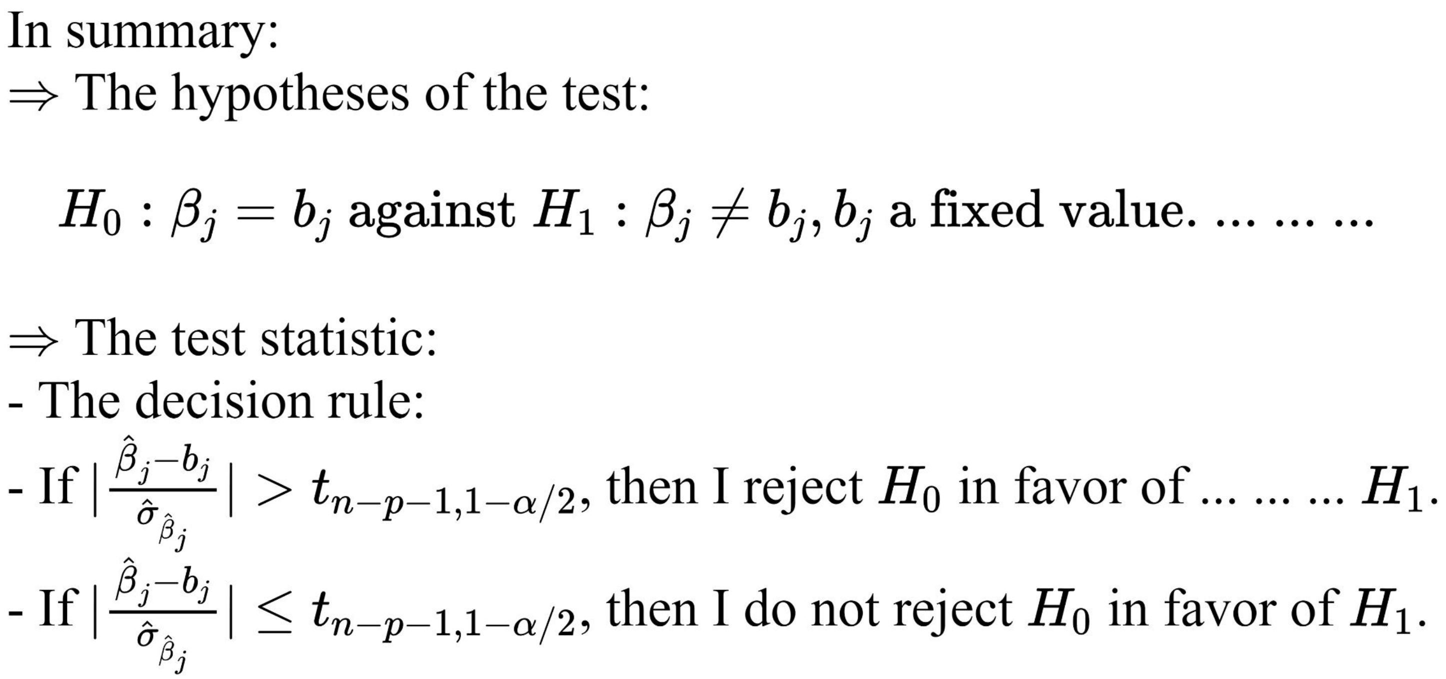
$\Rightarrow$ Consequently:

$$\frac{s_\alpha}{\hat{\sigma}_{\hat{\beta}_j}}=t_{n-p-1,1-\alpha/2}\Longrightarrow s_\alpha=\hat{\sigma}_{\hat{\beta}_j}\cdot t_{n-p-1,1-\alpha/2}.$$

Summary

Part 1: Introduction to Multiple Linear Regression

1.15 Hypothesis testing: significance of a coefficient Bj



In summary:

$\Rightarrow$ The hypotheses of the test:

$$H_0 : \beta_j = b_j \text{ against } H_1 : \beta_j \neq b_j, b_j \text{ a fixed value. ... ...}$$

$\Rightarrow$ The test statistic:

- The decision rule:

- If $|\frac{\hat{\beta}_j - b_j}{\hat{\sigma}_{\hat{\beta}_j}}| > t_{n-p-1,1-\alpha/2}$, then I reject H_0 in favor of ... ... H_1.

- If $|\frac{\hat{\beta}_j - b_j}{\hat{\sigma}_{\hat{\beta}_j}}| \leq t_{n-p-1,1-\alpha/2}$, then I do not reject H_0 in favor of H_1.

Note

Part 1: Introduction to Multiple Linear Regression

1.15 Hypothesis testing: significance of a coefficient Bj

$\Rightarrow$ The specific case $\beta_j = 0$ corresponds to the test statistic:

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}.$$

- This quantity is called the t-statistic and is automatically calculated by most statistical software.
- Note that the decision rule mentioned above can be extended to the cases of one-sided tests as follows:
- $H_0 : \beta_j = b_j$ against $H_1 : \beta_j > b_j$: we reject H_0 when $\frac{\hat{\beta}_j - b_j}{\tilde{\sigma}_{\beta_j}} > t_{n-p-1,1-\alpha}$ and do not reject H_0 otherwise.
- $H_0 : \beta_j = b_j$ against $H_1 : \beta_j < b_j$: we reject H_0 when $\frac{\hat{\beta}_j - b_j}{\hat{\sigma}_{\beta_j}} < t_{n-p-1,1-\alpha}$ and do not reject H_0 otherwise.

Definition

Part 1: Introduction to Multiple Linear Regression

1.16 Notion of p-value

⇒ Many software programs allow conducting a test without having to consult statistical tables.
- They provide a probability associated with the t-statistic, known as the critical probability or significance level (or p-value).
- The p-value corresponds to the level of Type I error risk α^* for which we would be undecided between the two decisions: rejecting or not rejecting H_0.
- Therefore, we can conclude a test using the p-value with the following decision rule:
- If $\alpha^* < \alpha$, then I reject H_0 in favor of H_1.
- If $\alpha^* \geq \alpha$, then I do not reject H_0 in favor of H_1.

