MAP553 Regression

Chapter 2: Linear regression model

Contents

2.1. Introduction 1
2.2. Modelisation 2
2.3. Assumed postulates 5
2.4. Estimation of the model under the assumption of rank 6
2.5. Statistical Properties of OLSE under Postulates [P1] - [P3] 7
2.6. General Cochran theorem 8
2.7. Residuals and variance estimation 9
2.7. Inference in the Gaussian linear regression model 10
2.7.1. Maximum likelihood estimator and its properties 10
2.7.2. Intervals and regions of confidence 12
2.7.3. Hypothesis tests 13
2.7.4. Confidence interval and bootstrap 15
2.8. Prediction 15
2.9. R^2-coefficient 16

Let formalize the ideas seen in the previous chapter, *i.e.* write a probabilistic model to model the data, study the properties of this model, estimate the parameters and study the properties of these estimators (bias, variance, properties asymptotic . . .). Finally we will develop decision tools (confidence intervals, tests).

Modeling the observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ means we suppose that $\boldsymbol{y}$ is the realization of a random variable $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ whose probability law is described.

2.1. Introduction

The aim of this course is to study the relation between a random variable, denoted Y, with other variables $X_1, \ldots, X_p$ called *predictors* (*regressor* or *explanatory variables*. . .). We focus on two

objectives : **explain** and **predict**. For this, we try to construct a function f such that

$$Y \approx f(X_1, \ldots, X_p).$$

The construction will be based on the n observations $(y_i)_{i=1,\cdots,n}$ of the variable Y and n observations $(X_{i1}, \cdots, X_{ip})_{i=1,\cdots,n}$ of the variables $X_1, \ldots, X_p$.

Example 1 *We try to explain and predict gasoline consumption (in liters per 100 km) of different automobile models based on several variables. For this, we have the following characteristics for 31 different cars:*

- `Consommation` = *Fuel consumption in liters per 100 km.*
- `Prix` = *Vehicle price in Swiss francs.*
- `Cylindree` = *Cylinder capacity in cm3.*
- `Puissance` = *Power in kW.*
- `Poids` = *Weight in kg.*

In this example, Y is the variable `Consommation`*. The variables X_j correspond to the other 4 variables.*

In this course, we are interested in the *linear model*, specifying a linear relation between the observed variable Y and the predictors stocked in the matrix X, so f is a linear function in our course. The simple framework of linear regression makes it possible to obtain very rich results, which justifies its indepth study and its widespread use among practitioners.

2.2. Modelisation

In the linear regression setting, the function f is linear. It's mean that there exists $\beta = (\beta_1, \ldots, \beta_p)^T$ such that for all $(X_1, \ldots, X_p)$,

$$f(X_1, \ldots, X_p) = \beta_1 X_1 + \cdots + \beta_p X_p.$$

The vocabulary is the following :

- The variable to explain Y is called the **response variable**,
- The explanatory variables X_j are called the **predictors** or **regressor** or **covariates**,...
- The β_j are the **regression coefficients**,

Regression linear model We can write the regression linear model as follows:

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

Comments:

- ☛ The ε_i are called the **error**.
- ☛ For each i, we modelize $\varepsilon_i = Y_i - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_p X_{ip}$ as a random variable. This stochastic term models the measurement errors and the impact of all the variables not taken into account by the model. We will see that the hypothesis $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ will often be assumed.
- ☛ The value of coefficients β_j measure the importance of the effect of each predictors X_j on the variable to be explained Y. We will see how to test if a predictor has a significant influence on the variable to explain, we say that the predictor is relevant. But also and above all, this model is constructed to predict the typical values that can take a new observation Y_{n+1} of which we only know the values of the associated predictors.

Matrix form of the regression linear model Let us denote $X_j = (X_{1j}, \ldots, X_{nj})^T \in \mathbb{R}^n$ and $x_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$.

Then define the matrix X of size $n \times p$ such that

$$X = \underbrace{(X_1, \cdots, X_p)}_{p \text{ predictors}} = \underbrace{\begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_i^T \\ \vdots \\ x_n^T \end{pmatrix}}_{n \text{ observations}} = \underbrace{\begin{pmatrix} X_{11} & \cdots & X_{1p} \\ X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots \\ X_{i1} & \cdots & X_{ip} \\ \vdots & \vdots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}}_{n \text{ lines} \times p \text{ columns}}.$$

When the first column is only composed by 1, the β_1 parameter is called *intercept*

$$X = \begin{pmatrix} 1 & X_{12} & \cdots & X_{1p} \\ 1 & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{i2} & \cdots & X_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

By using the matrix forms, we obtain the following definition of the regression linear model.

$$Y = X\beta + \varepsilon,$$

where
- $Y = (Y_1, \ldots, Y_n)^T$ is a random vector in $\mathbb{R}^n$.
- $\beta = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is the unknown parameters vector.
- $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$ is the vectors of errors.
- The matrix X is supposed to be deterministic (not random) in this course.

