Contents

Part I

Introduction to Linear Models

Chapter 1

Simple Linear Regression

1.1 Introduction

Originally, the term "regression" was coined by Sir Francis Galton. In 1885, while working on heredity, he sought to explain the height of sons in terms of their fathers. He noted that when a father was *taller than mediocrity* (average), his son tended to be shorter than he was, and, inversely, when the father was *shorter than mediocrity*, his son tended to be taller than him. These results led him to consider his theory of *regression toward mediocrity*. Nevertheless, the analysis of causality between several variables is much older and dates back from the middle of the XVIII$^{\text{th}}$ century. In 1757, R. Boscovich, proposed a method minimizing the sum of absolute values between a model and observations. Later, Legendre, in his famous 1805 article "On the Determination of the Orbit of Comets", introduced the method of least squares estimation. At the same time, Gauss published his work on the motion of celestial bodies thus developing the least squares method, which he argued to have used as early as 1795.

In this chapter, we introduce simple linear regression which can be seen as a statistical method for modeling a linear relationship between an explanatory variable (denoted X) and a response variable (denoted Y). This chapter presents a simple case of linear regression in order to better understand what is at stake, the problems raised, and the answers given.

This example uses data provided by UR2PI and CIRAD forest. When foresters want to assess a forest's health, they often consider the height of its trees. The taller the trees, the more productive the forest or plantation and we want to quantify production by wood volume. We therefore need to know the tree height in order to calculate the wood volume of the forest. We can do so by using a "truncated cone" formula. Nevertheless, measuring a 20 meter-tall tree is not easy and requires a dendrometer which measures the angle between the ground and the top of the tree. It therefore requires a clear view of the tree top and to be able to stand back far enough from the tree in order to obtain a precise measure. In many cases, it is impossible to measure the height because these two conditions are not met or the forester may not even have a dendrometer. We therefore need to estimate the height using a simpler measure, that of the circumference at 1.3 meter from the ground.

We collect data on eucalyptus in a plantation and want to build a model predicting the height from these observations. The planted areas are huge and spending a lot of time measuring is out of question. Therefore, estimating height by circumference is a satisfactory method which allows us to predict stand volume. In this plantation area, we have measured n=1429 pairs of circumference-height. The first 5 individuals can be found below:

Individual	ht	circ
1	18.25	36
2	19.75	42
3	16.50	33
4	18.25	39
5	19.50	43

Table 1.1 – Height and circumference (ht and circ) of the first 5 eucalyptus.

We want to explain the height by circumference. The data is plotted prior to modeling. Each of the points plotted in 1.1 is a circumference/height data pair for one tree.

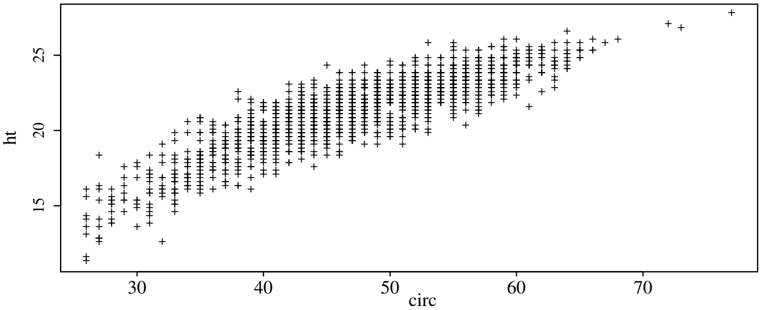


Figure 1.1 – Plot of $n = 1429$ measured eucalyptus.

To predict the height as a function of circumference, we look for a function f such as

$$y_i \approx f(x_i)$$

for each individual $i \in \{1, \ldots, 1429\}$.

1.2 Mathematical Modeling

In order to quantify the symbol ≈, we are going to choose a class of functions $\mathcal{G}$. This class represents all the possible functions of fit for modeling height as a function of circumference. Then we look for the function $\mathcal{G}$ which is as close as possible to the data according to a cost function. This is expressed:

$$\underset{f \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y_i - f(x_i)), \tag{1.1}$$

where n represents the number of individuals to analyze and $l(.)$ is called the *loss function* and $\mathcal{G}$ a set of given functions. In the following section, we will discuss how to select the loss function and set $\mathcal{G}$.

1.2.1 Selecting the loss function

There are many loss functions $l(.)$, but the two main ones are as follows:

- $l(u) = u^2$ quadratic loss;
- $l(u) = |u|$ absolute loss.

These two functions are plotted in the following figure:

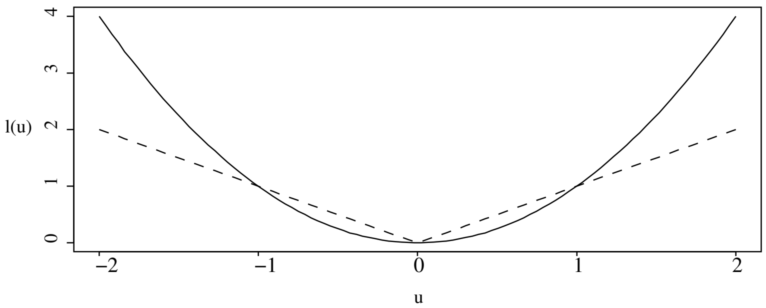


Figure 1.2 – Plot of absolute (dashed line) and quadratic (solid line) cost.

These functions are positive and symmetric, therefore producing the same value as the output whether the error is positive or negative, and canceling each other out when u is equal to zero. As can be seen in figure 1.2, compared to the absolute loss, quadratic loss gives greater importance to points which are further from the line of best fit, the distance being squared.

Despite its non-robust nature, the quadratic cost is the most commonly-used cost for several reasons: historical, ease of calculation, mathematical properties. In 1800, there were no computers and the quadratic cost made it possible to explicitly calculate estimates from data. Concerning the use of other cost functions, Gauss (1809) said: “But of all these principles, least squares is the simplest; with the others, we would need to carry out more complex calculations”.

1.2.2 Selecting the set of functions

If the class $\mathcal{G}$ is too large, for example the class of continuous functions ($\mathcal{C}_0$), then a large number of these functions minimizes the criterion (1.1). As a result, all the functions of the class go through all the points (interpolation) and, when possible, cancel the quantity $\sum_{i=1}^{n} l(y_i - f(x_i))$. The class of continuous functions may be too large.

We will start with the straight line and the simplest class $\mathcal{G}$ is the set of affine functions. By abuse of notation, we use the term linear functions. Other classes of functions can be chosen and this selection is usually given by pre-existing knowledge of the phenomenon and/or examining the data.

Simple linear regression analysis always starts by plotting the data (x, y). This plot allows us to find out whether a linear model is appropriate.

1.3 Statistical Model

When fiting the data to a straight line, we implicitly assume that

$$Y = \beta_1 + \beta_2 X.$$

In the tree example, we assume a model where height depends linearly on circonference. We clearly know that not all the observations will lie on the straight line. We cannot realistically believe that height of an eucalyptus linearly depends its circonference. The observations depend on instrument precision as well as the operator, and we may find that for identical values of the variable X, we observe different values of Y. We thus assume that height depends linearly on circonference but that this relationship is affected by an "error". We assume in fact that the data can be modeled as follows:

$$Y = \beta_1 + \beta_2 X + \varepsilon. \tag{1.2}$$

The equation (1.2) is called a **linear regression model** and in the case at hand, a **simple linear regression model**. β_j, the "model parameters" (intercept and regression coefficient), are fixed but unknown and we want to estimate them. The quantity denoted ε or, the "error", is random and unknown.

In order to estimate the unknown model parameters, we measure a single explanatory or independent variable X and a response or dependent variable Y as part of a simple regression. Variable X is often taken to be fixed, unlike Y. We therefore measure n observations of variable X, denoted x_i, where i varies from 1 to n, and n values of the explanatory variable Y denoted y_i.

We assume that we have collected n data pairs (x_i, y_i) where y_i is the realization of random variable Y_i. By abuse of notation, we blur the distinction between random variable Y_i and its realization, observation y_i. We also write ε_i, referring to both the random variable and its realization. According to the model (1.2), we can write

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \qquad i = 1, \cdots, n$$

where

- x_i are known, fixed values;
- model parameters β_j, $j = 1, 2$ are unknown;
- ε_i are realizations of an unknown random variable;
- y_i are observations of a random variable.

1.4 Least Squares Estimators

Definition 1.1 (OLS estimators)
The estimators of β_1 and β_2, estimators $\hat{\beta}_1$ and $\hat{\beta}_2$, are referred to as ordinary least squares (OLS) estimators, and are obtained by minimizing the expression

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2 = \|Y - \beta_1 \mathbb{1} - \beta_2 X\|^2,$$

where $\mathbb{1}$ is a vector in $\mathbb{R}^n$ where all the coefficients are equal to 1. The estimators can also be written as follows:

$$(\hat{\beta}_1, \hat{\beta}_2) = \underset{(\beta_1, \beta_2) \in \mathbb{R} \times \mathbb{R}}{\text{argmin}} \; S(\beta_1, \beta_2).$$

The function $S(\beta_1, \beta_2)$ is strictly convex. If there is a singular point, then it corresponds to the unique minimum.Solving the partial derivatives to 0, we obtain

$$\left\{ \begin{array}{lcl} \dfrac{\partial S(\hat{\beta}_1, \hat{\beta}_2)}{\partial \beta_1} & = & -2 \displaystyle\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0, \\ \dfrac{\partial S(\hat{\beta}_1, \hat{\beta}_2)}{\partial \beta_2} & = & -2 \displaystyle\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0. \end{array} \right.$$

The first equation yields

$$\hat{\beta}_1 n + \hat{\beta}_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

and we have the estimator for the intercept

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \tag{1.3}$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$. The second equation yields

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

Replacing $\hat{\beta}_1$ by its expression (1.3) we get

$$\hat{\beta}_2 = \frac{\sum x_i y_i - \sum x_i \bar{y}}{\sum x_i^2 - \sum x_i \bar{x}},$$

Using that the sum $\sum (x_i - \bar{x})$ is zero, we get

$$\hat{\beta}_2 = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})(x_i - \bar{x})} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}. \tag{1.4}$$

To obtain this result, we assume that there are at least two different x_i. This hypothesis, denoted $\mathcal{H}_1$ is formulated as $x_i \neq x_j$, for at least two individuals. This ensures that the coefficient estimators $\hat{\beta}_1, \hat{\beta}_2$ are unique.
With $\hat{\beta}_1$ and $\hat{\beta}_2$, we can estimate the regression line using the formula

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X.$$

We plot the original points and the regression line. The regression line goes through the center of gravity $(\bar{x}, \bar{y})$ of the data as shown in equation(1.3).

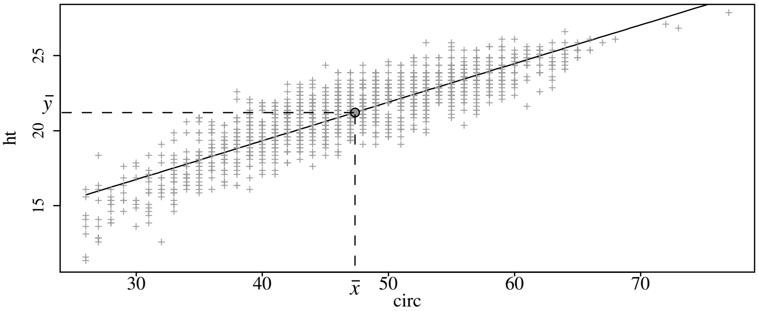


Figure 1.3 – Plot of the circumference/height data and the fitted line obtained.