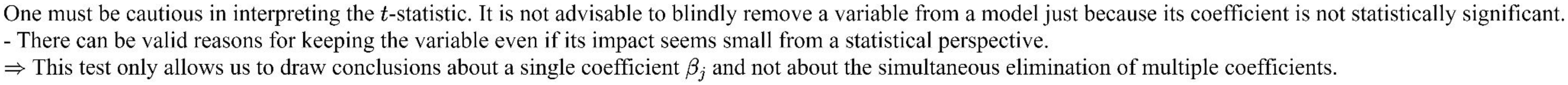
⇒ In addition to the p-value, software provides more detailed conclusions about the significance level of β_j with a precision of 0.1, 0.01, 0.001, etc.
- This level of precision is indicated by stars next to the p-value.
- The p-value helps us have more or less confidence in the hypothesis H_0.

⇒ The larger the p-value, the more we believe that H_0 is true. The smaller the p-value, the more tempted we are to reject H_0.

Note

Part 1: Introduction to Multiple Linear Regression

1.16 Notion of p-value



One must be cautious in interpreting the t-statistic. It is not advisable to blindly remove a variable from a model just because its coefficient is not statistically significant.
- There can be valid reasons for keeping the variable even if its impact seems small from a statistical perspective.
$\Rightarrow$ This test only allows us to draw conclusions about a single coefficient β_j and not about the simultaneous elimination of multiple coefficients.

Equation

Part 1: Introduction to Multiple Linear Regression

1.17 Test of a linear constraint on the coefficients

$\Rightarrow$ We want to test a linear restriction of the form:

$$\begin{cases} H_0 : a_0\beta_0 + a_1\beta_1 + \cdots + a_p\beta_p = b \\ \text{against} \\ H_1 : a_0\beta_0 + a_1\beta_1 + \cdots + a_p\beta_p \neq b \end{cases}$$

Examples:

$$(1) \begin{cases} H_0 : \beta_1 + \beta_2 = 1 \\ \text{against} \\ H_1 : \beta_1 + \beta_2 \neq 1 \end{cases}$$

$$(2) \begin{cases} H_0 : \beta_2 - \beta_3 = 0 \\ \text{against} \\ H_1 : \beta_2 - \beta_3 \neq 0 \end{cases}$$

Hypotheses

Part 1: Introduction to Multiple Linear Regression

1.17 Test of a linear constraint on the coefficients

The test hypotheses can be written in matrix form as:

$$(1)\begin{cases} H_0 : a^\top \beta = b \\ \text{against} \\ H_1 : a^\top \beta \neq b \end{cases} \quad \text{where} \quad a^\top = (a_0, \ldots, a_p).$$

$\Rightarrow$ ... ... Reminder:

$$\frac{a^\top \hat{\beta} - a^\top \beta}{\sqrt{\widehat{\sigma^2} a^\top \left(X^\top X\right)^{-1} a}} \sim \mathcal{T}_{n-p-1}$$

Decision

Part 1: Introduction to Multiple Linear Regression

1.17 Test of a linear constraint on the coefficients

$\Rightarrow$ Thus, we can establish the following decision rule:
- If

$$\left|\frac{a^\top \hat{\beta} - b}{\sqrt{\widehat{\sigma^2} a^\top \left(X^\top X\right)^{-1} a}}\right| > t_{n-p-1,1-\alpha/2},$$

then I reject H_0 in favor of H_1.
- if

$$\left|\frac{a^\top \hat{\beta} - b}{\sqrt{\widehat{\sigma^2} a^\top \left(X^\top X\right)^{-1} a}}\right| \leq t_{n-p-1,1-\alpha/2},$$

then I do not reject H_0 in favor of H_1.

Example

Part 1: Introduction to Multiple Linear Regression

1.17 Test of a linear constraint on the coefficients

Example.

Consider the model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

with:
$n = 40, \hat{\beta}_0 = 1.37, \hat{\beta}_1 = 0.632, \hat{\beta}_2 = 0.452$

Perform the following tests:

$$(1) \begin{cases} H_0 : \beta_1 = \beta_2 \\ \text{against} \\ H_1 : \beta_1 \neq \beta_2 \end{cases},$$

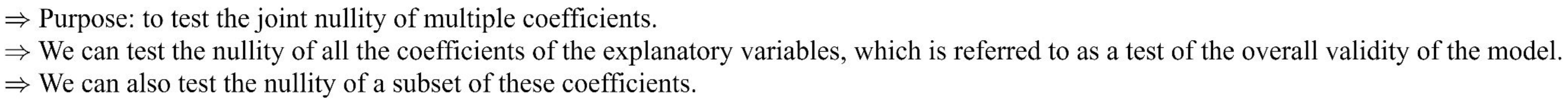
$$(2) \begin{cases} H_0 : \beta_1 + \beta_2 = 1 \\ \text{against} \\ H_1 : \beta_1 + \beta_2 \neq 1 \end{cases},$$

$$(3) \begin{cases} H_0 : \beta_2 = 0 \\ \text{against} \\ H_1 : \beta_2 \neq 0 \end{cases}.$$

Testing the joint nullity

Part 1: Introduction to Multiple Linear Regression

1.18 Test of significance of multiple coefficients



⇒ Purpose: to test the joint nullity of multiple coefficients.

⇒ We can test the nullity of all the coefficients of the explanatory variables, which is referred to as a test of the overall validity of the model.

⇒ We can also test the nullity of a subset of these coefficients.