Identifiability problems : The model is identifiable if the matrix X satisfies the following property: if $\beta \in \mathbb{R}^p$ and $\beta' \in \mathbb{R}^p$ satisafy $X\beta = X\beta'$ then $\beta = \beta'$. We have then, uniqueness of the vector of the parameters.This is a very important hypothesis and is realized as soon as the matrix X is injective, which is equivalent to $\text{rang}(X) = p$. Most of the time, we'll assume $p \leq n$ and $\text{rang}(X) = p$. Therefore, we will make the following hypothesis.

Rank assumption : The matrix X is full rank, $\text{rang}(X) = p$.

Comments:

- ☛ Most of the time, we'll assume ε centered. Other hypotheses (**Postulates**) (gaussianity, homoscedasticity, ...) can also be formulated.
- ☛ In the introductory example, we could have used other predictors, such as the square of the logarithm of one of the predictor, for example. This is possible by involving other variables.
- ☛ Thus, a linear model does not mean that the relationship between the predictors and the response variable is linear, but that the model is linear in the β_j parameters.

Definition 1

• *We define the* ***linear regression model*** *as follow :*

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

• *The matrix form is the following*

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

✎ *If we considère an intercept in our model, our model is written in its matrix form*

$$Y = \beta_0 \mathbb{1}_n + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \text{ or } Y = \beta_1 \mathbb{1}_n + \beta_2 X_2 + \cdots + \beta_p, X_p + \varepsilon.$$

where $\mathbb{1}_n = (1, \cdots, 1)^T \in \mathbb{R}^n$.

2.3. Assumed postulates

Note that we speak of *postulates* in the sense that we can not formally show that they are verified by statistical tests. We will use graphical tools to test them.

Postulat [P1] : $\forall i = 1, \cdots, n \quad \mathbb{E}_\beta[\varepsilon_i] = 0.$

Errors are centered. In practice, this means that the model is correct and that we have not forgotten a relevant term : the model is linear.

Postulat [P2] : $\forall i = 1, \cdots, n \quad \mathbb{V}\text{ar}_\beta[\varepsilon_i] = \sigma^2 > 0.$

The errors are of constant variance. We speak of a *homoscedastic* model, as opposed to a *heteroscedastic* model where the error term would not have the same variance for all observations.

Postulat [P3] : $\forall i \neq j \quad \mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_j) = 0.$

Errors are uncorrelated. Thus, the observations are assumed to be uncorrelated, *i.e.* independent sampling or the results of a physical experiment conducted under independent conditions. Problems can arise when time has an importance in the phenomenon.

Postulat [P4] : $\forall i = 1, \cdots, n \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$

The errors are Gaussian. This postulate is the least important as we will be able to do without it if the number of observations is large (larger than 20 or 30 observations). It is difficult to detect the non gaussianity of errors. We will see that **R** propose graphical tools to try to validate or not this postulate.

2.4. Estimation of the model under the assumption of rank

Some notations:

- Let X be the *design* matrix of size $n \times p$ and of full rank p, we denote by

$$[X] := \mathcal{I}m(X)$$

  the space generated by the p columns of X.

- Let denote by P_X the orthogonal projection into $[X]$. Then, $P_{X^\perp} = I - P_X$ is the orthogonal projection matrix into $[X]^\perp$ the orthogonal space at $[X]$.

Definition 2 *We define the **ordinary least square estimator (OLSE)** of β in the model* (1)*, the vector $\widehat{\beta} \in \mathbb{R}^p$ such that*

$$\widehat{\beta} = \arg\min_{u \in \mathbb{R}^p} \|Y - Xu\|^2.$$

✎ *Where $\|u\|^2 = \sum_k u_k^2$ denote the euclidian norm*

Proposition 1 *In the model* (1) *and under the Rank assumption, the design matrix X is injective and*

$$\widehat{\beta} = (X^T X)^{-1} X^T Y.$$

Proof : By the definition of the projected orthogonal, we have

$$\|Y - P_X Y\|^2 = \min_{v \in [X]} \|Y - v\|^2.$$

Thus,

$$P_X Y = \arg\min_{v \in [X]} \|Y - v\|^2.$$

Which implies $\exists \widehat{\beta} \in \mathbb{R}^p$ such that $P_X Y = X\widehat{\beta}$. For all $k \in \{1, \cdots, p\}$, we note X_k the $k - th$ column of X, then it comes

$$\langle X_k, P_X Y - Y\rangle = 0 \Leftrightarrow \langle X_k, X\widehat{\beta} - Y\rangle = 0 \Leftrightarrow X_k^T(X\widehat{\beta} - Y) = 0.$$

Thus,

$$X^T(X\widehat{\beta} - Y) = 0_p \quad \Leftrightarrow \quad X^T X\widehat{\beta} = X^T Y$$

As X is full rank, X^TX$ is inversible and for all $u \in \mathbb{R}^p \Leftrightarrow Xu \in [X]$. As we consider the Euclidean norm, $X\widehat{\beta}$ is the orthogonal projection of Y into $[X]$:

$$X\widehat{\beta} = X(X^T X)^{-1} X^T Y.$$

By injectivity of X, we get the result $\widehat{\beta} = (X^T X)^{-1} X^T Y$. □

Exercise 1 *Proof the result using the calculation of the partial derivatives of*

$$u \longmapsto \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} u_j X_{ij})^2.$$

Comments:

- It may be noted that we have shown the relation:

$$P_X Y = X\widehat{\beta} = \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_p X_p.$$