We carried out an experiment and observed n values (x_i, y_i). From these n values, we obtained an estimate of β_1 and β_2. If we carry out another experiment, we would observe other data pairs (x_i, y_i) and would obtain another estimate of β_1 and β_2. Estimators depend on observations and therefore vary with the data. The true values of β_1 and β_2 on the other hand are unknown and do not vary.

Definition 1.2 (Estimation/prediction)

With the n values (x_i, y_i), we estimate $\hat{\beta}_1$ and $\hat{\beta}_2$.

- *if we use one x_i to calculate $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ we say that $\hat{y}_i$ is an estimated value.*
- *if we use a new x^* to calculate $\hat{y}^* = \hat{\beta}_1 + \hat{\beta}_2 x^*$ we say that $\hat{y}^*$ is an predicted value.*

In average, the estimation error (which will be defined) is smaller than the prediction error.

1.5 Basic properties

In general, statisticians seek to check that the estimators have specific properties such as:

- is the estimator $\hat{\beta}$ unbiased?
- is the estimator $\hat{\beta}$ of minimal variance among all the estimators of a defined class?

To ensure this, we formulate a second hypothesis, $\mathcal{H}_2$: the errors are centered, of the same variance (homoscedasticity) and are uncorrelated with each other. Using $\mathcal{H}_2$, it is possible to derive the statistical properties of the estimators.
$\mathcal{H}_2 : \mathbb{E}(\varepsilon_i) = 0$, for $i = 1, \cdots, n$ and $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$, where $\mathbb{E}(\varepsilon)$ is the expectation of ε, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j)$ is the covariance between ε_i and ε_j and $\delta_{ij} = 1$ when $i = j$ and $\delta_{ij} = 0$ when $i \neq j$.

Definition 1.3 (Estimator Bias)
The bias of an estimator $\hat{\beta}$ of a parameter β is calculated with $\mathbb{E}(\hat{\beta}) - \beta$.

Proposition 1.1 ($\hat{\boldsymbol{\beta}}$ unbiased)
$\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased estimators of β_1 and β_2, in other words $\mathbb{E}(\hat{\beta}_1) = \beta_1$ and $\mathbb{E}(\hat{\beta}_2) = \beta_2$.

Estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ being unbiased, we are going to focus on their variance. In order to show that these estimators are of minimum variance in their class, we first calculate their variance. We shall see this in the next proposition.

Proposition 1.2 (Variances of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$)
The variances and covariances of the estimators are equal to:

$$\begin{aligned}
\mathrm{V}(\hat{\beta}_2) &= \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \\
\mathrm{V}(\hat{\beta}_1) &= \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2} \\
\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) &= -\frac{\sigma^2 \bar{x}}{\sum(x_i - \bar{x})^2}.
\end{aligned}$$

This proposition makes it possible to assess the accuracy of the estimates using the variance. The smaller the variance, the more accurate the estimate. To obtain small variances, we need a small numerator and (or) a large denominator. The estimates therefore have small variances when:

- the variance σ^2 is small. This means that the variance of Y is small and the observations are close to the line of fit;
- the quantity $\sum(x_i - \bar{x})^2$ is large, the observations x_i must be spread about their mean;
- the quantity $\sum x_i^2$ must not be too large, the points must have a small mean in absolute value. Indeed, we have

$$\frac{\sum x_i^2}{\sum(x_i - \bar{x})^2} = \frac{\sum x_i^2 - n\bar{x}^2 + n\bar{x}^2}{\sum(x_i - \bar{x})^2} = 1 + \frac{n\bar{x}^2}{\sum(x_i - \bar{x})^2}.$$

The equation (1.3) shows that the line of least squares goes through the center of gravity $(\bar{x}, \bar{y})$. Suppose $\bar{x}$ positive, then if we increase the slope, the intercept will decrease and vice versa. We therefore find a negative sign for the covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$.

We conclude this section concerning the properties by the Gauss-Markov theorem which states that, among all linear unbiased estimators, the least squares estimator has the smallest variance.

Theorem 1.1 (Gauss-Markov)
Among all the linear unbiased estimators in Y, the estimators $\hat{\beta}_j$ have the smallest variances.

1.6 Residuals and Residual Variance

We have estimated β_1 and β_2. The variance σ^2 of ε_i is the last unknown parameter to be estimated. To do so, we use residuals: they are the estimates of the unknown errors ε_i.

Definition 1.4 (Residuals)
Residuals are defined as

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

where $\hat{y}_i$ is the model fitted value of y_i.

We have the following property:

Proposition 1.3
In a simple linear regression model, the sum of the residuals is zero.

We now propose to estimate σ^2 and construct an unbiased estimator $\hat{\sigma}^2$

Proposition 1.4 (Estimator of the error variance)
The statistic $\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2/(n-2)$ is an unbiased estimator of σ^2.

Chapter 2

Multiple Linear Regression

2.1 Introduction

Air quality forecasting is an important issue. Being able to anticipate it, should allow to adjust public policies in order to prevent possible illnesses. Air Breizh, while measuring ozone concentration, also measures meteorological variables which may have an influence on it. A program to collect data to characterize air pollution has been set up and it was measured at a given point, the concentration of ozone, temperature, cloudiness, wind speed and wind direction at noon. Some of this data are given below:

Individual	O3	T12	Vx	Ne12
1	63.6	13.4	9.35	7
2	89.6	15	5.4	4
3	79	7.9	19.3	8
4	81.2	13.1	12.6	7
5	88	14.1	-20.3	6

Table 2.1 – 5 daily data.

The variable Vx is a synthetic variable which represents wind. Wind is usually measured in degree (direction) and meter per second (speed). The variable created is the projection of wind on the east-west axis, and takes into account the direction and speed. The variable Ne12 represents cloud cover and T12 the temperature measured at 12 AM.

In order to analyses the relationship between temperature (T12), wind (Vx), cloud cover (Ne12) and ozone (O3), we are looking for a function f such as

$$\mathtt{O3}_i \approx f(\mathtt{T12}_i, \mathtt{Vx}_i, \mathtt{Ne12}_i).$$

In order to make the meaning of $\approx$ more precise, we need to define a positive criterion which qualifies the quality of the fit of the function f to the data called a loss function. Minimizing a loss requires knowledge of the space in which we minimize, hence the class of functions $\mathcal{G}$ in which we assume the true unknown function lies. The mathematical problem can be written as follows :

$$\underset{f\in\mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y_i - f(x_{i1}, \cdots, x_{ip})),$$

where n represents the number of observations to analyze and $l(.)$ is called the loss function ans we will consider again the quadratic loss. Concerning the choice of the class $\mathcal{G}$, we are going to use first the class of linear functions :
$\mathcal{G}\left\{f : f(x_1, \cdots, x_p) = \sum_{j=1}^{p} \beta_j x_j \quad \text{with} \quad \beta_j \in \mathbb{R}, j \in \{1, \ldots, p\}\right\}.$

2.2 Modeling

The multiple regression model generalizes the simple regression model when the number of explanatory variables is finite. We therefore suppose that the data collected is consistent with the following model :

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \cdots, n \tag{2.1}$$

where

- The x_{ij} are known and fixed numbers. The variable x_{i1} may be 1 for each i varying from 1 to n. In this case, β_1 represents the intercept. In statistics, this column of 1 is almost always used;
- the model parameters estimates β_j are unknown ;
- the ε_i are unknown random variables.

In matrix notation (2.1), we obtain the following definition:

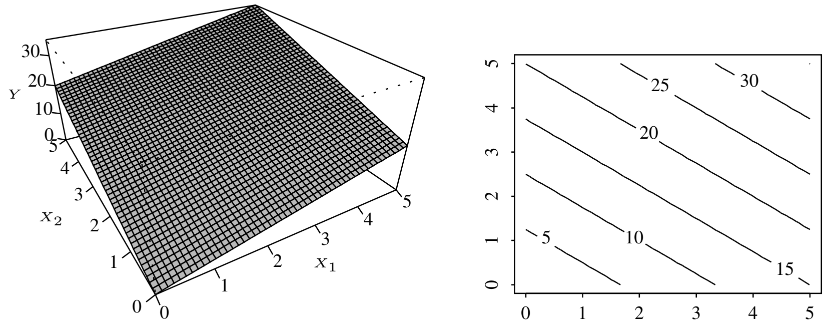
Definition 2.1 (Multiple regression model)
A linear regression model is defined by an equation of the form

$$Y_{n\times 1} = X_{n\times p}\ \beta_{p\times 1} + \varepsilon_{n\times 1}. \tag{2.2}$$

where :
• *Y is a random vector of dimension n ;*
• *X is a matrix of size $n \times p$ known, called design matrix, X is the concatenation of the p variables X_j : $X = (X_1|X_2|\ldots|X_p)$. We write the i^{th} row of the matrix X by the vector row $x_i' = (x_{i1}, \ldots, x_{ip})$;*
• *β is the p dimensional vector of the unknown parameters of the model ;*
• *ε is the centered n dimensional vector of the errors.*

We suppose that the matrix X is of full rank. This hypothesis is formulated as $\mathcal{H}_1$. Since, in general, the number of individuals n is larger than the number of explanatory variables p, the rank of the matrix X is less or equal than p.

The preceding description is equivalent to saying that the function relating Y to the explanatory variables X is an hyperplane as illustrated below (fig. 2.1).



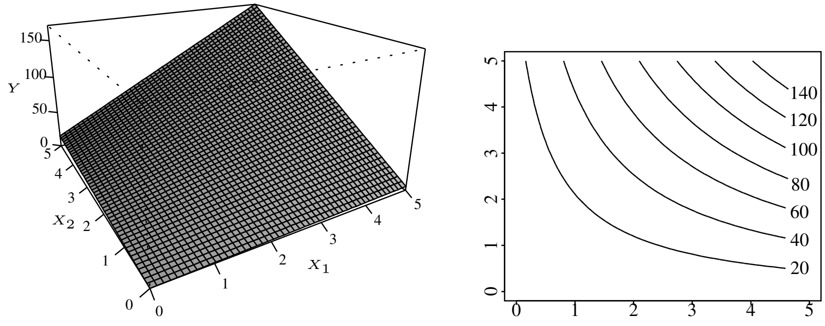
(a) Surface. **(b)** Contour.

Figure 2.1 – Geometrical representation of the relation $Y = 3X_1 + 4X_2$.

It is natural to assume that there exist in many problems interactions between the explanatory variables. In the ozone example, we may think that temperature and wind interact. In order to model this interaction, we generally write the model as a product between the explanatory variables which interact with each other. So, for two variables, we have the following model:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i, \qquad i = 1, \cdots, n.$$

Products carried out between two variables define interactions of order 2, between three variables, interactions of order 3, etc. From a geometrical point of view, this gives (fig. 2.2) :



(a) Surface. **(b)** Contour.

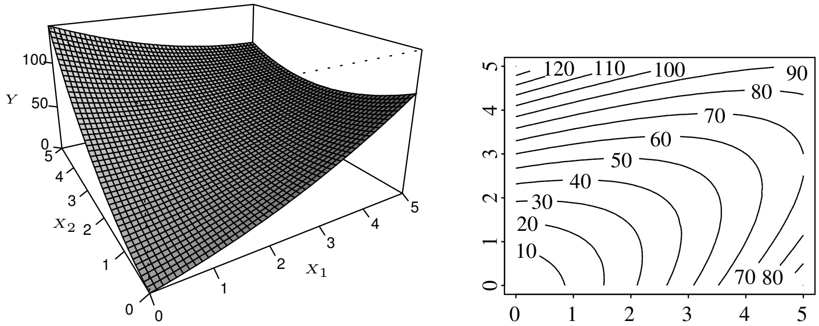
Figure 2.2 – Geometrical representation of the relationship $Y = X_1 + 3X_2 + 6X_1X_2$.