Test of overall significance of the model

Test of overall significance of the model
⇒ The hypotheses of the test:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ \text{against} \\ H_1 : \exists j \in \{1, \dots, p\} \text{ such that } \beta_j \neq 0 \end{cases}$$

⇒ The test statistic is given by:

$$F = \frac{S......C_{reg}/p}{SC_{res}/n-p-1} = \frac{R^2(n-p-1)}{p\big(1-R^2\big)} \sim ...\mathcal{F}_{p,n-p-1}......underH_0.$$

⇒ The decision rule is:
- If $F > f_{p,n-p-1,1-\alpha}$, then I reject H_0 in favor of H_1.
- If $F \leq f_{p,n-p-1,1-\alpha}$, then I do not reject H_0 in favor of H_1.

Comparing the models

Part 1: Introduction to Multiple Linear Regression

1.18 Test of significance of multiple coefficients

$\Rightarrow$ If the null hypothesis H_0 is not rejected, it means that the model without explanatory variables is preferable to the original model.
$\Rightarrow$ The variables x_j explain very little of the variations in Y, so it is better to eliminate them.
$\Rightarrow$ When we reject H_0 in favor of H_1, we say that the model is globally significant.

Part 1: Introduction to Multiple Linear Regression

1.18 Test of significance of multiple coefficients

- As with the t statistic, statistical software provides the value of the F statistic, as well as the associated p-value.
- The calculation of F is sometimes presented in a table called ANOVA, which shows several terms:

Source	Df	Sum Sq	Mean Sq	F value
Regression	p	$S.........C_{\text{reg}}$	$MC_{\text{reg}} = \frac{SC_{\text{reg}}}{p}$	$F = \frac{MC_{\text{reg}}}{MC_{\text{res}}}$
Residuals	... ... $n-p-1$	SC_{res}	$MC_{\text{res}}\ \frac{SC_{res}}{n-p-1}$	

Test of comparison of two nested models

Part 1: Introduction to Multiple Linear Regression

1.18 Test of significance of multiple coefficients

Comparison test of two nested models
The hypotheses of the test:
For $q \leq p$, we have:
$H_0 : \beta_{j_1} = \beta_{j_2} = \beta_{jq} = 0$

against

$H_1 : \exists j \in \{j_1, \ldots, j_q\}$ such that $\beta_j \neq 0$
$\Rightarrow$ The test statistic:

$$\begin{aligned} F &= \frac{[SC_{\text{res}}(P) - SC_{\text{res}}(G)]/q}{SC_{\text{res}}(G)/n-p-1} = \frac{[SC_{\text{reg}}(G) - SC_{\text{reg}}(P)]/q}{SC_{\text{res}}(G)/n-p-1} \\ &= \frac{\left(R_G^2 - R_P^2\right)(n-p-1)}{q\left(1-R_G^2\right)} \sim \mathcal{F}_{q,n-p-1}, \text{ under } H_0.^{a} \end{aligned}$$

$\Rightarrow$ The decision rule:
- If $F > f_{q,n-p-1,1-\alpha}$, then I reject H_0 in favor of H_1.
- If $F \leq f_{q,n-p-1,1-\alpha}$, then I do not reject H_0.

a. G = large model and P = small model

Note

Part 1: Introduction to Multiple Linear Regression

1.18 Test of significance of multiple coefficients

We have seen that in the case $q = p$, the small model is

$$Y_i = \beta_0 + \varepsilon_i.$$

In this case,

$$\widehat{Y}_i = \hat{\beta}_0 = \bar{Y}$$

Therefore,

$$SC_{\text{reg}}\left(P\right) = 0 \text{ and } SC_{\text{res}}\left(P\right) = 1$$

Thus, we obtain the result of the overall significance test of the model.

Testing multiple linear restrictions

Part 1: Introduction to Multiple Linear Regression

1.19 Test of multiple linear constraints on the coefficients

⇒ We want to test multiple linear restrictions of the form:

$$H_0 : \begin{cases} \ell_{10}\beta_0 + \ell_{11}\beta_1 + \cdots + \ell_{1p}\beta_p = b_1 \\ \ell_{20}\beta_0 + \ell_{21}\beta_1 + \cdots + \ell_{2p}\beta_p = b_2 \\ \vdots \\ \ell_{q0}\beta_0 + \ell_{q1}\beta_1 + \cdots + \ell_{qp}\beta_p = b_q \end{cases}$$

against $\exists j \in \{1, \ldots, q\}$ such that the jth equation is not satisfied, with $q \leq p + 1$.
⇒ Examples:
(1) $H_0 : \begin{cases} \beta_0 + \beta_1 + \beta_2 + \beta_5 = 1 \\ \beta_1 + \beta_2 = 0 \end{cases}; (2) H_0 : \begin{cases} \beta_2 - \beta_3 = 0 \\ \beta_0 - \beta_1 = 1 \end{cases}.$

Hypotheses

Part 1: Introduction to Multiple Linear Regression

1.19 Test of multiple linear constraints on the coefficients

- The test hypotheses can be written in matrix form as:

$$(1) \begin{cases} H_0 : L\beta = b \\ \text{against} \\ H_1 : L\beta \neq b \end{cases}$$

where $b^\top = (b_1, \dots, b_q)$ and L is a $q \times (p+1)$ matrix.
$\Rightarrow$ Reminder:

$$F = \frac{[L(\hat{\beta} - \beta)]^\top \left[L\left(X^\top X\right)^{-1} L^\top \right]^{-1} L(\hat{\beta} - \beta)}{q\widehat{\sigma^2}} \sim \mathcal{F}_{q,n-p-1}$$

Test of multiple linear constraints on the coefficients

Part 1: Introduction to Multiple Linear Regression

1.19 Test of multiple linear constraints on the coefficients

The test hypotheses:

$$\begin{cases} H_0 : L\beta = b \\ \text{ against} \\ H_1 : L\beta \neq b \end{cases}$$

where $b^\top = (b_1, \ldots, b_q)$ and L is a $q \times (p+1)$ matrix.
$\Rightarrow$ The test statistic:

$$F = \frac{(L\hat{\beta} - b)^\top \left[L \left(X^\top X \right)^{-1} L^\top \right]^{-1} (L\hat{\beta} - b)}{q\widehat{\sigma^2}} \sim \mathcal{F}_{q,n-p-1}.$$

- The decision rule:
- if $F > f_{q,n-p-1,1-\alpha}$, then reject H_0 in favor of H_1.
- if $F \leq f_{q,n-p-1,1-\alpha}$, then do not reject H_0.