- Note that

$$\widehat{Y} = P_X Y \text{ and } P_X = X(X^T X)^{-1} X^T.$$

Exercise 2 *To which condition(s) the formula $P_{X_j} Y = \widehat{\beta}_j X_j$ is true?*

2.5. Statistical Properties of OLSE under Postulates [P1] - [P3]

Study now the statistical properties of the EMCO:

Proposition 2 *In the model* (1)*, under the Rank assumption and under **[P1]**, the OSLE $\widehat{\beta}$ is an unbiased estimator of β:*

$$\forall \beta \in \mathbb{R}^p, \quad \mathbb{E}_\beta[\widehat{\beta}] = \beta.$$

Proof : As X is a deterministic matrix and $Y = X\beta + \varepsilon$, it comes

$$\mathbb{E}_\beta[\widehat{\beta}] = \mathbb{E}_\beta[(X^T X)^{-1} X^T Y] = \mathbb{E}_\beta[(X^T X)^{-1} X^T (X\beta + \varepsilon)] = \underbrace{(X^T X)^{-1} X^T X}_{identity}\beta + (X^T X)^{-1} X^T \underbrace{\mathbb{E}_\beta[\varepsilon]}_{0_n} = \beta. \quad \square$$

Theorem 1 (Gauss-Markov theorem) *In the model* (1)*, under the Rank assumption and under **[P1]–[P3]**, the OLSE $\widehat{\beta}$ is such that*

$$\mathbb{V}ar_\beta(\widehat{\beta}) = \sigma^2(X^TX)^{-1}$$

*and $\widehat{\beta}$ is the estimator of "minimum variance" **among linear and unbiased estimators.***

Proof :

• Note that $\mathbb{V}\text{ar}_\beta(Y) = \mathbb{V}\text{ar}_\beta(X\beta + \varepsilon) = \mathbb{V}\text{ar}_\beta(\varepsilon) = \sigma^2\mathbb{I}_n$, then

$$\mathbb{V}\text{ar}_\beta(\widehat{\beta}) = \mathbb{V}\text{ar}_\beta\left((X^TX)^{-1}X^TY\right) = (X^TX)^{-1}X^T\underbrace{\mathbb{V}\text{ar}_\beta(Y)}_{\sigma^2\mathbb{I}_n}X(X^TX)^{-1} = \sigma^2\underbrace{(X^TX)^{-1}X^TX}_{\mathbb{I}_p}(X^TX)^{-1} = \sigma^2(X^TX)^{-1}.$$

• Let $\tilde{\beta} = CY$ be any linear and unbiased estimators where C is a matrix of size $p \times n$ then as $\tilde{\beta}$ is unbiased we have:

$$\mathbb{E}_\beta[CY] - \beta = 0 \Leftrightarrow CX\beta - \beta = 0_p \Leftrightarrow (CX - \mathbb{I}_p)\beta = 0_p \Leftrightarrow CX = \mathbb{I}_p \text{ and } \mathrm{X^TC^T} = \mathbb{I}_\mathrm{p}$$

Therefore,

$$\begin{aligned}\mathbb{V}\text{ar}_\beta(\tilde{\beta}) - \mathbb{V}\text{ar}_\beta(\widehat{\beta}) &= \sigma^2(CC^T - (X^TX)^{-1}) = \sigma^2C(\mathbb{I}_n - X(X^TX)^{-1}X^T)C^T \\ &= \sigma^2C(\mathbb{I}_n - P_X)C^T = \sigma^2CP_{X^\perp}C^T = \sigma^2CP_{X^\perp}P_{X^\perp}^TC^T,\end{aligned}$$

$P_{X^\perp}$ is the orthogonal projection into the orhtogonal space at $[X]$. Then, $\forall u \in \mathbb{R}^p$ we get

$$u^T(\mathbb{V}\text{ar}_\beta(\tilde{\beta}) - \mathbb{V}\text{ar}_\beta(\widehat{\beta}))u = \sigma^2u^TCP_{X^\perp}P_{X^\perp}^TC^Tu = \|P_{X^\perp}^TC^Tu\|^2 \geq 0. \quad \square$$

2.6. General Cochran theorem

Theorem 2 (General Cochran theorem) *Let $W \sim \mathcal{N}(m, \sigma^2I_d)$ be a gaussian vector in $\mathbb{R}^d$ and $E_1 \oplus \cdots \oplus E_r$ a decomposition of $\mathbb{R}^d$ into two-by-two orthogonal subspaces of dimension $d_1, \ldots, d_r$. For all $j = 1, \cdots, r$, we define the random vectors $W_{E_1}, \ldots, W_{E_r}$ such that*

$$W_{E_j} = P_{E_j}\left(\frac{W - m}{\sigma}\right)$$

is the orthogonal projection of $\frac{W-m}{\sigma}$ into E_j. Then:

The random vectors W_{E_j} are mutually independent. For all $j = 1, \cdots, r$, the random vectors $\|W_{E_j}\|^2$ are mutually independent and $\|W_{E_j}\|^2 \sim \chi^2_{d_j}$.