Nevertheless, this type of modeling falls perfectly within the framework of multiple regression. The interaction variables are the product of known variables and are therefore known. In the preceding example, the third explanatory variable X_3 is simply the product of X_1X_2 and we recover once again the model proposed in the preceding section.

Similarly, other extensions can be used such as polynomial regression. Using our preceding example with two explanatory variables X_1 and X_2, we propose the following polynomial model of degree 2 :

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2 + \varepsilon_i, \qquad i = 1, \cdots, n.$$

This model can be also be subsumed by the framework developed in the preceding section by setting $X_3 = X_1X_2$, $X_4 = X_1^2$ and $X_5 = X_2^2$. The hypersurface then looks like (fig. 2.3) :



(a) Surface. **(b)** Contour.

Figure 2.3 – Representation of the relationship $Y = 10X_1 + 8X_2 - 6X_1X_2 + 2X_1^2 + 4X_2^2$.

In conclusion, any **known and fixed transformation** of the explanatory variables (logarithm, exponential, product, etc.) can be used under the multiple regression model. So, the transformation of an explanatory variable X_1 by the function log for example becomes $\tilde{X}_1 = \log(X_1)$ and the model therefore remains a multiple regression model. On the other hand, a transformation such as $\exp\{-r(X_1 - k)\}$ which is a non linear function of two unknown parameters r and k does not fall within this framework. In fact, not knowing r and k, it is impossible to calculate $\exp\{-r(X_1 - k)\}$ and therefore to write it as $\tilde{X}_1$.

Thus a linear model does not necessarily mean that the relationship between the explanatory variables and the dependent variable is linear but rather that the *model is linear in the parameters β_j.*

2.3 Least squares estimators

Definition 2.2 (OLS estimator)
We call the least squares estimator (OLS) $\hat{\beta}$ of β the following value :

$$\hat{\beta} = \underset{\beta_1,\cdots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|^2.$$

Theorem 2.1 (Expression the OLS estimator)
If the hypothesis $\mathcal{H}_1$ is verified, the OLS estimator $\hat{\beta}$ of β is equal to

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

The following section is entirely devoted to this result.

2.3.1 Calculating and Interpreting $\hat{\beta}$

Consider the vectors in the variables space ($\mathbb{R}^n$). Thus, Y, a column vector, defines in $\mathbb{R}^n$ a vector $\overrightarrow{OY}$ with origin O and end point Y. This vector has coordinates $(y_1, \cdots, y_n)$. The design matrix X is formed of p column vectors. Each vector X_j defines in $\mathbb{R}^n$ a vector $\overrightarrow{OX_j}$ of origin O and end point X_j. This vector has for coordinates $(x_{1j}, \cdots, x_{nj})$. The p linearly independent vectors (hypothesis $\mathcal{H}_1$) span a subspace of $\mathbb{R}^n$, written from now $\Im(X)$, of dimension p.

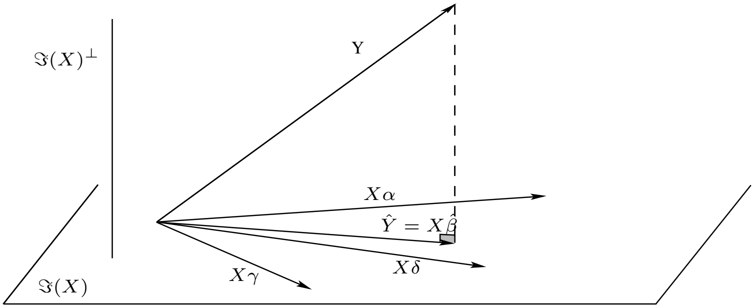


Figure 2.4 – Representation in the variables space.

This space $\Im(X)$, called image of X, is spanned by the columns of X. It is sometimes called the solution space. All vectors $\overrightarrow{v}$ of $\Im(X)$ are uniquely written as:

$$\overrightarrow{v} = \alpha_1 \overrightarrow{X_1} + \cdots + \alpha_p \overrightarrow{X_p} = X\alpha,$$

where $\alpha = [\alpha_1, \cdots, \alpha_p]'$. According to the model (2.2), the vector Y is the sum of an element of $\Im(X)$ and of an error, an element of $\mathbb{R}^n$, which has no reason to belong to $\Im(X)$. Minimizing $S(\beta)$ is equivalent to looking for an element of $\Im(X)$

which is closest to Y, in Euclidean norm. By definition, this unique element is called orthogonal projection of Y on $\Im(X)$. It is written $\hat{Y} = P_X Y$, where P_X is the orthogonal projection matrix on $\Im(X)$. This matrix is often written H and called "hat matrix" as it puts a "hat" on Y and we write h_{ij} the (i,j) $^{\text{th}}$ element of P_X. The element $\hat{Y}$ of $\Im(X)$ is given by $X\hat{\beta}$, where $\hat{\beta}$ is the OLS estimator of β. The space orthogonal to $\Im(X)$, written $\Im(X)^{\perp}$, is often called the residuals space. The vector $\hat{Y} = P_X Y$ contains the model fitted values of Y.

• Calculating $\hat{\beta}$ by projection :
Three options for calculating $\hat{\beta}$ are given.

- The first consists in knowing the analytic form of P_X. The orthogonal projection matrix on $\Im(X)$ is given by :

  $$P_X = X(X'X)^{-1}X'$$

  and, as $P_X Y = X\hat{\beta}$, we obtain $\hat{\beta} = (X'X)^{-1}X'Y$.

- The second method uses the fact that the vector Y of $\mathbb{R}^n$ can be uniquely decomposed into a part on $\Im(X)$ and a part on $\Im(X)^{\perp}$, we write this as :

  $$Y = P_X Y + (I - P_X)Y.$$

  The quantity $(I - P_X)Y$ being an element of $\Im(X)^{\perp}$ is orthogonal to any element v of $\Im(X)$. Recall that $\Im(X)$ is the space spanned by the columns of X. In other words all linear combinations of the variables $X_1, \cdots, X_p$ are elements of $\Im(X)$ or similarly, for all $\alpha \in \mathbb{R}^p$, we have $X\alpha \in \Im(X)$. The two vectors v and $(I - P_X)Y$ being orthogonal, the scalar product between these two quantities is zero, thus :

  $$\begin{aligned} \langle v, (I - P_X)Y \rangle &= 0 \quad \forall v \in \Im(X) \\ \langle X\alpha, (I - P_X)Y \rangle &= 0 \quad \forall \alpha \in \mathbb{R}^p \\ \alpha' X'(I - P_X)Y &= 0 \\ X'Y &= X'P_X Y \quad \text{with} \quad P_X Y = X\hat{\beta} \\ X'Y &= X'X\hat{\beta} \qquad X \text{ of full rank} \\ \hat{\beta} &= (X'X)^{-1}X'Y. \end{aligned}$$

  We recover $P_X = X(X'X)^{-1}X'$, the orthogonal projection matrix on the subspace spanned by the columns of X. The main properties of an orthogonal projector ($P_X' = P_X$ and $P_X^2 = P_X$) are verified.

- The last way of proceeding is to write that the vector $(I - P_X)Y$ is orthogonal to each of the columns of X :

$$\left\{\begin{array}{rcl}\langle X_1, Y - X\hat{\beta}\rangle & = & 0 \\ \vdots & & \\ \langle X_p, Y - X\hat{\beta}\rangle & = & 0\end{array}\right. \quad \Leftrightarrow \quad X'Y = X'X\hat{\beta}.$$

Given $P_X = X(X'X)^{-1}X'$ the orthogonal projection matrix on $\Im(X)$, the orthogonal projection matrix on $\Im(X)^{\perp}$ is $P_{X^{\perp}} = (I - P_X)$.

• Matrix operations
We can also derive the preceding result analytically by writing the function $S(\beta)$ to minimize as follows:

$$\begin{array}{rcl}S(\beta) & = & \|Y - X\beta\|^2 \\ & = & Y'Y + \beta'X'X\beta - Y'X\beta - \beta'X'Y \\ & = & Y'Y + \beta'X'X\beta - 2Y'X\beta.\end{array}$$

The necessary condition for finding an optimum is that the first derivative with respect to β is canceled. The derivative here is written as follows:

$$\frac{\partial S(\beta)}{\partial \beta} = -2X'Y + 2X'X\beta,$$

whence, if it exists, the optimum, written $\hat{\beta}$, verifies

$$-2X'Y + 2X'X\hat{\beta} = 0$$

in other words $\hat{\beta} = (X'X)^{-1}X'Y$.
To ensure that this point $\hat{\beta}$ is indeed a strict minimum, the second derivative must be a positive definite matrix. Here the second derivative is written

$$\frac{\partial^2 S(\beta)}{\partial \beta^2} = 2X'X,$$

and X is of full rank so $X'X$ is invertible and does not have any zero eigenvalue. The matrix $X'X$ thus is definite. Furthermore, for $\forall z \in \mathbb{R}^p$, we have

$$z'2X'Xz = 2\langle Xz, Xz\rangle = 2\|Xz\|^2 \geq 0$$

$(X'X)$ is thus positive definite and $\hat{\beta}$ is indeed a strict minimum.
We have seen that $\hat{Y}$ is the projection of Y on the subspace spanned by the columns of X. This projection exists and is unique even if the hypothesis $\mathcal{H}_1$ is not verified. The hypothesis $\mathcal{H}_1$ allows us in fact to obtain a unique $\hat{\beta}$. In this case, being interested in the coordinates of $\hat{\beta}$ is meaningful, and these coordinates are the

coordinates of $\hat{Y}$ in the coordinate system $X_1, \cdots, X_p$. This coordinate system does not have to be orthogonal and so $\hat{\beta}_j$ is not the coordinate of the projection of Y on X_j. We have

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

Calculating the projection of Y on X_j. yields

$$\begin{aligned} P_{X_j} Y &= P_{X_j} P_X Y \\ &= \hat{\beta}_1 P_{X_j} X_1 + \cdots + \hat{\beta}_p P_{X_j} X_p \\ &= \hat{\beta}_j X_j + \sum_{i \neq j} \hat{\beta}_i P_{X_j} X_i. \end{aligned}$$

This last quantity is different from $\hat{\beta}_j X_j$ except if X_j is orthogonal to all the other variables. When all the variables are orthogonal, it is clear that $(X'X)$ is a diagonal matrix

$$(X'X) = \operatorname{diag}(\|X_1\|^2, \cdots, \|X_p\|^2). \tag{2.3}$$

2.3.2 Some Statistical Properties

The statistician wish that the OLS estimators have nice properties in a statistical sense. Within our framework, this can be summarized into two parts: is the OLS estimator unbiased and of minimum variance in its class of estimators?
In order to answer these questions, we formulate a second hypothesis, noted $\mathcal{H}_2$, indicating that the errors are centered, of same variance (homoscedasticity) and uncorrelated with each other. We write this hypothesis as $\mathcal{H}_2 : \mathbb{E}(\varepsilon) = 0, \quad \Sigma_\varepsilon = \sigma^2 \mathbb{I}_n$, with $\mathbb{I}_n$ the identity matrix of order n. This hypothesis enables us to calculate

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}((X'X)^{-1}X'Y) = (X'X)^{-1}X'\mathbb{E}(Y) = (X'X)^{-1}X'X\beta = \beta.$$

The OLS estimator is therefore unbiased. Let us now calculate its variance

$$\mathrm{V}(\hat{\beta}) = \mathrm{V}((X'X)^{-1}X'Y) = (X'X)^{-1}X'\,\mathrm{V}(Y)X(X'X)^{-1} = \sigma^2(X'X)^{-1}.$$

Proposition 2.1 ($\hat{\beta}$ unbiased)
The OLS estimator $\hat{\beta}$ is an unbiased estimator of β and its variance is equal to $\mathrm{V}(\hat{\beta}) = \sigma^2(X'X)^{-1}$.