Proof

Part 1: Introduction to Multiple Linear Regression

1.19 Test of multiple linear constraints on the coefficients

$\Rightarrow$ If H_0 is true, then $L\hat{\beta}$ should be close to b. Therefore, we will reject H_0 if $L\hat{\beta}$ is far from b, i.e. for a threshold s

$$||L\hat{\beta} - b||^2 > s.$$

$\Rightarrow$ Now, we recall that if H_0 is true, then:

$$F = \frac{(L\hat{\beta} - b)^\top \left[L\left(X^\top X\right)^{-1} L^\top\right]^{-1} (L\hat{\beta} - b)}{q\widehat{\sigma^2}} \sim \mathcal{F}_{q,n-p-1}.$$

Hence, the decision rule obtained using a similar reasoning as that used to establish the simultaneous confidence region.

Two related questions

Part 1: Introduction to Multiple Linear Regression

1.20 Analysis of results and validation of model sumptions

⇒ We can distinguish two types of questions:
- Those related to the influence of specific observations
- Those related to the influence of the chosen factors on the estimates.

Study of residuals

Part 1: Introduction to Multiple Linear Regression

1.21 Analysis of residuals

⇒ The study of residuals is fundamental. It allows us to identify:

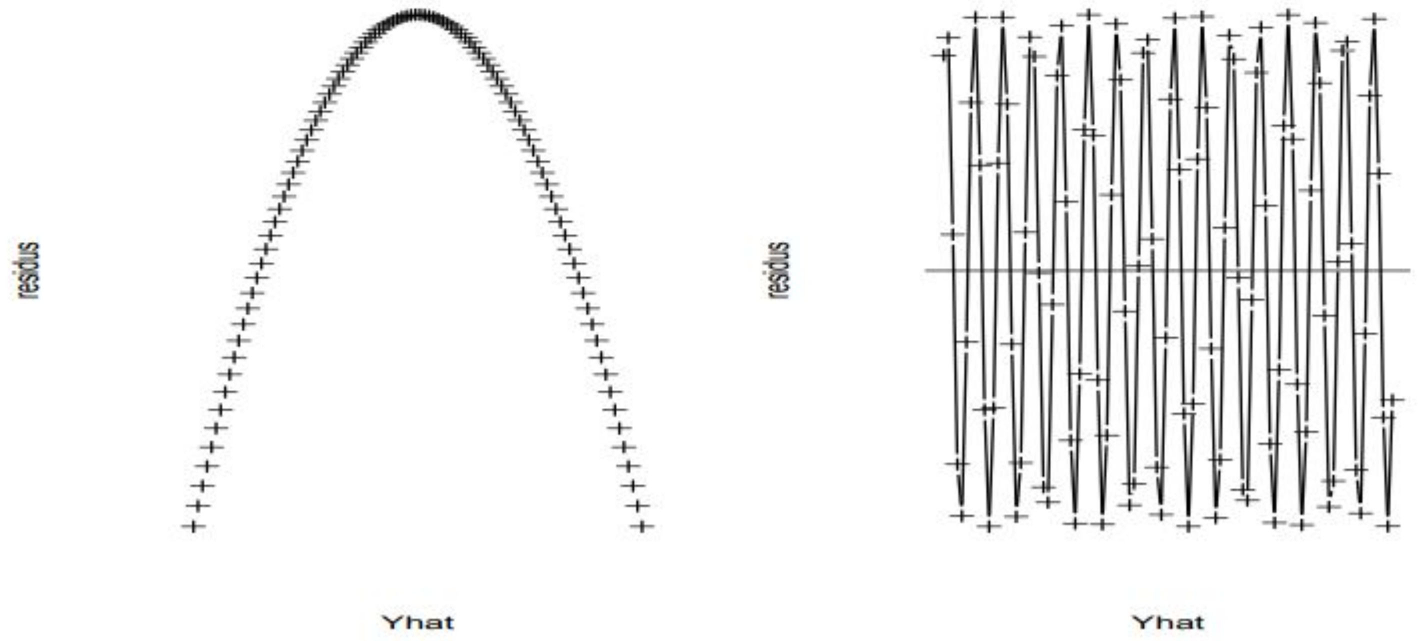
- Potentially outlier observations.
- Observations that play an important role in the regression estimation.
- Furthermore, the study of residuals is often the only way to empirically verify the validity of the model assumptions.
- The representation of residuals is done based on i (or based on $\widehat{Y}_i$).
- This representation should not exhibit any particular trend.

Curvature in the form of residuals

Part 1: Introduction to Multiple Linear Regression

1.21 Analysis of residuals

⇒ Curvature in the shape of the residuals based on $\widehat{Y}_i$ suggests that the linearity assumption is not appropriate.
- Monotonic behavior of the variability of the residuals based on $\widehat{Y}_i$ can indicate non-constant variance.