Comment:

☛ Cochran's theorem is most often applied with $W \sim \mathcal{N}(0, \sigma^2 I_d)$, $\mathbb{R}^d = E_1 \oplus E_2$ and $E_2 = E_1^{\perp}$. In this case, W_{E_1} and $W_{E_1^{\perp}}$ are indenpendent vectors and

$$\sigma^{-2}\|W_{E_1}\|^2 \sim \chi^2_{d_1}, \quad \sigma^{-2}\|W_{E_1^{\perp}}\|^2 \sim \chi^2_{d-d_1}.$$

2.7. Residuals and variance estimation

Recall first that the vector ε is the vector of errors. It is also called the vector of theoretical residues.

Definition 3 *In the model* (1)*, we define the* ***residuals*** *(or estimated residuals) as follows:*

$$\widehat{\varepsilon} = Y - \widehat{Y} = Y - X\widehat{\beta} = Y - P_X Y = (I - P_X)Y = P_{X^{\perp}}Y = P_{X^{\perp}}\varepsilon.$$

Proposition 3 *In the model* (1)*, under the Rank assumption and under* ***[P1]–[P3]****, we have*

- $\mathbb{E}_\beta[\widehat{\varepsilon}] = 0_n$
- $\mathbb{V}ar_\beta[\widehat{\varepsilon}] = \sigma^2 P_{X^{\perp}}$.
- $\mathbb{C}ov_\beta(\widehat{\varepsilon}, \widehat{Y}) = 0_{n\times n}$.

Proof : By using the fact that $\widehat{\varepsilon} = P_{X^{\perp}}Y$, the two first points are obvious. Then, as ε is centered

$$\mathbb{C}\mathrm{ov}_\beta(\widehat{\varepsilon}, \widehat{Y}) = \mathbb{C}\mathrm{ov}_\beta(P_{X^{\perp}}Y, P_X Y) = 0_{n\times n}. \quad \square$$

Proposition 4 *In the model* (1)*, under the Rank assumption and under* ***[P1]–[P3]****, an unbiased estimator of* σ^2 *is* $\widehat{\sigma}^2$ *defined by*

$$\widehat{\sigma}^2 = \frac{\|\widehat{\varepsilon}\|^2}{n-p} = \frac{\|Y - X\widehat{\beta}\|^2}{n-p} = \frac{\|Y - P_X Y\|^2}{n-p} = \frac{\|\widehat{\varepsilon}\|^2}{n-p} = \frac{\|P_{X^{\perp}}Y\|^2}{n-p}.$$

Proof : Note that $\sigma^{-1}\varepsilon \sim \mathcal{N}(O_n, \mathbb{I}_n)$ and $P_{X^{\perp}}$ is an orthogonal projector of rank $(n-p)$ then by Cochran theorem

$$\|P_{X^{\perp}}(\sigma^{-1}\varepsilon)\|^2 \sim \chi^2_{(n-p)}.$$

Then, it comes :

$$\mathbb{E}_\beta[\|\widehat{\varepsilon}\|^2] = \mathbb{E}_\beta[\|P_{X^{\perp}}\varepsilon\|^2] = \sigma^2\mathbb{E}_\beta[\|P_{X^{\perp}}(\sigma^{-1}\varepsilon)\|^2] = \sigma^2(n-p)$$

as the expectation of the Chi-2 random variable of $(n-p)$ degree of freedom. $\square$

2.7. Inference in the Gaussian linear regression model

In this section, we assume the Rank assumption and **[P1]–[P4]** satisfied, then

$$\varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n).$$

In others words the ε_i are i.i.d. This assumption allow us to consider maximum likelihood estimators and to build confidence regions and tests.

Comments:

- ☛ In this section, $\widehat{\beta}$ is the ordinary maximum likelihood estimator which is in this gaussian setting exactly the least squares estimator.
- ☛ Note that the maximum likelihood estimator of σ^2 is $\widehat{\sigma}^2_{ML} = \frac{\|Y - X\widehat{\beta}\|^2}{n} = \frac{\|\widehat{\varepsilon}\|^2}{n}$. which is a biased estimator. We recall that an unbiased estimator of σ^2 is $\widehat{\sigma}^2 = \frac{\|Y - P_X Y\|^2}{n-p} = \frac{\|Y - X\widehat{\beta}\|^2}{n-p} = \frac{\|\widehat{\varepsilon}\|^2}{n-p}$.

2.7.1. Maximum likelihood estimator and its properties

Proposition 5 *the maximum likelihood estimatorof* (β, σ^2) *est le vecteur*

$$(\widehat{\beta}, \widehat{\sigma}^2_{ML}) = \left(X^T X)^{-1} X^T Y, (n-p)/n \times \widehat{\sigma}^2\right) = \left(X^T X)^{-1} X^T Y, \frac{\|Y - X\widehat{\beta}\|^2}{n}\right)$$

Proof : The demonstration follows from the calculation. □

Proposition 6 *In the model* (1)*, under the Rank assumption and under **[P1]–[P4]**, we have*

- $\widehat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$
- $\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$
- $\widehat{\beta}$ *and* $\widehat{\sigma}^2$ *are indenpendent.*

Proof : The first point is obvious. For the second point, we can write :

$$\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} = \frac{\|Y - P_X Y\|^2}{\sigma^2} = \frac{\|P_{X^\perp} Y\|^2}{\sigma^2} = \left\|P_{X^\perp}(\sigma^{-1}\varepsilon)\right\|^2.$$