Remark 1
When the variables are orthogonal two by two, the components of $\hat{\beta}$ are not correlated with each other since the matrix $(X'X)$ is a diagonal (2.3) matrix.

The Gauss-Markov theorem, shows that among all the linear unbiased estimators of β, the OLS estimator has the smallest variance:

Theorem 2.2 (Gauss-Markov)
The OLS estimator is optimal among all the unbiased linear estimators of β.

2.3.3 Residuals and Residual Variance

The residuals are defined by the following relation:

$$\hat{\varepsilon} = Y - \hat{Y}.$$

It follows from the model, $Y = X\beta + \varepsilon$ and the fact that $X\beta \in \Im(X)$, that residuals can be re-expressed as:

$$\hat{\varepsilon} \;=\; Y - X\hat{\beta} = Y - X(X'X)^{-1}X'Y = (I - P_X)Y = P_{X^\perp}Y = P_{X^\perp}\varepsilon.$$

The residuals therefore belong to $\Im(X)^\perp$ and this space is also called the residuals space. The residuals are always orthogonal to $\hat{Y}$.

We have the following properties

Proposition 2.2 (Properties of $\hat{\varepsilon}$ and $\hat{Y}$)
Under the hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$, we have

$$\begin{aligned}
\mathbb{E}(\hat{\varepsilon}) &= P_{X^\perp}\mathbb{E}(\varepsilon) = 0\\
\mathrm{V}(\hat{\varepsilon}) &= \sigma^2 P_{X^\perp} I P'_{X^\perp} = \sigma^2 P_{X^\perp}\\
\mathbb{E}(\hat{Y}) &= X\mathbb{E}(\hat{\beta}) = X\beta\\
\mathrm{V}(\hat{Y}) &= \sigma^2 P_X\\
\mathrm{Cov}(\hat{\varepsilon}, \hat{Y}) &= 0.
\end{aligned}$$

The estimates of the residuals $\hat{\varepsilon}$ of ε possess the same expectation than ε. We will look later at residuals in more details.
We have discussed an estimator of σ^2 written $\hat{\sigma}^2$. A “natural” estimator of the residual variance is given by

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n}\|\hat{\varepsilon}\|^2.$$

But since $\|\hat{\varepsilon}\|^2$ is a scalar expression, then this scalar expression is equal to its trace, and using the properties of traces, we obtain

$$\mathbb{E}(\|\hat{\varepsilon}\|^2) = \mathbb{E}[\mathrm{tr}(\hat{\varepsilon}'\hat{\varepsilon})] = \mathbb{E}[\mathrm{tr}(\hat{\varepsilon}\hat{\varepsilon}')] = \mathrm{tr}(\mathbb{E}[\hat{\varepsilon}\hat{\varepsilon}']) = \mathrm{tr}(\sigma^2 P_{X^\perp}) = \sigma^2(n-p).$$

The last equality comes from the fact that the trace of a projector is equal to the dimension of the subspace onto which we project. This “natural” estimator is biased. In order to obtain an unbiased estimator we define

$$\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n-p} = \frac{\mathrm{SSR}}{n-p},$$

where SSR is the residuals sum of squares.

Proposition 2.3 ($\hat{\sigma}^2$ unbiased)
The statistic $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

From this estimator of the residual variance, we obtain an estimator of the variance of $\hat{\beta}$ in a straightforward way by replacing σ^2 by its estimator:

$$\hat{\sigma}^2_{\hat{\beta}} = \hat{\sigma}^2 (X'X)^{-1} = \frac{\text{SSR}}{n-p}(X'X)^{-1}.$$

We thus have an estimator of the standard error of the estimator $\hat{\beta}_j$ for each coefficient of the regression β_j

$$\hat{\sigma}_{\hat{\beta}_j} = \sqrt{\hat{\sigma}^2 [(X'X)^{-1}]_{jj}}.$$

2.3.4 Predictions

One of the aims of regression is to provide predictions for the response variable y when we have new values of x. Given a new value $x'_{n+1} = (x_{n+1,1}, \cdots, x_{n+1,p})$, we can predict y_{n+1}. But

$$y_{n+1} = x'_{n+1}\beta + \varepsilon_{n+1},$$

with $\mathbb{E}(\varepsilon_{n+1}) = 0$, $\mathrm{V}(\varepsilon_{n+1}) = \sigma^2$ and $\mathrm{Cov}(\varepsilon_{n+1}, \varepsilon_i) = 0$ for $i = 1, \cdots, n$. We can predict the corresponding value using our fitted model

$$\hat{y}^p_{n+1} = x'_{n+1}\hat{\beta}.$$

Two types of error are going to taint our prediction, the first due to the uncertainty surrounding ε_{n+1} and the other due to the estimates. Let us calculate the prediction error

$$\begin{aligned} \mathrm{V}\left(y_{n+1} - \hat{y}^p_{n+1}\right) &= \mathrm{V}(x'_{n+1}\beta + \varepsilon_{n+1} - x'_{n+1}\hat{\beta}) = \sigma^2 + x'_{n+1}V(\hat{\beta})x_{n+1} \\ &= \sigma^2(1 + x'_{n+1}(X'X)^{-1}x_{n+1}). \end{aligned}$$

We thus recover the uncertainty due to the errors σ^2.

Remark 2
Since the estimator $\hat{\beta}$ is an unbiased estimator of β and the expectation of ε is equal to zero, then the expectations of y_{n+1} and $\hat{y}^p_{n+1}$ are identical. The variance of the prediction error is written:

$$\mathrm{V}\left(y_{n+1} - \hat{y}^p_{n+1}\right) = \mathbb{E}\left[y_{n+1} - \hat{y}^p_{n+1} - \mathbb{E}(y_{n+1}) + \mathbb{E}(\hat{y}^p_{n+1})\right]^2 = \mathbb{E}(y_{n+1} - \hat{y}^p_{n+1})^2.$$

2.4 Geometrical Interpretation

The following equality is directly derived from the Pythagoras theorem :

$$\begin{aligned} \|Y\|^2 &= \|\hat{Y}\|^2 + \|\hat{\varepsilon}\|^2 \\ &= \|X\hat{\beta}\|^2 + \|Y - X\hat{\beta}\|^2. \end{aligned}$$

If a constant is part of the model, then we still have according to the Pythagoras theorem

$$\begin{aligned} \|Y - \bar{y}\mathbb{1}\|^2 &= \|\hat{Y} - \bar{y}\mathbb{1}\|^2 + \|\hat{\varepsilon}\|^2 \\ \text{total SS} &= \text{SS explained by the model} + \text{residual SS} \\ \text{SST} &= \text{SSE} + \text{SSR}\,. \end{aligned}$$

Definition 2.3 (R^2)
The (multiple) coefficient of determination R^2 *is defined by*

$$\mathrm{R}^2 = \frac{\|\hat{Y}\|^2}{\|Y\|^2} = \cos^2\theta_0$$

and if a constant is part of $\Im(X)$ *by*

$$\mathrm{R}^2 = \frac{\textit{V. explained by the model}}{\textit{Total variation}} = \frac{\|\hat{Y} - \bar{y}\mathbb{1}\|^2}{\|Y - \bar{y}\mathbb{1}\|^2} = \cos^2\theta.$$

R^2 can also be written as a function of the residuals:

$$\mathrm{R}^2 = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{y}\mathbb{1}\|^2}.$$

This coefficient measures the cosine squared of the angle between the vectors Y and $\hat{Y}$ at the origin or at $\bar{y}$ (seefig. 2.5). This latter angle is always larger than the first and R^2 calculated when the constant is part of $\Im(X)$ is therefore smaller than R^2 directly calculated.

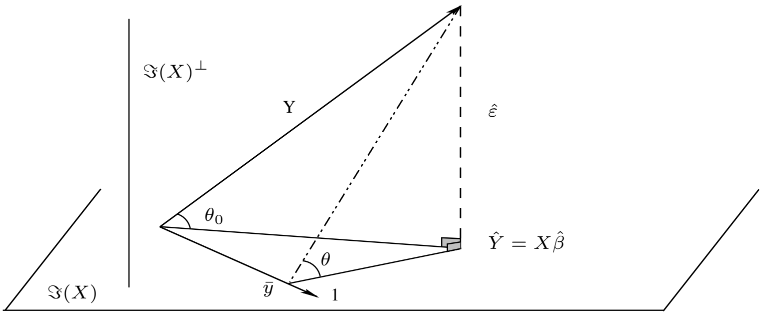


Figure 2.5 – Representation of the variables and geometrical interpretation of R^2.

However, this coefficient does not take into account the dimension of $\Im(X)$, adjusted R^2 is thus defined :

Definition 2.4 (Ajusted $\mathbf{R^2}$)
The adjusted coefficient of determination $\mathrm{R}_{\mathrm{a}}^2$ *is defined by*

$$\mathrm{R}_{\mathrm{a}}^2 = 1 - \frac{n}{n-p}\frac{\|\hat{\varepsilon}\|^2}{\|Y\|^2}$$

and, if the constant is part of $\Im(X)$, *by*

$$\mathrm{R}_{\mathrm{a}}^2 = 1 - \frac{n-1}{n-p}\frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{y}\mathbb{1}\|^2}.$$

The adjustment corresponds to the division of the norms squared by their respective degrees of freedom (or the sub-space dimension to which the vector belongs).

Chapter 3

Model Diagnostic

3.1 Introduction

Let us remind the context, we supposed that

$$Y_{n\times 1} = X_{n\times p}\ \beta_{p\times 1} + \varepsilon_{n\times 1},$$

under the hypotheses

- $\mathcal{H}_1 : rank(X) = p.$
- $\mathcal{H}_2 : \mathbb{E}(\varepsilon) = 0, \quad \mathrm{V}(\varepsilon) = \sigma^2 \mathbb{I}_n.$

The different stages of a regression analysis can be summarized as follows :

1. Modelisation: we assume the regression model $Y = X\beta + \varepsilon$;
2. Estimation: using the data, we estimate β;
3. Validation: which is the the objective of this chapter.

3.2 Residuals Analysis

Analyzing the residuals is an essential stage in linear regression. This stage is mainly based on graphical methods, and it is therefore difficult to have precise decision rules. The objective of this section is to present these graphical methods. Let us first recall the definitions of the different residuals.

3.2.1 Residuals

We estimate ε_i by $\hat{\varepsilon}_i = y_i - \hat{y}_i$. We have

Assumptions	Estimation
$\mathbb{E}(\varepsilon_i) = 0$	$\mathbb{E}(\hat{\varepsilon}_i) = 0$
$\mathrm{V}(\varepsilon) = \sigma^2 I_n$	$\mathrm{V}(\hat{\varepsilon}) = \sigma^2 (I - P_X)$

In order to have the same variance for each residual, we use the normalized residuals defined by

$$r_i = \frac{\hat{\varepsilon}_i}{\sigma\sqrt{1-h_{ii}}},$$

where h_{ij} is the (i,j) component of the matrix P_X. However σ is unknown, we replace σ by $\hat{\sigma}$, and obtain the standardized residuals

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}.$$

Their distribution is difficult to evaluate since the numerator and denominator are correlated. They possess the same unit variance, they are therefore useful to detect large residual values. Nevertheless, we prefer to use the studentized residuals by cross validation (CV)

$$t_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}},$$

where $\hat{\sigma}_{(i)}$ is the estimator of σ in the linear model without the observation i. If we suppose that $\varepsilon \sim \mathbb{N}(0, \sigma^2 \mathbb{I}_n)$ we can prove that

$$t_i^* = \frac{y_i - \hat{y}_i^p}{\hat{\sigma}_{(i)}\sqrt{1 + x_i'(X_{(i)}'X_{(i)})^{-1}x_i}} \sim \mathcal{T}(n-1-p),$$

where $X_{(i)}$ is the matrix X without its i^{th} row. We have $(n-1)$ observations and therefore loose a degree of freedom.