Definition + Note

Part 1: Introduction to Multiple Linear Regression

1.22 Notion of leverage

Definition:
The hat matrix associated with X is called the matrix H_X defined by

$$H_X = X\left(X^\top X\right)^{-1} X^\top$$

Note:
The i-th element on the diagonal of H_X is nothing but the leverage of observation number i, i.e:

$$h_{ii} = a_i^\top \left(X^\top X\right)^{-1} a_i$$

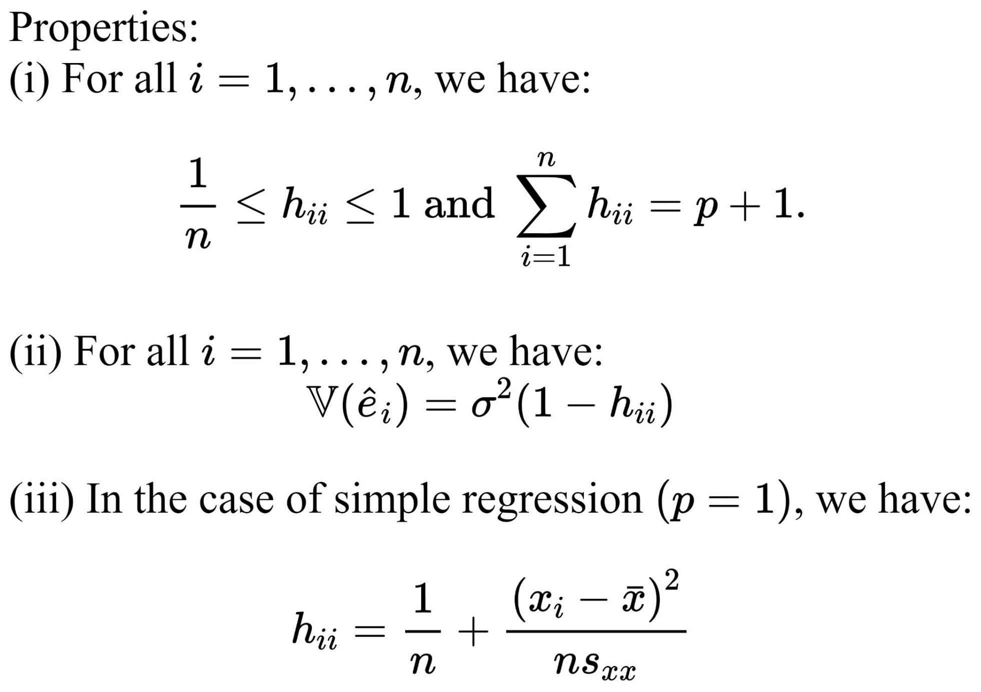
where $a_i^\top$ is the i-th row of the matrix X, given by:

$$a_i^\top = [1 \quad x_{i1} \quad x_{i2} \ldots x_{ip}]$$

Properties

Part 1: Introduction to Multiple Linear Regression

1.22 Notion of leverage



Properties:
(i) For all $i = 1, \ldots, n$, we have:

$$\frac{1}{n} \leq h_{ii} \leq 1 \text{ and } \sum_{i=1}^{n} h_{ii} = p + 1.$$

(ii) For all $i = 1, \ldots, n$, we have:

$$\mathbb{V}(\hat{e}_i) = \sigma^2 (1 - h_{ii})$$

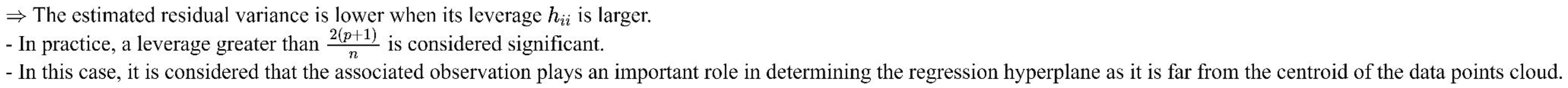
(iii) In the case of simple regression $(p = 1)$, we have:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{n s_{xx}}$$

Note

Part 1: Introduction to Multiple Linear Regression

1.22 Notion of leverage



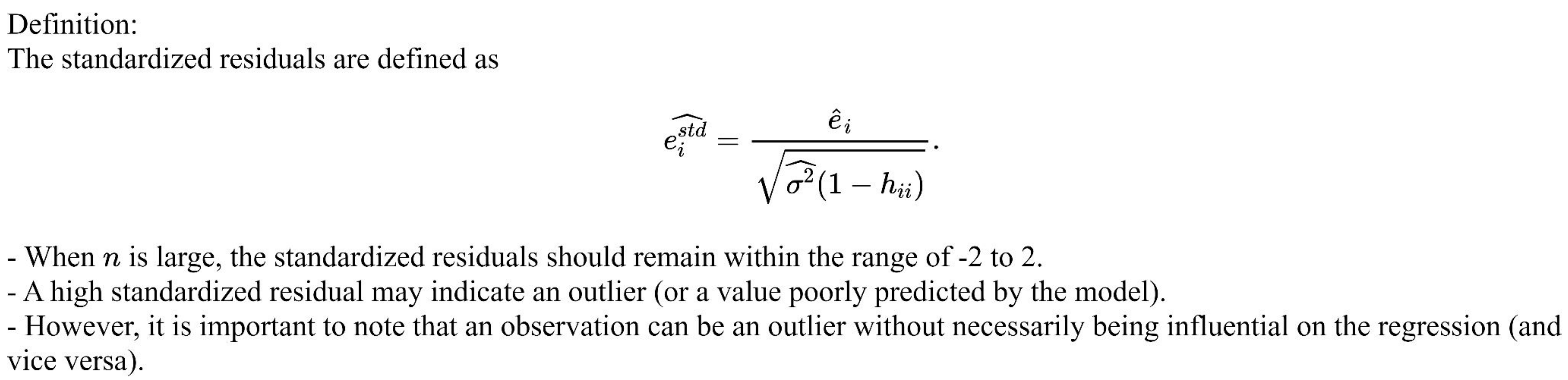
⇒ The estimated residual variance is lower when its leverage h_{ii} is larger.

- In practice, a leverage greater than $\frac{2(p+1)}{n}$ is considered significant.
- In this case, it is considered that the associated observation plays an important role in determining the regression hyperplane as it is far from the centroid of the data points cloud.