We conclude by Cochran theorem. Then, for the last point we show that

$$\widehat{\sigma}^2 = \frac{\|P_{X^\perp}\varepsilon\|^2}{n-p} \quad \text{and} \quad \widehat{\beta} = \beta + (X^T X)^{-1} X^T P_X \varepsilon.$$

Then

$$\mathbb{C}\text{ov}_\beta(\widehat{\sigma}^2, \widehat{\beta}) = \frac{(X^T X)^{-1} X^T}{n-p} \mathbb{C}\text{ov}_\beta\left(\|P_{X^\perp}\varepsilon\|^2, P_X \varepsilon\right) = O_p. \quad \square$$

The following result will be very important for building trust regions and hypothesis testing.

Theorem 3 *In the model* (1)*, under the Rank assumption and under* ***[P1]–[P4]****, let* $\widehat{\sigma}^2$ *and* $\widehat{\beta}$ *the estimator defined in proposition 6.*

• *For all vectors* $c \in \mathbb{R}^p$*, we have*

$$\frac{c^T\widehat{\beta} - c^T\beta}{\widehat{\sigma}\sqrt{c^T(X^TX)^{-1}c}} \sim t_{(n-p)}.$$

• *For all matrices* C *of size* $q \times p$ *and of full rank* q $(q \leq p)$*, we have*

$$\frac{(C\widehat{\beta} - C\beta)^T(C(X^TX)^{-1}C^T)^{-1}(C\widehat{\beta} - C\beta)}{q\widehat{\sigma}^2} \sim \mathcal{F}(q, n-p).$$

• *We set two vector subspaces* V *and* W *where* W *is a vector subspace of* V*. We assume* $q = dim(W) < p = dim(V)$*. If* $X\beta \in W \subset V$*, then*

$$F = \frac{\|P_WY - P_VY\|^2/(p-q)}{\|Y - P_VY\|^2/(n-p)} \sim \mathcal{F}(p-q, n-p),$$

where P_VY *denote the orthogonal projection of* Y *into* V *and* P_WY *denote the orthogonal projection of* Y *into* W*.*

Proof :

• The first point is immediate by applying a probability result

$$\begin{cases} A \sim \mathcal{N}(0,1), \\ B \sim \chi^2(n-p) \\ A \text{ independent of } B \end{cases} \Leftrightarrow \frac{A}{\sqrt{B/(n-p)}} \sim t_{(n-p)}.$$

• For the second point : As the rank of C is $q \leq p$, the matrix, of size $q \times q$ and of rank q, $C(X^TX)^{-1}C^T$ is inversible. There exists a inversible and symetric matrix Δ such that $C(X^TX)^{-1}C^T =: \Delta^2$. Then $\Delta^{-1}(C\widehat{\beta} - C\beta) \sim \mathcal{N}(0, \sigma^2 I)$. And finally by using the following probability result

$$\begin{cases} A \sim \mathcal{N}(0_q, \mathbb{I}_q), \\ B \sim \chi^2(n-p) \\ A \text{ independent of } B \end{cases} \Leftrightarrow \frac{\|A\|^2/q}{B/(n-p)} \sim F(q, n-p).$$

$$\frac{(C\widehat{\beta} - C\beta)^T(C(X^TX)^{-1}C^T)^{-1}(C\widehat{\beta} - C\beta)}{q\widehat{\sigma}^2} \sim \mathcal{F}(q, n-p).$$

• The third point stems from the application of Cochran's theorem which implies that

$$P_WY - P_VY \underbrace{=}_{as\ X\beta \in W} P_W\varepsilon - P_V\varepsilon \in V$$

is independent of $Y - P_VY = P_{V^\perp}\varepsilon$. □

2.7.2. Intervals and regions of confidence

By using the previous theorem, we can determine the intervals and confidence regions of the unknownparameters. We have the following theorem which is based on the Theorem~3 :

Theorem 4 *In the model* (1)*, under the Rank assumption and under **[P1]–[P4]**, let $\widehat{\sigma}^2$ and $\widehat{\beta}$ the estimator defined in proposition 6. Let $\alpha \in]0,1[$.*

• *For all $c \in \mathbb{R}^p$,*

$$\left[c^T\widehat{\beta} - t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{c^T(X^TX)^{-1}c}\ ,\ c^T\widehat{\beta} + t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{c^T(X^TX)^{-1}c}\right],$$

where $t_{n-p,1-\alpha/2}$ is the quantile of order $1-\alpha/2$ the Student's law at $n-p$ degrees of freedom, is a confidence interval for $c^T\beta$ of exactly $1-\alpha$ level.

• *Let q_1 and q_2 such that if $Z \sim \chi^2(n-p)$, then $P(q_1 \leq Z \leq q_2) = 1-\alpha$. Then*

$$\left[\frac{(n-p)\widehat{\sigma}^2}{q_2}\ ,\ \frac{(n-p)\widehat{\sigma}^2}{q_1}\right]$$

is a confidence interval for σ^2 of exactly $1-\alpha$ level.