Theorem 3.1 (Distribution of studentized residuals par VC)
If the matrix X is of full rank, if the $\varepsilon_i \sim \mathcal{N}$~~are~~$(0, \sigma^2)$ ~~distributed~~ and if deleting a row i does not modify the rank of the matrix, then CV studentized residuals, noted t_i^*, have a Student distribution with $(n-p-1)$ degrees of freedom.

Remark 3
The calculations carried out in the proof clearly show the relationship between the prediction error $y_i - \hat{y}_i^p$ and the estimation error $y_i - \hat{y}_i$. We have

$$y_i - \hat{y}_i^p = \frac{y_i - \hat{y}_i}{1 - h_{ii}}. \tag{3.1}$$

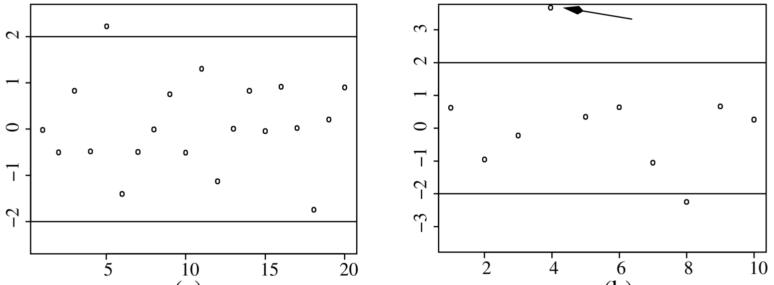
This result allows us to derive the prediction error without having to compute $\hat{\beta}_{(i)}$ for each observation i.

3.2.2 Fitting Individual Observation

To analyses the quality of fit of an observation, we only need to look at its corresponding residual. If this residual is very large, then the individual i is called an outlier. We therefore need to understand the reason why (error of measurement, individual belonging to a sub population) and eventually eliminate this point since it can modify the estimates. *An outlier* is an observation which is not well explained by the model and has a large residual. This notion is defined by :

Definition 3.1 (Outlier)
An outlier is a point (x_i', y_i) which has a large t_i^* value (compared to the critical value provided by the Student distribution) : $|t_i^*| > t_{n-p-1}(1 - \alpha/2)$.

In general outliers are detected by plotting the t_i^*. Detecting outliers does not only depend on the size of the residuals. Let us look at an randomly generated or simulated example .



(a) (b)

Figure 3.1 – Studentized residuals (a) and studentized residuals with an outlier as indicated by an arrow (b).

The plot (3.1.a) shows a satisfactory fit. None of the absolute values of the residuals are large. We note that in theory α % of the observations are outliers. We therefore look for residuals whose absolute values are clearly higher than $t_{n-p-1}(1 - \alpha/2)$. We therefore (3.1.b) only consider the individual indicated by an arrow as being an outlier.
Once identified and noted, it is important to understand why we have these outliers : is it a measurement error or a recording error? We recommend to delete these points from the analysis. But if you want to retain them then it is necessary to ensure that these values are not influential: the coefficients and the interpretations drawn from the model must not vary too much with or without these observations.

Conclusion

It is necessary to plot the residuals on the y axis and either $\hat{Y}$, or the number of the observation, or the time or any other potential factor of non independence on the x- axis. This type of plot allows us to identify the outliers, as well as verify the hypotheses concerning the structure of the vector of variance ε. Other graphs, such as those having the absolute value of the residuals on the y-axis allow us to look at the structure of the variance. The analysis of the residuals allows us to detect significant differences between the observed values and the fitted / predicted values. Nevertheless, it doesn't tell us anything about the variations of the parameter estimators due to omitting an observation and hence the robustness of the estimates. In the next section we consider measures adapted.

3.3 Analysis of the Projection Matrix

$P = P^2$ $\quad h_{ii} = \sum_{j=1}^{n} h_{ij}^2$

$h_{ii}(1-h_{ii}) = \sum_{j \neq i} h_{ij}^2$

The projection matrix

$$P_X = X(X'X)^{-1}X',$$

is the matrix which is used to calculate fitted values. More specifically,

$$\hat{Y} = P_X Y.$$

For row i, where h_{ij} is the element of the i^{th} row and j^{th} column of the projection matrix P_X, we write

$$\hat{y}_i \quad = \quad \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j.$$

this last expression allows us to measure the weight of the observation on its own fitted value *via* h_{ii}.

Definition 3.2 (Weight of an observation *i*)
The "weight" of an observation i on its own estimate is h_{ii}.

The orthogonal projection matrix P_X on the space spanned by the columns of X whose elements are the h_{ij} has in particular the following properties
if $h_{ii} = 1$ then $h_{ij} = 0$ for each $j \neq i$ ✓
and if $h_{ii} = 0$, then $h_{ij} = 0$ for each $j \neq i$. ✓
We have then the following extreme cases:

- if $h_{ii} = 1$, $\hat{y}_i$ is fully specified by y_i since $h_{ij} = 0$ for all j ;
- if $h_{ii} = 0$, y_i has no leverage?? on $\hat{y}_i$ (which is equal to zero).

We also know that $\text{tr}(P_X) = \sum h_{ii} = p$, therefore the average or mean of h_{ii} is p/n. Thus if h_{ii} is "large", y_i has a large impact on $\hat{y}_i$. Different authors have worked on this criterion and the following is their definition of "large" but we do have our own definition

Definition 3.3 (Leverage points)
A point i is a leverage point if the value h_{ii} of the projection matrix is much bigger than most of the values.

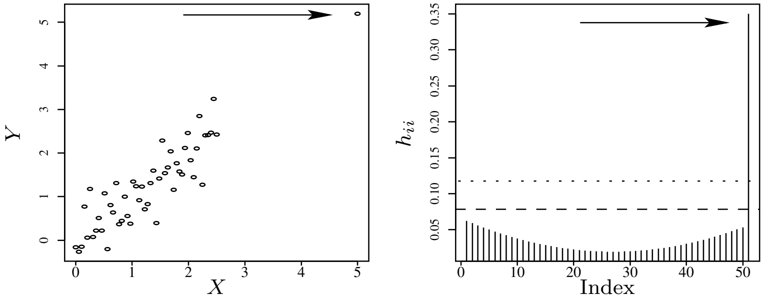


Figure 3.2 – Example of a leverage point, indicated by the arrow, for a simple regression model.

For a simple regression model whose points are represented on figure (3.2) the point indicated by an arrow is a leverage point. Its position on the x axis is different from the other points and its weight h_{ii} is much higher than the others values. A high value of h_{ii} indicates that the corresponding individual is far away from the center of gravity of points. Note that this is a leverage point but its it not an outlier as it lies on the extension of the regression line and therefore its residual is small.

Leverage points are therefore unusual points considering the explanatory variables. Here again it is useful to identify them and to understand why these points are different: measurement error, recording error, or they might belong to another population. Even if these points are not influential, i.e. without these points the estimates do not change, we can ask our self about the model validity. After some consideration, these values can be omitted or kept. In the first case, we take no risk at the edge of the domain, even if we delete a few points. In the second case, the model is implicitly extended to include these points.

The analysis of the residuals allows us to find these unusual values in terms of the values of the explanatory variable. The analysis of the projection matrix allows us to find those unusual individuals as a function of the explanatory variables (observations far from the mean/ overall average). Other criteria are combining both analyses such as the Cook distance.

Part II

Inference

Chapter 4

Inference in Regression

4.1 Introduction

Recall the following from the previous lessons:

$$Y_{n\times 1} = X_{n\times p}\ \beta_{p\times 1} + \varepsilon_{n\times 1},$$

under the assumptions

- $\mathcal{H}_1 : rank(X) = p.$
- $\mathcal{H}_2 : \mathbb{E}(\varepsilon) = 0, \quad \mathrm{V}(\varepsilon) = \sigma^2 \mathbb{I}_n.$

We are now assuming that the errors are normally distributed, therefore $\mathcal{H}_2$ becomes

- $\mathcal{H}_3 : \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n).$

We note that $\mathcal{H}_3$ implies $\mathcal{H}_2$. Furthermore, in the case of the Gaussian distribution, no correlation means independence. The hypothesis $\mathcal{H}_3$ is formulated as $\varepsilon_1, \cdots, \varepsilon_n$ are i.i.d. and $\mathcal{N}(0, \sigma^2)$ distributed. The Gaussian hypothesis allows us to calculate the likelihood and maximum likelihood estimators (MLE). This hypothesis also allows us to calculate the confidence regions and provide tests. This is the objective of the chapter

4.2 Maximum Likelihood Estimators

We start by calculating the likelihood of the sample. The likelihood is the density of the sample as a function of the parameters. Because the errors are independent, the observations are independent and the likelihood is written :

$$L(Y, \beta, \sigma^2) \quad = \quad \prod_{i=1}^{n} f_Y(y_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\big)^2\right].$$

We have therefore

$$L(Y,\beta,\sigma^2)=\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left[-\frac{1}{2\sigma^2}\|Y-X\beta\|^2\right],$$

which yields, taking the log

$$\mathcal{L}(Y,\beta,\sigma^2) \quad = \quad \log L(Y,\beta,\sigma^2)=-\frac{n}{2}\log\sigma^2-\frac{n}{2}\log 2\pi-\frac{1}{2\sigma^2}\|Y-X\beta\|^2.$$

In order to get the maximum, we differentiate according to β and σ^2 and obtain

$$\frac{\partial\mathcal{L}(Y,\beta,\sigma^2)}{\partial\beta} = \frac{1}{2\sigma^2}\frac{\partial}{\partial\beta}\left(\|Y-X\beta\|^2\right), \tag{4.1}$$

$$\frac{\partial\mathcal{L}(Y,\beta,\sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\|Y-X\beta\|^2. \tag{4.2}$$

From(4.1), we evidently have $\hat\beta_{ML}=\hat\beta$ and from (4.2) we have

$$\hat\sigma^2_{ML}=\frac{\|Y-X\hat\beta_{ML}\|^2}{n}$$

therefore $\hat\sigma^2_{ML}=(n-p)\hat\sigma^2/n$. The ML estimator is therefore biased as opposed to the estimator $\hat\sigma^2$ obtained using least squares. In order to verif y that we have a maximum, we need to look at the second derivatives.From now, $\hat\sigma^2$ will design the LS estimator. Under the additional hypothesis $\mathcal{H}_3$, the properties given previously are still valid (unbiased and minimum variance). However, we can show a set of new properties.

4.3 New Statistical Properties

Because of the Gaussian hypothesis, we can “improve” on the Gauss-Markov theorem. The optimally of the estimators is widened and we no longer only consider the unbiased linear estimators, but a larger class of unbiased estimators. Furthermore, the theorem now subsumes the estimator of σ^2. The proof of this proposition is given as an exercise.

Proposition 4.1 (Estimators distribution : variance known)
Under assumptions $\mathcal{H}_1$ and $\mathcal{H}_3$, we have
i) $\hat\beta$ is a Gaussian vector with expectation the vector β and variance matrix $\sigma^2(X'X)^{-1}$,
ii) $(n-p)\hat\sigma^2/\sigma^2$ is χ^2 distributed with $n-p$ df (χ^2_{n-p}),
iii) $\hat\beta$ and $\hat\sigma^2$ are independent.