Definition

Part 1: Introduction to Multiple Linear Regression

1.23 Standardized residuals



Definition:
The standardized residuals are defined as

$$\widehat{e_i^{std}} = \frac{\hat{e}_i}{\sqrt{\widehat{\sigma^2}(1-h_{ii})}}.$$

- When n is large, the standardized residuals should remain within the range of -2 to 2.
- A high standardized residual may indicate an outlier (or a value poorly predicted by the model).
- However, it is important to note that an observation can be an outlier without necessarily being influential on the regression (and vice versa).

Definition

Definition
The studentized residuals are defined as

$$\widehat{t}_i^* = \frac{\hat{e}_i}{\sqrt{\widehat{\sigma^2_{(-i)}}(1-h_{ii})}},$$

where $\widehat{\sigma^2_{(-i)}}$ is the estimator of σ^2 obtained without using the ith observation, namely

$$\widehat{\sigma^2_{(-i)}} = \frac{SC_{res,(-i)}}{n-p-2} = \frac{1}{n-p-2}\sum_{\substack{k=1\\k\neq i}}^{n-1} \hat{e}_k^2 = \frac{\widehat{\sigma^2}(n-p-1) - \hat{e}_i^2}{n-p-2}.$$

When n is small, we use the studentized residuals.

Properties

Properties

We have the following relationship:

⇒ An observation is considered an outlier with a confidence level of $1-\alpha$ ($\alpha = 0.05$ typically), when

$$\left|\widehat{t_i^*}\right| > t_{n-p-2,1-\alpha/2}.$$

Definition

Part 1: Introduction to Multiple Linear Regression

1.25 Cook's distance

⇒ The Cook's distance allows us to assess the importance of standardized residuals in relation to leverage.

Definition
The Cook's distance associated with observation number i is defined as:

$$C_i = \frac{\widehat{e_i^{std}}^2 h_{ii}}{(p+1)(1-h_{ii})}.$$

- A Cook's distance greater than 1 generally indicates an abnormal influence of the observation Y_i on the regression performed.

DFFITS

Part 1: Introduction to Multiple Linear Regression

1.25 Cook's distance

$\Rightarrow$ We can also examine the DFFITS defined as

$$D_i = \widehat{t_i^*}\sqrt{\frac{h_{ii}}{1-h_{ii}}}.$$

- A high value of D_i indicates that the parameter estimation is strongly affected if we remove observation number i from the sample.
- In practice, an observation is considered influential when

$$|D_i| > \frac{2(p+1)}{n}$$

KEYWORDS (NEW)

- Projecteur
- Modèle linéaire gaussien
- Régression multiple modèle multiplicatif
- Régression linéaire multiple
- Relation fonctionnelle
- Rang matriciel
- Hyperplan
- Normalité des erreurs
- Régression multiple modèle additif
- Modèle linéaire hiérarchique
- Régression multiple
- Notion de levier
- Normalité des données
- Droite des moindres carrés ordinaires (MCO)
- Modèle linéaire généralisé (GLM)
- Notion de linéarité
- Homoscédasticité
- Matrice orthogonale
- Matrice symétrique
- Décomposition de variance et test de linéarité
- Intervalle de confiance asymptotique
- Modèles emboités
- Erreur résiduelle
- Linéarité de l'espérance
- Matrice définie-positive
- Distance de Cook
- Indépendance des observations
- Degré de liberté
- Distribution normale des résidus
- Indépendance des erreurs
- Distribution des résidus

Use Case with python

Use Case: Predictive Modeling in Real Estate

In the context of the course on Multiple Linear Regression, which covers topics related to understanding multiple linear regression, estimation and inference, model evaluation and selection, interaction effects, and model validation, let's explore a use case related to predictive modeling in the real estate industry.

Description:

In this use case, we will apply multiple linear regression techniques to predict real estate property prices based on various independent variables. Accurate price prediction is crucial in real estate for both buyers and sellers to make informed decisions. Multiple linear regression allows us to model the relationship between property attributes and its price.

Key Components:

Introduction to Multiple Linear Regression: Understanding the fundamentals of multiple linear regression, including the purpose, formulation of the regression equation, and interpretation of coefficients.