• *Let C be a matrix of size $q \times p$ and rank $q \leq p$. Then*

$$R = \left\{a :\quad \frac{(C\widehat{\beta}-a)^T(C(X^TX)^{-1}C^T)^{-1}(C\widehat{\beta}-a)}{q\widehat{\sigma}^2} \leq f_{q,n-p,1-\alpha}\right\}$$

with $f_{q,n-p,1-\alpha}$ the quantile of order $1-\alpha$ of the Fisher law at $(q, n-p)$ degrees of freedom, is a confidence region (set) for $C\beta$ of exactly $1-\alpha$ level.

Proof : Immediate. □

Corollary 1 *In the model* (1)*, under the Rank assumption and under **[P1]–[P4]**, let $\widehat{\sigma}^2$ and $\widehat{\beta}$ the estimator defined in proposition 6. Let $\alpha \in]0,1[$. Then for all $j = 1, \cdots, p$*

$$\left[\widehat{\beta}_j - t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}\ ,\ \widehat{\beta}_j + t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}\right],$$

where $t_{n-p,1-\alpha/2}$ is the quantile of order $1-\alpha/2$ the Student's law at $n-p$ degrees of freedom, is a confidence interval for β_j of exactly $1-\alpha$ level.

Proof : Direct application of the first point of the theorem~4 with $c \in \mathbb{R}^p$ such that $c_k = 0$ for all $k \neq j$ et $c_j = 1$. □

Corollary 2 *In the model* (1)*, under the Rank assumption and under **[P1]–[P4]**, let $\widehat{\sigma}^2$ and $\widehat{\beta}$ the estimator defined in proposition 6. Let $\alpha \in]0,1[$. Let $x_0 = (x_{01}, \cdots, x_{0p})^T$, then*

$$\left[x_0^T\widehat{\beta} - t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0}\ ,\ x_0^T\widehat{\beta} + t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0} \right],$$

where $t_{n-p,1-\alpha/2}$ is the quantile of order $1-\alpha/2$ the Student's law at $n-p$ degrees of freedom, is a confidence interval for $x_0\beta = \mathbb{E}_\beta[Y_0]$ of exactly $1-\alpha$ level.

Proof : The proof is left in exercice. □

Exercise 3 *Assume $p \geq 2$ and set $c_{ij} := ((X^TX)^{-1})_{ij}$.*

1. *Determine a confidence interval for β_1 and β_2 of level $1-\alpha$ based on the $\widehat{\beta}$, $\widehat{\sigma}$ and the c_{ij}.*
2. *Deduce a confidence region for the vector (β_1, β_2).*
3. *Answer the previous question but using the 3rd point of the theorem.*

2.7.3. Hypothesis tests

We want to first test

$$H_0 :\ c^T\beta = a \quad \text{vs} \quad H_1 :\ c^T\beta \neq a$$

for $c \in \mathbb{R}^p$ and $a \in \mathbb{R}$. Note that, within this framework, there is the following test

$$H_0 :\ \beta_j = 0 \quad \text{vs} \quad H_1 :\ \beta_j \neq 0, \qquad j \in \{1, \ldots, p\}.$$

Theorem 5 *Let $c \in \mathbb{R}^p$ and $\alpha \in]0,1[$, we set*

$$T := \frac{c^T\widehat{\beta} - a}{\widehat{\sigma}\sqrt{c^T(X^TX)^{-1}c}}$$

and $t_{n-p,1-\alpha/2}$ is the quantile of order $1-\alpha/2$ of the Student law of $n-p$ degrees of freedom. Then, $\phi(Y) = 1_{\{|T| > t_{n-p,1-\alpha/2}\}}$ is a test of size α of

$$H_0 :\ c^T\beta = a \quad vs \quad H_1 :\ c^T\beta \neq a.$$

Proof : Immediate. □

The previous test can be generalized as in the following theorem.

Theorem 6 *Let V and W be two vector subspaces where W is a vector subspace of V. Assume $q = dim(W) < p = dim(V)$. We set*

$$F := \frac{\|P_W Y - P_V Y\|^2/(p-q)}{\|Y - P_V Y\|^2/(n-p)}$$

Where $P_V Y$ is the orthogonal projection of Y into V and $P_W Y$ is the orthogonal projection of Y into W. Then,

$$\phi(Y) = 1_{\{F > f_{p-q,n-p,1-\alpha}\}},$$

where $f_{p-q,n-p,1-\alpha}$ is the quantile of order $1-\alpha$ of the Fisher law at $(p-q, n-p)$degrees of freedom, is a test of sizeα of

$$H_0 : X\beta \in W \subset V \quad vs \quad H_1 : X\beta \in V \setminus W.$$

Proof : Immediate. □

Comments:

- The idea of the previous theorem is to select the smallest model (the smallest number of needed predictors): Let denote by $\widehat{Y}_0$ the projection of Y under H_0 (the smaller than the model under H_1) and $\widehat{Y}_1$ the projection of Y under H_1. If $\widehat{Y}_0$ is "closed to" we keep $\widehat{Y}_0$ ("sparse selection"). "Close" in the sense of the euclidian distance $\|\widehat{Y}_0 - \widehat{Y}_1\|^2$ standardized by the estimation error $\|Y - \widehat{Y}_1\|^2$.
- The previous test is justified because if H_0 is true then the numerator is expected to be small.
- We call the **Global Fisher test** the following test

Global Fisher test $\quad H_0 : Y = \beta_0 \mathbb{1}_n + \varepsilon \quad$ vs $\quad H_1 : Y = \beta_0 \mathbb{1}_n + \sum_{j=1}^{p} \beta_j X_j + \varepsilon.$

  We test under H_0 the model reduced to the intercept aigainst the full model under H_1.