Proof
i) $\hat\beta$ is a linear function of Gaussian variables and is therefore normally distributed fully specified by its variance and expectation as calculated in the preceding chapter.

ii)

$$\hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|^2}{n-p} = \frac{1}{n-p}\|\hat{\varepsilon}\|^2 = \frac{1}{n-p}\|P_{X^\perp}\varepsilon\|^2 = \frac{1}{n-p}\varepsilon' P_{X^\perp}\varepsilon.$$

But $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ and $P_{X^\perp}$ is the orthogonal projection matrix on $\Im(X)^\perp$, a space of dimension $n-p$. The result is derived using Cochran theorem.
iii) Note that $\hat{\beta}$ is a function of $P_X Y$ ($\hat{\beta} = (X'X)^{-1}X'P_X Y$) and $\hat{\sigma}^2$ is a function of $(I - P_X)Y$. The Gaussian vectors $\hat{Y}$ and $\hat{\varepsilon}$ have 0 covariance and are therefore independent. Any fixed function of $\hat{Y}$ remains independent of any fixed function of $\hat{\varepsilon}$, whence the result. □

We also derive a more general result in order to build the confidence regions.

Proposition 4.2 (Estimators distribution : estimated variance)
Under the assumptions $\mathcal{H}_1$ and $\mathcal{H}_3$, we have

i) for $j = 1, \cdots, p$, $T_j = \dfrac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(X'X)^{-1}]_{jj}}} \sim \mathcal{T}(n-p)$,

ii) given R a matrix of size $q \times p$ rank q ($q \leq p$) then

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

Proof
i) The variance of the estimator $\hat{\beta}_j$ is equal to $\sigma^2[X'X]^{-1}_{jj}$, we have then

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{[(X'X)^{-1}]_{jj}}} \sim \mathcal{N}(0, 1).$$

σ^2 is unknown and its estimate is $\hat{\sigma}^2$. The next part follows from using (ii) and (iii) from the preceding proposition.
ii) The rank of R is equal to $q \leq p$ according to our hypothesis, therefore the rank of $R(X'X)^{-1}R'$ is equal to q. $R\hat{\beta}$ is a Gaussian mean vector $R\beta$ of variance $\sigma^2 R(X'X)^{-1}R'$. We therefore have

$$\frac{1}{\sigma^2}(R\hat{\beta} - R\beta)' \left[R(X'X)^{-1}R'\right]^{-1} (R\hat{\beta} - R\beta) \sim \chi^2_q. \tag{4.3}$$

However σ^2 is unknown. In order to eliminate σ^2 from the equation (4.3), we divide the left hand side by $\hat{\sigma}^2/\sigma^2$. Recall that from (ii) we know that $\hat{\sigma}^2/\sigma^2$ is χ^2 distributed divided by its degree of freedom and that from (iii) $\hat{\sigma}^2/\sigma^2$ is independent from the left hand side of the equation (4.3). The rest follows from the definition of the Fisher distribution (ratio of two independent χ^2 distributions divided by their respective degrees of freedom). □

4.4 Confidence Intervals and Regions

Computer packages and some textbooks give CI for the parameters taken separately. Nevertheless, these CI do not take into account the dependence of the estimates. We can obtain simultaneous CI for several parameters. The theorem below provides all the types of CR : simple or simultaneous. This is the main theorem of the interval estimation (the proof is devoted to an exercise). The value $t_{n-p}(1-\alpha/2)$ denote the fractile of order $1-\alpha/2$ of a Student distribution with $n-p$ df, that $P(. > t_{n-p}(1-\alpha/2)) = \alpha/2$.

Theorem 4.1 (*CI* and *CR* of the parameters)
i) A $1-\alpha$ bilateral CI for β_j with $j=1,\cdots,p$ is given by

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

ii) A $1-\alpha$ equal tails CI for σ^2 is given by

$$\left[\frac{(n-p)\hat{\sigma}^2}{c_2},\quad \frac{(n-p)\hat{\sigma}^2}{c_1}\right] \quad \textit{where} \quad P(c_1 \leq \chi^2_{n-p} \leq c_2) = 1-\alpha.$$

iii) A $1-\alpha$ CR for q $(q \leq p)$ parameters β_j written $(\beta_{j_1},\cdots,\beta_{j_q})$ is given,

- *when σ is known, by*

$$CR_\alpha(R\beta) = \left\{R\beta \in \mathbb{R}^q, \frac{1}{\sigma^2}[R(\hat{\beta}-\beta)]'[R(X'X)^{-1}R']^{-1}[R(\hat{\beta}-\beta)] \leq \chi^2_q(1-\alpha)\right\}$$

- *when σ is unknown, by*

$$\begin{aligned} CR_\alpha(R\beta) = \{&R\beta \in \mathbb{R}^q, \\ &\frac{1}{q\hat{\sigma}^2}\,[R(\hat{\beta}-\beta)]'[R(X'X)^{-1}R']^{-1}[R(\hat{\beta}-\beta)] \leq f_{q,n-p}(1-\alpha)\Big\}, \end{aligned} \quad (4.4)$$

  where R is a matrix of $q \times p$ whose elements are all equal to 0 except for the $[R]_{ij_i}$ which are equal to 1.

The critical values c_1 and c_2 are the fractiles of the χ^2_q distribution and $f_{q,n-p}(1-\alpha)$ the fractile $(1-\alpha)$ of the Fisher distribution with $(q, n-p)$ df.

4.5 Prediction

Given $x'_{n+1} = (x_{n+1,1},\cdots,x_{n+1,p})$ a new value, we want to predict y_{n+1}. The model shows that

$$y_{n+1} = x'_{n+1}\beta + \varepsilon_{n+1},$$

with the ε_i which are i.i.d. and $\mathcal{N}(0,\sigma^2)$ distributed. From the n observations, we have an estimate for $\hat{\beta}$ and we predict y_{n+1}

$$\hat{y}^p_{n+1} = x'_{n+1}\hat{\beta}.$$

The expectation and variance of the prediction error $\varepsilon^p_{n+1} = y_{n+1} - \hat{y}^p_{n+1}$ are :

$$\begin{aligned}
\mathbb{E}(y_{n+1} - \hat{y}^p_{n+1}) &= 0\\
\mathrm{V}(\hat{y}^p_{n+1} - y_{n+1}) &= \mathrm{V}(x'_{n+1}(\hat{\beta}-\beta) - \varepsilon_{n+1})\\
&= x'_{n+1}\,\mathrm{V}(\hat{\beta}-\beta)x_{n+1} + \sigma^2\\
&= \sigma^2\left[x'_{n+1}(X'X)^{-1}x_{n+1} + 1\right].
\end{aligned}$$

We obtain the following la proposition.

Proposition 4.3 (*CI* de prévision)
A $(1-\alpha)$ CI for y_{n+1} is given by

$$\left[x'_{n+1}\hat{\beta} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{x'_{n+1}(X'X)^{-1}x_{n+1}+1}\right].$$

Proof
$\hat{\beta}$ is normally distributed and x_{n+1} is fixed therefore $\hat{y}^p_{n+1}$ is normally distributed. The random value y_{n+1} to predict is normally distributed $\mathcal{N}(x'_{n+1}\beta,\sigma^2)$ and is independent of the $y_1,\cdots,y_n$ by hypothesis $\mathcal{H}_3$.
We therefore have that y_{n+1} is independent of $\hat{y}^p_{n+1} = x'_{n+1}\hat{\beta}$ because $\hat{\beta}$ is a linear combination of $y_1,\cdots,y_n$. The prediction error $y_{n+1} - \hat{y}^p_{n+1}$ is normally distributed and its mean and variance have been calculated. We have thus

$$N = \frac{\hat{y}^p_{n+1} - y_{n+1}}{\sigma\sqrt{x'_{n+1}(X'X)^{-1}x_{n+1}+1}} \sim \mathcal{N}(0,1).$$

However σ is unknown and its estimate is given by $\hat{\sigma}$. We use the definition of the Student distribution: if N has a Standard Normal distribution, if D is χ^2 distributed on d df and if N and D are independent then the ratio of $N/\sqrt{D/d}$ is distributed according to a Student distribution on d df.
Proposition 4.2 shows that $D = \hat{\sigma}^2(n-p)/\sigma^2$ is χ^2 on $(n-p)$ degrees of freedom and that D is independent of $\hat{\beta}$. However $\hat{\sigma}^2$ uniquely depends on the $y_1,\cdots,y_n$ and is therefore independent of y_{n+1}. The same goes for D. The randomness of N comes from $\hat{\beta}$ and of y_{n+1}, we deduct from this that N and D are independent hence

$$\frac{N}{\sqrt{\frac{D}{d}}} = \frac{\hat{y}^p_{n+1} - y_{n+1}}{\hat{\sigma}\sqrt{x'_{n+1}(X'X)^{-1}x_{n+1}+1}} \sim \mathcal{T}(n-p), \tag{4.5}$$

the confidence interval is derived from this result. □

4.6 Hypothesis Testing

4.6.1 Introduction

Let us illustrate hypotheses testing on the ozone example. We have explained ozone by `T12`, `Vx` and `Ne12`. It seems reasonable to ask the following questions:
(a) is the value of `O3` influenced by `Vx`?
(b) is there a cloud cover effect?
(c) is the value of `O3` influenced by `Vx` or `T12`?

Recall that the model adopted is the following:

$$\texttt{O3} = \beta_1 + \beta_2\texttt{T12} + \beta_3\texttt{Vx} + \beta_4\texttt{Ne12} + \varepsilon.$$

We can make the preceding three questions explicit using hypothesis tests:
(a) corresponds to $\mathrm{H}_0 : \beta_3 = 0$, against $\mathrm{H}_1 : \beta_3 \neq 0$;
(b) corresponds to $\mathrm{H}_0 : \beta_4 = 0$, against $\mathrm{H}_1 : \beta_4 \neq 0$;
(c) corresponds to $\mathrm{H}_0 : \beta_2 = \beta_3 = 0$, against $\mathrm{H}_1 :\ \beta_2 \neq 0$ or $\beta_3 \neq 0$.

We note that all these cases come down to testing the null hypothesis for all the parameters together. In the c) case we talk about simultaneous null hypothesis for all the coefficients. This means that under the H_0 hypothesis some coefficients are equal to zero, and therefore the variables corresponding to these coefficients are not needed. This case corresponds by definition to comparing two models one nested inside the other (one being a special case of the other).
The design matrix without these variables is noted X_0 and the columns of X_0 span a subspace noted $\Im(X_0)$. In order to simplify the notation, we write $\Im(X_0) = \Im_0$ and $\Im(X) = \Im_X$. The test level is fixed to the standard α level.

4.6.2 Test between Nested Models

First recall the model and the hypotheses used :

$$Y = X\beta + \varepsilon \quad \text{where} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n),$$

This means that $\mathbb{E}(Y) \in \Im_X$ the space spanned by the columns of X.
To simplify the notation, lets assume that we want to test the hypothesis that the model last q coefficients with $q \leq p$ are equal to zero. The problem is thus written as:

$$\mathrm{H}_0 : \quad \beta_{p-q+1} = \cdots = \beta_p = 0 \quad \text{against} \quad \mathrm{H}_1 :\ \exists j \in \{p-q+1, \cdots, p\} : \beta_j \neq 0.$$

What does $\mathrm{H}_0 :\ \beta_{p-q+1} = \cdots = \beta_p = 0$ mean in terms of the model? If the last q coefficients are zero then the model becomes

$$Y = X_0\beta_0 + \varepsilon_0 \quad \text{where} \quad \varepsilon_0 \sim \mathcal{N}(0, \sigma^2 I),$$

where the matrix X_0 is made of the first $p - q$ columns of X. The columns of X_0 span a space noted $\Im_0$ of dimension $p_0 = p - q$. This subspace is clearly included in $\Im_X$. Under the null hypothesis H_0, the expectation of Y belongs to this subspace. Once the hypothesis tests are stated, we need a test statistic. We are going to adopt a rather intuitive geometric approach here.