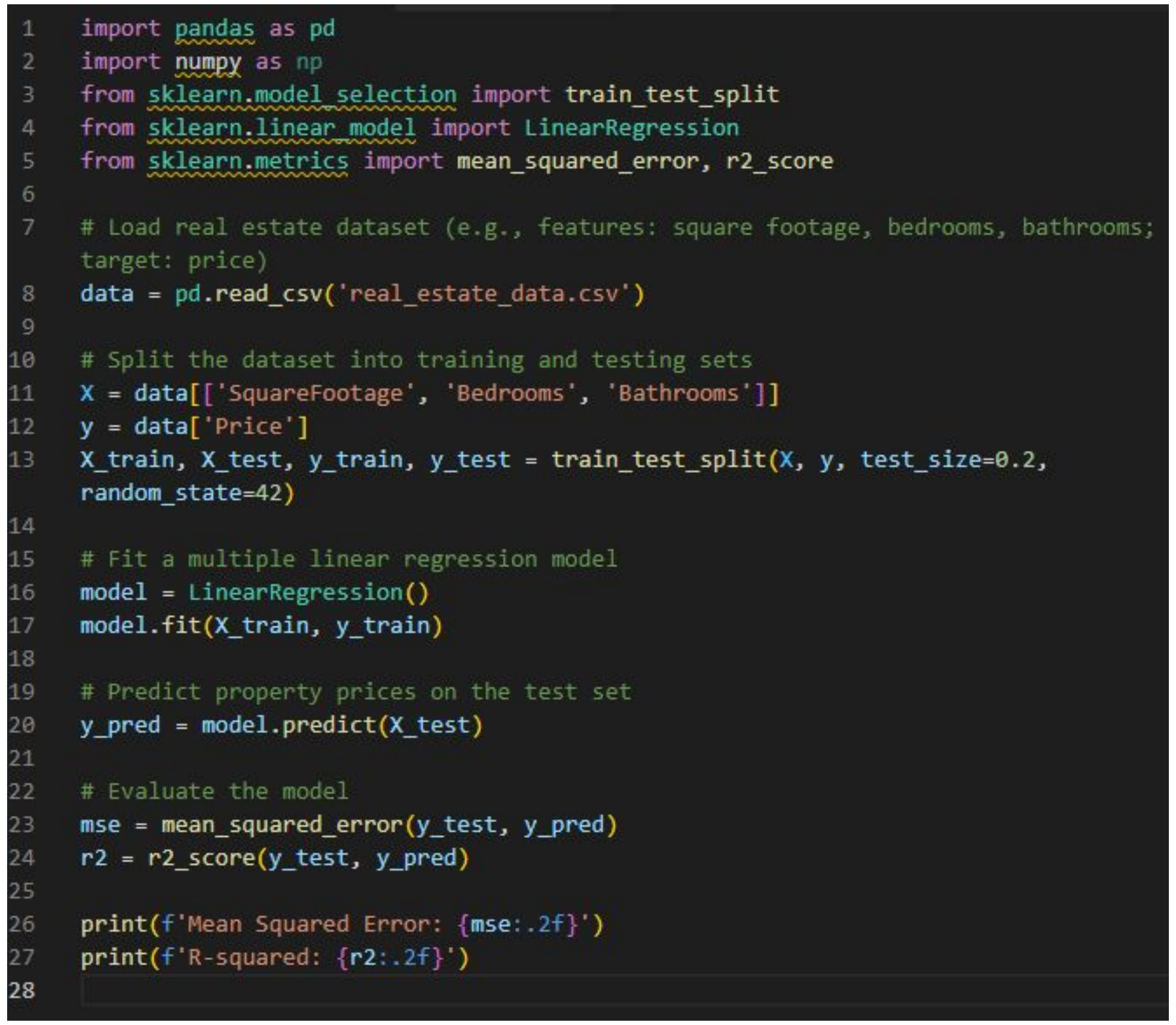
Estimation and Inference: Applying estimation methods such as least squares to calculate model coefficients. Conducting hypothesis tests to determine the significance of independent variables and constructing confidence intervals.

Model Evaluation and Selection: Assessing model fit using measures like R-squared and adjusted R-squared. Analyzing residuals to diagnose model fit and detect any patterns. Understanding multicollinearity and its effects.

Interaction Effects and Nonlinearity: Incorporating interaction terms to capture complex relationships between independent variables. Using polynomial regression to handle nonlinear relationships.

Model Validation and Assumptions: Employing cross-validation techniques to validate the model's performance and prevent overfitting. Identifying and handling outliers using robust regression methods.

Python Code Example (Real Estate Price Prediction):



```python
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Load real estate dataset (e.g., features: square footage, bedrooms, bathrooms; target: price)
data = pd.read_csv('real_estate_data.csv')

# Split the dataset into training and testing sets
X = data[['SquareFootage', 'Bedrooms', 'Bathrooms']]
y = data['Price']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Fit a multiple linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Predict property prices on the test set
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f'Mean Squared Error: {mse:.2f}')
print(f'R-squared: {r2:.2f}')

```

In this code, we load a real estate dataset containing property features and prices. We split the data into training and testing sets, fit a multiple linear regression model, and use it to predict property prices. We then evaluate the model's performance using mean squared error (MSE) and R-squared.

This use case demonstrates how multiple linear regression, as covered in the course, can be applied to real-world scenarios like predicting real estate property prices, providing valuable insights for buyers and sellers in the real estate market.

REFERENCES V2

TOP 20 REFERENCES

- Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied Linear Statistical Models. McGraw-Hill/Irwin.
- Draper, N. R., & Smith, H. (1998). Applied Regression Analysis. Wiley.
- Faraway, J. J. (2016). Linear Models with R. CRC Press.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. Wiley.
- Fox, J. (2015). Applied Regression Analysis and Generalized Linear Models. Sage Publications.
- Weisberg, S. (2014). Applied Linear Regression. Wiley.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied Linear Statistical Models. McGraw-Hill/Irwin.
- Cook, R. D., & Weisberg, S. (1999). Applied Regression Including Computing and Graphics. Wiley.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Routledge.
- Harrell, F. E. (2015). Regression Modeling Strategies. Springer.
- Chatterjee, S., & Hadi, A. S. (2015). Regression Analysis by Example. Wiley.
- Rencher, A. C., & Schaalje, G. B. (2008). Linear Models in Statistics. Wiley.
- Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). Semiparametric Regression. Cambridge University Press.
- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.
- Sheather, S. (2009). A Modern Approach to Regression with R. Springer.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.
- Gelman, A., & Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.
- Freedman, D. A. (2009). Statistical Models: Theory and Practice. Cambridge University Press.
- Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). Data Analysis: A Model Comparison Approach to Regression, ANOVA, and Beyond. Routledge.

TEXTE DE DESCRIPTION DU COURS

Explore the comprehensive course on Multiple Linear Regression, covering key concepts and techniques for understanding and applying regression models in data analysis. You'll start by gaining an overview of multiple linear regression, including its definition, purpose, and distinctions from simple linear regression. The course introduces you to multivariate data and variables, teaching you to identify dependent and independent variables in complex datasets.

You'll delve into the core of the multiple linear regression model, learning to formulate the regression equation and interpret coefficients. The course will guide you through the assumptions and diagnostic checks essential for reliable regression analysis.

Next, you'll explore estimation methods, focusing on least squares estimation in multiple linear regression. Hypothesis testing techniques are covered, allowing you to assess the overall significance of the model and the significance of individual coefficients. The course also equips you with the skills to construct and interpret confidence intervals for model coefficients.

In the realm of model evaluation and selection, you'll dive into measures of model fit, including R-squared and Adjusted R-squared, and understand how to compare different models for effective model selection. You'll also learn to perform residual analysis, diagnose residual patterns, and address model fit issues. Additionally, you'll explore the concept of multicollinearity and its effects on regression models, along with methods for detection and mitigation.