Exercise 4 (Test between nested models) *Let $1 \leq q < p$. How to test if regressors $X_{q+1}, \ldots, X_p$ are unrelevant ?*

Exercise 5 (Global Fisher test) *Let $p \geq 2$. Assume $X_1 = (1, \ldots, 1)^T$. We want to test if at least one predictor variable comes into play in the model. Prove that the Fisher test answers the problem and that in this case the variable F is written:*

$$F = \frac{\|\widehat{Y} - \overline{Y}\mathbb{1}_n\|^2/(p-1)}{\widehat{\sigma}^2} = \frac{\|\widehat{Y} - \overline{Y}\mathbb{1}_n\|^2/(p-1)}{\|Y - \widehat{Y}\|^2/(n-p)}$$

with $\widehat{Y} = P_X Y$. What is the law of F under H_0 ?

2.7.4. Confidence interval and bootstrap

The aim of this section is to present the regression bootstrap method in order to obtain a confidence interval for β without additional assumptions about the distribution of ε. The algorithm is as follows:

1. The method is built from the formulas seen previously. We estimate (β, ε) by $(\widehat{\beta}, \widehat{\varepsilon})$. We note $\widehat{F}_n$ the empirical distribution of $\widehat{\varepsilon}_i$:
$$\widehat{F}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\widehat{\varepsilon}_i}.$$
2. We sample n estimated residuals (say $\widehat{\varepsilon}_i^*$) with replacement from estimated residuals $\widehat{\varepsilon}_i$.
3. Based on these $\widehat{\varepsilon}_i^*$, we construct
$$Y^* = X\widehat{\beta} + \widehat{\varepsilon}^* \in \mathbb{R}^n$$
4. We estimate β from Y^*
$$\widehat{\beta}^* = (X^T X)^{-1} X^T Y^*.$$

The bootsrap theory gives
$$\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \overset{loi}{\approx} \sqrt{n}(\widehat{\beta} - \beta).$$
To estimate the distribution of $\sqrt{n}(\widehat{\beta}^* - \widehat{\beta})$, we repeat M times the previous operation. For $k = 1, \ldots, M$, we have at our disposal a vector $(\widehat{\beta}^*(k), \widehat{\varepsilon}^*(k))$. We obtain the distribution of $\widehat{\beta}_j$ using the histogram of the $(\widehat{\beta}_j^*(k))_{k=1,\ldots,M}$. We deduce the approximate quantiles.

2.8. Prediction

Consider a new individual $n+1$ such that
$$x_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p})^T$$
are the values of the regressors associated with this individual. We want to predict the value of Y_{n+1}. We set:
$$Y_{n+1} = x_{n+1}^T \beta + \varepsilon_{n+1}$$
where ε_{n+1} is a random variable independent of all the $(\varepsilon_i)_{1\le i\le n}$. Moreover ε_{n+1} has the same distribution as each ε_i. We estimate β by $\widehat{\beta}$ which is a function of $Y_1, \ldots, Y_n$. Then we estimate Y_{n+1} by
$$\widehat{Y}_{n+1} = x_{n+1}^T \widehat{\beta}.$$
We get :
$$\mathbb{E}_\beta[\widehat{Y}_{n+1}] = x_{n+1}^T \beta, \quad \mathbb{V}\mathrm{ar}_\beta(\widehat{Y}_{n+1}) = \sigma^2 (x_{n+1}^T (X^T X)^{-1} x_{n+1}).$$

Theorem 7 *In the model* (1)*, under the Rank assumption and under **[P1]–[P4]**. Assume* $\varepsilon = (\varepsilon_i)_{1\le i\le n} \sim \mathcal{N}(0, \sigma^2 I)$*. Un intervalle de confiance pour* Y_{n+1} *est donné par :*

$$\left[x_{n+1}^T \widehat{\beta} - t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{1 + x_{n+1}^T (X^T X)^{-1} x_{n+1}} \ , \ x_{n+1}^T \widehat{\beta} + t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{1 + x_{n+1}^T (X^T X)^{-1} x_{n+1}} \right],$$

où $t_{n-p,1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-p$ degrés de liberté.

Proof : We show that

$$Y_{n+1} - \widehat{Y}_{n+1} \sim \mathcal{N}\left(0, \sigma^2(1 + x_{n+1}^T (X^T X)^{-1} x_{n+1})\right). \quad \square$$

2.9. R^2-coefficient

In this section, we consider the model~(1), under the Rank assumption and under **[P1]–[P3]**. We also assume an intercept, *i.e* the first column of the matrix X is the one vector $X_1 := \mathbb{1}_n := (1, 1, \ldots, 1)^T \in \mathbb{R}^n$.