Geometrical Approach

Consider the subspace noted $\Im_0$. We have written that under $\mathrm{H}_0 : \mathbb{E}(Y) \in \Im_0$. In this case, the least squares method consists in projecting Y not so much on $\Im_X$ (and obtaining $\hat{Y}$) but on $\Im_0$ and obtain $\hat{Y}_0$. Let us visualize these different projections with the following graph:

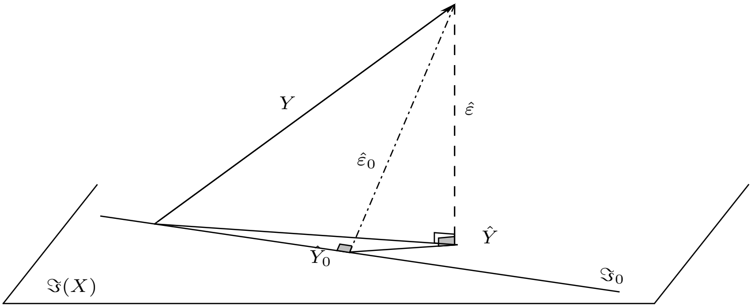
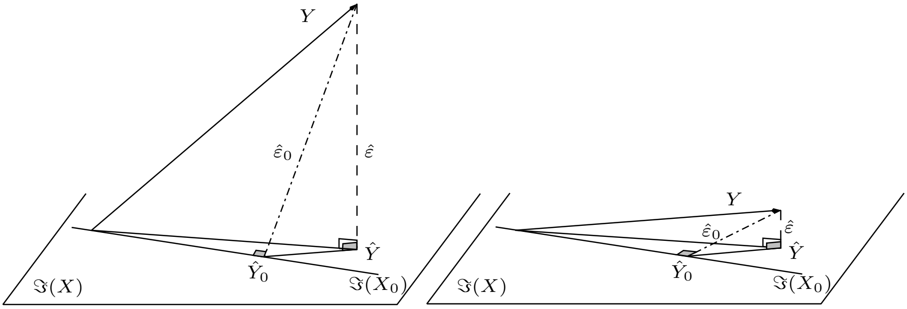


Figure 4.1 – Representation of the projections.

The intuitive idea of the test, and therefore of the choice to keep or reject H_0: if the projection of Y in $\Im_0$, noted $\hat{Y}_0$, is "close" to the projection of Y in $\Im_X$, noted $\hat{Y}$, then it seems logical to retain the null hypothesis. Equivalently, if the information provided by the two models is the "same", it is best to retain the smaller model (principle of parsimony). We must of course quantify the term "close". Normally, we would calculate the Euclidean distance between $\hat{Y}_0$ and $\hat{Y}$, or its square, $\|\hat{Y}_0 - \hat{Y}\|^2$. However, this distance varies according to the data and to the units used. For exemple compare these two different cases



To avoid this problem of scale we are going to "standardize" this distance by dividing it by the norm squared of the error $\hat{\varepsilon}$. The quantities $\hat{\varepsilon}$ and $\hat{Y}_0 - \hat{Y}$ do not belong to spaces of same dimensions, we thus divide each of the term by its respective degree of freedom. We thus have the following test statistic:

$$F = \frac{\|\hat{Y}_0 - \hat{Y}\|^2/q}{\|Y - \hat{Y}\|^2/(n-p)} = \frac{\|\hat{Y}_0 - \hat{Y}\|^2/(p-p_0)}{\|Y - \hat{Y}\|^2/(n-p)}.$$

In order to use this test statistic, we need to know its distribution at least under H_0. Note that this statistic is the ratio of two norms squared. We therefore need to decide on the distribution of the numerator and denominator and note their independence. We know that

$$\hat{Y}_0 - \hat{Y} = P_{X_0}Y - P_X Y,$$

or $\Im_0 \subset \Im_X$ therefore

$$\hat{Y}_0 - \hat{Y} = P_{X_0}P_X Y - P_X Y = (P_{X_0} - \mathbb{I}_n)P_X Y = -P_{X_0^\perp}P_X Y.$$

We deduce from this that $(\hat{Y}_0 - \hat{Y}) \in \Im_0^\perp \cap \Im_X$ and therefore that $(\hat{Y}_0 - \hat{Y}) \perp (Y - \hat{Y})$. The graph (4.1) allows us to visualize these notions of orthogonality. The random vectors $\hat{Y}_0 - \hat{Y}$ and $Y - \hat{Y}$ are elements of orthogonal spaces, therefore their covariance is zero. These two vectors are Gaussian, they are therefore independent and any fixed function of these 2 vectors remains independent, in particular, the norms of the numerator and denominator are independent.
Using the normal hypothesis $\mathcal{H}_3$ and applying the geometry Cochran theorem, we deduce from it that these two norms follow a χ^2 distribution.

$$\begin{aligned} \frac{1}{\sigma^2}\|P_{X^\perp}Y\|^2 &\sim \chi^2_{n-p}, \\ \frac{1}{\sigma^2}\|P_{X_0^\perp \cap X}Y\|^2 &\sim \chi^2_{p-p_0}\left(\frac{1}{\sigma^2}\|P_{X_0^\perp \cap X}X\beta\|^2\right), \end{aligned}$$

where the de-centering parameter $\|P_{X_0^\perp \cap X}X\beta\|^2/\sigma^2$ is zero under H_0 since in this case $X\beta \in \Im_0$. We can conclude with the following theorem.

Theorem 4.2 (Test between nested models)
Given a regression model with p variables $Y = X\beta + \varepsilon$ which satisfies $\mathcal{H}_1$ and $\mathcal{H}_3$. We want to test the validity of a reduced model (nested model) where one or more of the coefficients are zero. The design matrix without these variables is written X_0, the p_0 columns of X_0 span a subspace noted $\Im_0$ and the reduced model is $Y = X_0\beta_0 + \varepsilon_0$. We write the null hypothesis (reduced model) $\mathrm{H}_0 : \mathbb{E}(Y) \in \Im_0$ *and the alternative hypothesis (full model)* $\mathrm{H}_1 : \mathbb{E}(Y) \in \Im(X)$. *To test these two hypotheses, we use the test statistic F below which is distributed under* H_0 *as:*

$$F = \frac{\|\hat{Y}_0 - \hat{Y}\|^2/(p-p_0)}{\|Y - \hat{Y}\|^2/(n-p)} \sim \mathcal{F}_{p-p_0,n-p}.$$

Note also an often used and equivalent way of writing it

$$F = \frac{n-p}{p-p_0}\frac{\mathrm{SSR}_0 - \mathrm{SSR}}{\mathrm{SSR}} \sim \mathcal{F}_{p-p_0,n-p}.$$

The hypothesis H_0 *is rejected in favor of* H_1 *if the observed statistic* F *exceeds the critical value* $f_{p-p_0,n-p}(1-\alpha)$, α *being the test level.*

Proof
How the test statistic F is obtained comes from the construction which precedes the theorem. Recall that if $N \sim \chi^2$ on n df and $D \sim \chi^2$ on p df and if N and D are independent then

$$\frac{N}{D}\frac{d}{n} \sim \mathcal{F}_{n,p}.$$

or equivalently is obtained using the SSR notation by noticing that

$$\begin{aligned} \|Y - \hat{Y}_0\|^2 &= \|Y - P_X Y + P_X Y - P_{X_0} Y\|^2 \\ &= \|P_{X^\perp} Y + (\mathbb{I}_n - P_{X_0}) P_X Y\|^2 \\ &= \|P_{X^\perp} Y\|^2 + \|P_{X_0^\perp \cap X} Y\|^2 \\ &= \|Y - \hat{Y}\|^2 + \|\hat{Y} - \hat{Y}_0\|^2. \end{aligned}$$

This geometric approach appears without connection with the classic tests statistics, but we can show that the test F is simply a ratio test of maximum likelihoods.

Student Test for a Coefficient β_j

We want to test $\mathrm{H}_0 : \beta_j = 0$ against $\mathrm{H}_1 : \beta_j \neq 0$ (two tailed test for β_j). According to theorem 4.2, the test statistic is

$$F = \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\hat{\sigma}^2}.$$

We reject H_0 if the observed statistic F, noted $F(w)$, is such that

$$F(\omega) > f_{1,n-p}(1-\alpha),$$

since F has a Fisher distribution on 1 and $(n-p)$ df.
This test is equivalent to the T-test on $(n-p)$ df which allows us to test $\mathrm{H}_0 : \beta_j = 0$ against $\mathrm{H}_1 : \beta_j \neq 0$ with the same test statistic

$$T = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$$

which is under H_0 Student distributed on $(n-p)$ df. We reject H_0 if the observed T statistic, noted $T(w)$, is such that

$$|T(\omega)| > t_{n-p}(1-\alpha/2).$$

This is the form of the test we encounter in linear regression computer software.

Fisher Test of Goodness of fit

If *a priori* knowledge of the phenomenon studied shows that a constant term exists in the regression, then to test the influence on the response of the regressor terms which are not constant in the regression, we test whether $\mathbb{E}(Y) = \mu$ belongs to the diagonal $\Im_0(X) = \Delta$ of $\mathbb{R}^n$. We therefore test the overall model validity. This is equivalent to saying all the coefficients are assumed to be zero, except the constant. This test is called the goodness of fit test. In this case, $\hat{Y}_0 = \bar{Y}\mathbb{1}$ and we have the following test statistic :

$$\frac{\|P_{\Im_X}Y - P_{\Im_0}Y\|^2/(p-1)}{\|Y - P_{\Im_X}Y\|^2/(n-p)} = \frac{\|P_{\Im_X}Y - \bar{Y}\mathbb{1}\|^2/(p-1)}{\|Y - P_{\Im_X}Y\|^2/(n-p)} \sim \mathcal{F}_{p-1,n-p}.$$

If we write the test statistic using R^2, we obtain the ratio

$$F = \frac{\mathrm{R}^2}{1-\mathrm{R}^2}\frac{n-p}{p-1}.$$

This test is called R^2 test in some statistical packages.

Part III

Dimension reduction

Chapter 5

Variable Selection

5.1 Introduction

In the preceding lessons, we have assumed that the model proposed

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

was correct and that all the explanatory variables $(X_1, \cdots, X_p)$ are useful. Nevertheless, in many statistical analyses, we have at hand a set of explanatory variables to explain one variable and nothing tells us whether the variables play a role in the modelling. The user therefore has a set of potential explanatory variables or candidate variables available.
We have p variables $(p < n)$ available and we assume, that the constant (variable $\mathbb{1}$) is among the candidate variables, in other words, that one of the X_i is equivalent to $\mathbb{1}$. If one wants to keep this special variable in his/her model, he/she needs therefore to analyse (2^{p-1}) potential model. Amongst these variables, we suppose that there could exist variables which have been transformed (features engineering) such as polynomial.

The main goal of this introduction is to understand the figure 7.2 below.

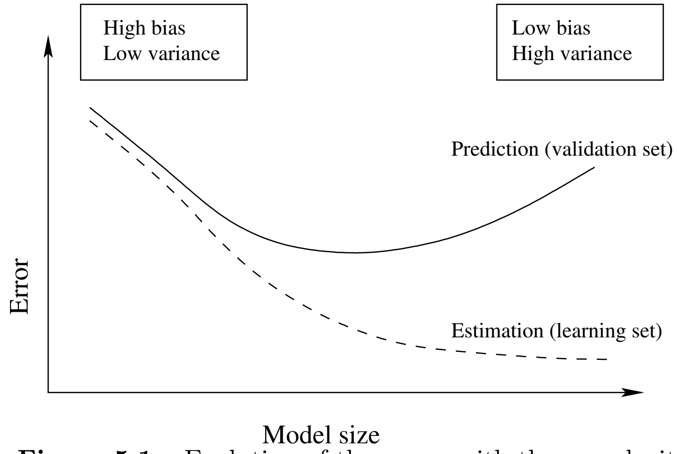


Figure 5.1 – Evolution of the errors with the complexity.