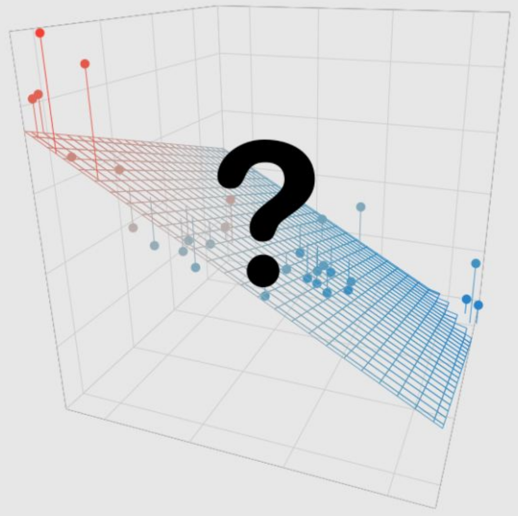
The course delves into interaction effects and nonlinearity, showing you how to incorporate interaction terms and polynomial regression to handle complex relationships in your data.

Finally, you'll explore model validation and assumptions, including cross-validation techniques like K-Fold Cross-Validation and methods for outlier detection and handling in multiple regression.

By the end of this course, you'll possess a robust understanding of multiple linear regression, enabling you to build, evaluate, and refine regression models, make data-driven decisions, and address various challenges in regression analysis.

DESIGN DE LA CARD

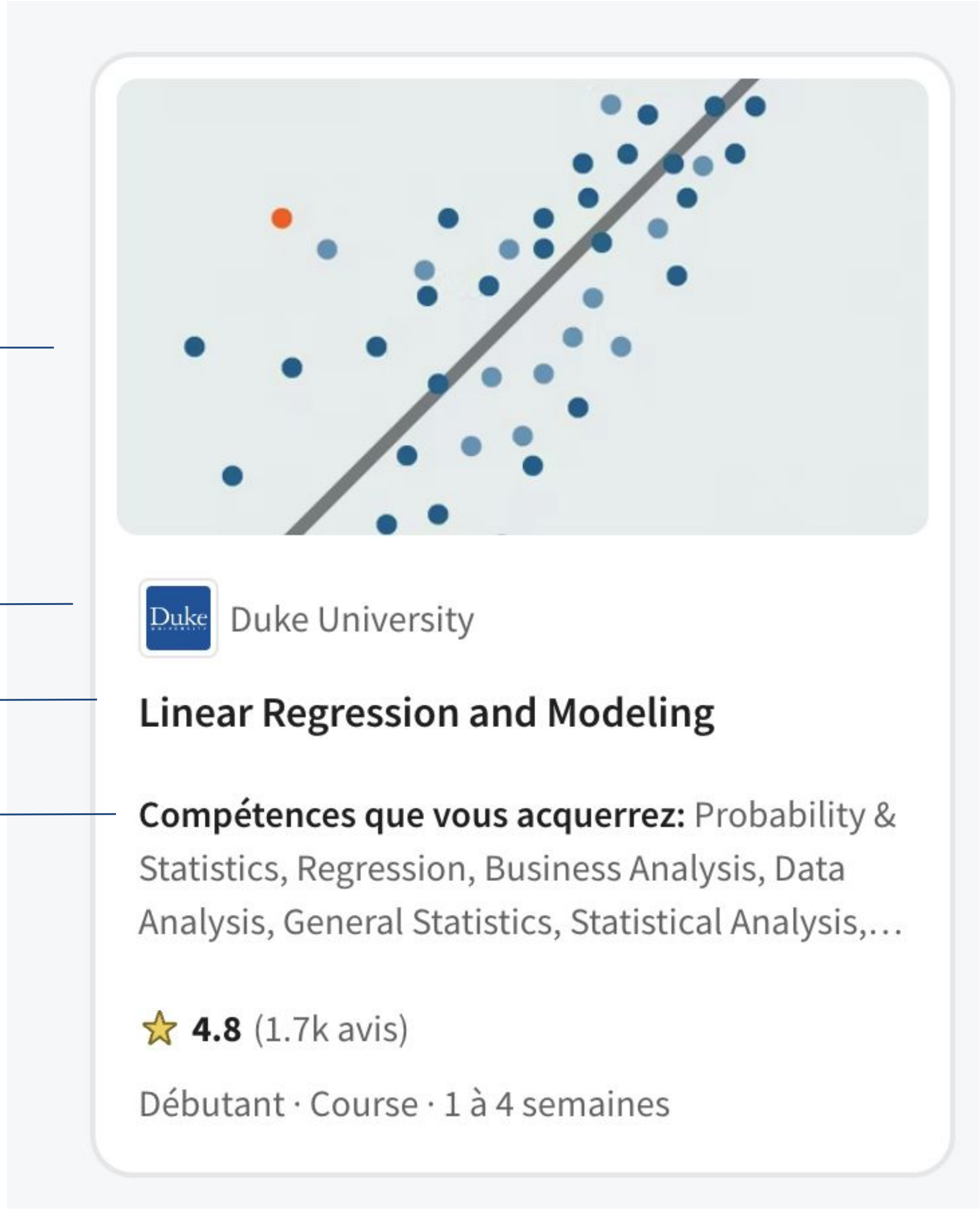
Multiple Linear Regression



Author: Baptiste Mokas, Weeki

Course Name: Simple Linear Regression

#MultipleLinearRegression
#RegressionAnalysis
#ModelEvaluation



Duke University

Linear Regression and Modeling

Compétences que vous acquerrez: Probability & Statistics, Regression, Business Analysis, Data Analysis, General Statistics, Statistical Analysis,...

☆ **4.8** (1.7k avis)

Débutant · Course · 1 à 4 semaines