Comment:

- First note that the projection of $\left(Y - \overline{Y}\mathbb{1}_n\right)$ into $[X] = \mathcal{I}m(X)$ is such that

$$P_X\left(Y - \overline{Y}\mathbb{1}_n\right) = P_X Y - P_X \overline{Y}\mathbb{1}_n = \widehat{Y} - \overline{Y}\mathbb{1}_n.$$

- Then as $\left(Y - \widehat{Y}\right) \in [X]^\perp$ and $\left(\widehat{Y} - \overline{Y}\mathbb{1}_n\right) \in [X]$ are orhtogonal vectors, we get by Pythagoras formula

$$\|Y - \overline{Y}\mathbb{1}_n\|^2 = \|Y - \widehat{Y} + \widehat{Y} - \overline{Y}\mathbb{1}_n\|^2 = \|Y - \widehat{Y}\|^2 + \|\widehat{Y} - \overline{Y}\mathbb{1}_n\|^2 \tag{2}$$

In this section $\overline{y} = n^{-1}\sum_{i=1}^n y_i$ and $\widehat{y} = X\widehat{\beta}$, where the y_i are the observations.

Definition 4 *We denote by TSS (SCT in french) the total sum of squares*

$$TSS := \|y - \overline{y}\mathbb{1}_n\|^2 = \sum_{i=1}^n (y_i - \overline{y})^2.$$

Definition 5 *We denote by RSS (SCR in french) the residual sum of squares*

$$RSS := \|y - \widehat{y}\|^2 = \sum_{i=1}^n (y_i - \widehat{y}_i)^2.$$

Definition 6 *We denote by MSS ($SCEM$ or SCM or SCE in french) model sum of squares by the model*

$$MSS := \|\widehat{y} - \overline{y}\mathbb{1}_n\|^2 = \sum_{i=1}^{n} (\widehat{y_i} - \overline{y})^2.$$

Comments:

- The MSS corresponds to the variability (variance) explained by the model.
- The RSS corresponds to the variability (variance) not explained by the model.
- The TSS corresponds to the total variability (variance).

Definition 7 *From* (2)*, we have the following relation*

$$TSS = MSS + RSS$$

Comment:

- If the model is "good" enough, the part of total variability must be explained by the model and inversely the part of variability not explained by the model shall be small. It is the role of the determination coefficient to quantify this notion.

Definition 8 *We define the* ***determination coefficient*** R^2 *as follows*

$$R^2 = \frac{MSS}{TSS} = \frac{\|\widehat{y} - \overline{y}\mathbb{1}_n\|^2}{\|y - \overline{y}\mathbb{1}_n\|^2}.$$

We can rewrite the R^2-coefficient

$$R^2 = 1 - \frac{RSS}{TSS}.$$

Comments:

- First note that $R^2 \in [0, 1]$.
- Note that in the simplest model reduced to the intercept $Y = \beta\mathbf{1} + \varepsilon$, The least square estimator of the real parameter β is $\widehat{\beta} = \overline{Y}$.
- The linear model is interesting only if the errors "$y_i - \widehat{y_i}$" are small relative to the errors "$y_i - \overline{y}$" that we would make if we took a model without regressors, i.e. a model reduced to the intercept. The linear model has an interest if

$$\frac{RSS}{TSS} = \frac{\|y - \widehat{y}\|^2}{\|y - \overline{y}\mathbb{1}_n\|^2} \to 0.$$

This is equivalent to

$$R^2 \to 1.$$

- ☛ Note that $R^2 \in [0,1]$. R^2-coefficient close to 1 "means" the predictors $X_2, \ldots, X_p$ "well" explain the model.
- ☛ Note that for the global Fisher test,

$$F = \frac{\|P_X Y - \overline{Y}\mathbb{1}_n\|^2/(p-1)}{\widehat{\sigma}^2} = \frac{(n-p)R^2}{(p-1)(1-R^2)}$$

  Thus, H_0 is rejected for large values of R^2.
- ☛ The R^2-coefficient is not a good criterion as it depends on the dimension p. Indeed R^2 increase with p or equivalentely RSS decrease with p. Take two models, the first one with p predictors and the second one with same predictors plus one, so a model of size $p+1$. Denote by RSS_p and RSS_{p+1} the RSS in the first and the second models, it comes as $\widehat{Y} = X\widehat{\beta}$

$$RSS(p+1) = \|Y - X\widehat{\beta}\|^2 = \min_{\beta \in \mathbb{R}^{p+1}} \|Y - X\beta\|^2 = \min_{\beta \in \mathbb{R}^{p+1}} \|Y - \sum_{j=1}^{p+1} X_j \beta_j\|^2 \leq \min_{\beta \in \mathbb{R}^p} \|Y - \sum_{j=1}^{p} X_j \beta_j\|^2 = RSS(p)$$

  So the R^2 coefficient is increasing with p. If we want to get rid of the dimension p, we use the adjusted R^2 define bellow.

Definition 9 *We define the **adjusted determination coefficient** R_a^2 as follows*

$$R_a^2 = 1 - \frac{(n-1)\|Y - \widehat{Y}\|^2}{(n-p)\|Y - \overline{Y}\mathbb{1}_n\|^2} = 1 - \frac{(n-1)\|\widehat{\varepsilon}\|^2}{(n-p)\|Y - \overline{Y}\mathbb{1}_n\|^2}.$$