- Let us explain the dashed line: as the number of explanatory variables increases (ie model size or complexity), the model will adapt to the data and the estimation error (the error done on the individual which have been used in estimation) decreases and could be 0 in the interpolation cases. This easily follows from the fact that the estimation error is defined by $\|Y-\hat{Y}\|^2 = \|P_{X^\perp}Y\|^2$ and when we add variables $dim(\Im(X))$ increases and thus $\|P_{X^\perp}Y\|^2$ decreases to 0 and can reach it when there is enough variables or transformed variables.

- On the opposite the solid line shows the prediction error (the error done on new individual, $\|Y^* - X^*\hat{\beta}\|^2$) and have a convex form: high with a few variables (the model is too simple), decreases with reasonable number of "suitable" variables, and increases again with a high number of variables: as there is too many variables the estimation of $\hat{\beta}$ is less precise and the predictions get worse.

In conclusion to the analysis of figure 7.2 we need to find a trade-off between having a lot of variables (high complexity model difficult to estimate, high variability but a small estimation error since the model *overfits* the data) and too few variables (small complexity, model easy to estimate, low variability and high estimation error as the model is too simple, it *underfits*). This last sentence is summarised by 5.2.

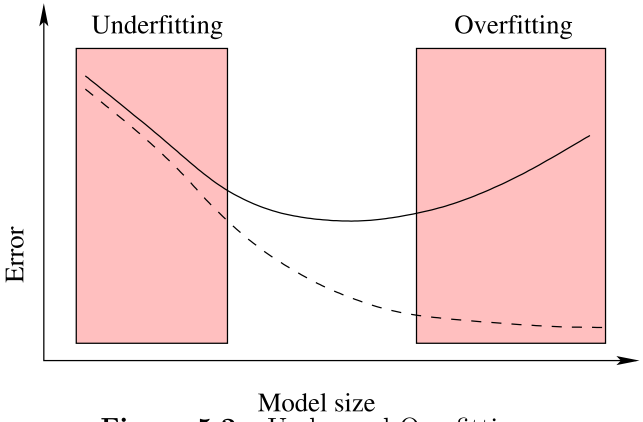


Figure 5.2 – Under and Overfitting.

To have an idea of the prediction error in a suitable way, we need to have a set of validation (a dataset which was not be used for estimation). If this set does not exist, it is necessary to create it and therefore to reduce the training data. Unfortunately there are numerous cases where the dataset is (very) small and cannot be split: the model/variable selection have to be conducted on the same dataset preventing us to estimate easily the prediction error.
Let us summarise the 3 possibilities to estimate the prediction error:

- We can split the data in estimation/prediction sets and on the prediction set (denoted with *) we calculate $\|Y^* - X^*\hat{\beta}\|^2$ (or another criterion). This is the best choice when n is big enough,

- We can split the data in a cross validation way (coming soon in the next sequence named ridge regression). This can be seen as an intermediate possibilities between the two others;

- Without splitting, use the "classical way" to choose variables. This can be seen as a last choice strategy.

In the sequel we will focus en the last choice, the "classical variable selection". Before presenting the different criteria and procedures used in this approach, we think that it is important to understand the consequences of a bad choice among the set of selected variables, supposing that this set exists.

5.2 Notations

The notation used in this chapter are the following:

- X is the matrix composed of all the explanatory variables (of size $n \times p$),

- ξ is a subset (of indices/subscripts) of $\{1, 2, \dots, p\}$, its cardinal is denoted by $|\xi|$, and its complementary (with respect of $\{1, 2, \dots, p\}$) is denoted $\bar{\xi}$
- X_ξ is a sub-matrix of X whose columns correspond to the subscripts contained in ξ,
- in the model ξ selecting $|\xi|$ variables, the corresponding parameters are written β_ξ,
- the coordinates corresponding to ξ in the vector $\hat{\beta}$ are $[\hat{\beta}]_\xi$. In general, $[\hat{\beta}]_\xi \neq \hat{\beta}_\xi$ except if $\Im(X_\xi) \perp \Im(X_{\bar{\xi}})$,
- with a new observation $x^{\star\prime} = [x_\xi^{\star\prime}, x_{\bar{\xi}}^{\star\prime}]$, we have the following predictions:

$$\hat{y}^p = x^{\star\prime}\hat{\beta} \qquad\qquad \hat{y}_\xi^p = x_\xi^{\star\prime}\hat{\beta}_\xi.$$

5.3 Incorrect Variable Selection: consequences

The objective of this section is to understand the consequences of an incorrect choice of explanatory variables. By incorrect, we mean either selecting too few, or selecting the right number but not the correct ones, or selecting too many. We obviously find all the ideas shown in section 5.1. We first analyse a simple example and then generalise the results. The example treated here is the following : Assume that we have three potential explanatory variables X_1, X_2 and X_3 and that the true model is

$$Y \quad = \quad \beta_1 X_1 + \beta_2 X_2 + \varepsilon = X_{12}\beta_{12} + \varepsilon.$$

One of the variables is therefore not useful but this fact is unknown to the person carrying out the regression. We can therefore analyse 7 different models, 3 models with one variable; 3 with two variables and 1 with the three variables. We only make the calculations explicit when $\xi = \{1\}$. We obtain therefore as estimators :

$$\begin{aligned} \hat{\beta}_1 &= (X_1'X_1)^{-1}X_1'Y \\ \hat{Y}_1 &= P_{X_1}Y \\ \hat{\sigma}_1^2 &= \|P_{X_1^\perp}Y\|^2/(n-1). \end{aligned}$$

5.3.1 Estimators Bias

We analyse the bias of these estimators using the true model $\mathbb{E}Y = \beta_1 X_1 + \beta_2 X_2 = X_{12}\beta_{12}$.

$$\begin{aligned} \mathbb{E}\hat{\beta}_1 &= (X_1'X_1)^{-1}X_1'\mathbb{E}Y = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 \\ \mathbb{E}\hat{Y}_1 &= X_1\beta_1 + P_{X_1}X_2\beta_2. \end{aligned}$$

The bias is therefore:

$$\begin{aligned} B(\hat{\beta}_1) &= \mathbb{E}(\hat{\beta}_1) - \beta_1 = (X_1'X_1)^{-1}X_1'X_2\beta_2 \\ B(\hat{Y}_1) &= \mathbb{E}(\hat{Y}_1) - \mathbb{E}(Y) = P_{X_1}X_2\beta_2 - X_2\beta_2 = -P_{X_1^\perp}X_2\beta_2. \end{aligned}$$

The orthogonal projection matrix $P_{X_1^\perp}$ is not random (the choice of X_1 is not made as a function of the data), and since the trace of a projector is the dimension of its image space, we have

$$\begin{aligned} \mathbb{E}\hat{\sigma}_1^2 &= \frac{1}{n-1}\mathbb{E}\,\mathrm{tr}(Y'P_{X_1^\perp}Y) = \frac{1}{n-1}\,\mathrm{tr}(P_{X_1^\perp}\mathbb{E}(YY')) \\ &= \frac{1}{n-1}\,\mathrm{tr}(P_{X_1^\perp}(V(Y) + \mathbb{E}(Y)\mathbb{E}(Y)')) = \sigma^2 + \frac{1}{n-1}\beta_{12}'X_{12}'P_{X_1^\perp}X_{12}\beta_{12} \\ &= \sigma^2 + \frac{1}{n-1}\beta_2^2\|P_{X_1^\perp}X_2\|^2. \end{aligned}$$

The bias is equivalent to :

$$B(\hat{\sigma}_1^2) = \frac{1}{n-1}\beta_2^2\|P_{X_1^\perp}X_2\|^2.$$

In carrying out the calculations for the 7 possible models, we have the table 5.1.

model	estimations	properties
$Y_1 = X_1\beta_1 + \varepsilon$	$\hat{Y}_1 = X_1\hat{\beta}_1$	$B(\hat{Y}_1) = -P_{X_1^\perp}X_2\beta_2$
	$\hat{\sigma}_1^2 = \frac{\|P_{X_1^\perp}Y\|^2}{n-1}$	$B(\hat{\sigma}_1^2) = \frac{1}{n-1}\beta_2^2\|P_{X_1^\perp}X_2\|^2$
$Y = X_2\beta_2 + \varepsilon$	$\hat{Y}_2 = X_2\hat{\beta}_2$	$B(\hat{Y}_2) = -P_{X_2^\perp}X_1\beta_1$
	$\hat{\sigma}_2^2 = \frac{\|P_{X_2^\perp}Y\|^2}{n-1}$	$B(\hat{\sigma}_2^2) = \frac{1}{n-1}\beta_1^2\|P_{X_2^\perp}X_1\|^2$
$Y = X_3\beta_3 + \varepsilon$	$\hat{Y}_3 = X_3\hat{\beta}_3$	$B(\hat{Y}_3) = -P_{X_3^\perp}X_{12}\beta_{12}$
	$\hat{\sigma}_3^2 = \frac{\|P_{X_3^\perp}Y\|^2}{n-1}$	$B(\hat{\sigma}_3^2) = \frac{1}{n-1}\beta_{12}'X_{12}'P_{X_{12}^\perp}X_{12}\beta_{12}$
$Y = X_{12}\beta_{12} + \varepsilon$	$\hat{Y}_{12} = X_{12}\beta_{12}$	$B(\hat{Y}_{12}) = 0$
	$\hat{\sigma}_{12}^2 = \frac{\|P_{X_{12}^\perp}Y\|^2}{n-2}$	$B(\hat{\sigma}_{12}^2) = 0$
$Y = X_{13}\beta_{13} + \varepsilon$	$\hat{Y}_{13} = X_{13}\hat{\beta}_{13}$	$B(\hat{Y}_{13}) = -P_{X_{13}^\perp}X_{12}\beta_{12}$
	$\hat{\sigma}_{13}^2 = \frac{\|P_{X_{13}^\perp}Y\|^2}{n-2}$	$B(\hat{\sigma}_{13}^2) = \frac{1}{n-2}\beta_{12}'X_{12}'P_{X_{13}^\perp}X_{12}\beta_{12}$
$Y = X_{23}\beta_{23} + \varepsilon$	$\hat{Y}_{23} = X_{23}\hat{\beta}_{23}$	$B(\hat{Y}_{23}) = -P_{X_{23}^\perp}X_{12}\beta_{12}$
	$\hat{\sigma}_{23}^2 = \frac{\|P_{X_{23}^\perp}Y\|^2}{n-2}$	$B(\hat{\sigma}_{23}^2) = \frac{1}{n-2}\beta_{12}'X_{12}'P_{X_{23}^\perp}X_{12}\beta_{12}$
$Y = X_{123}\beta_{123} + \varepsilon$	$\hat{Y}_{123} = X_{123}\hat{\beta}_{123}$	$B(\hat{Y}_{123}) = 0$
	$\hat{\sigma}_{123}^2 = \frac{\|P_{X_{123}^\perp}Y\|^2}{n-3}$	$B(\hat{\sigma}_{123}^2) = 0$

Table 5.1 – Bias of the different estimators.

We note then that in the models which are “too small” (here with 1 variable), in other words which have fewer variables than the “correct” model, the estimators obtained are biased. On the other hand, when the models are “too big” (here with 3 variables), the estimators are not biased.

Proposition 5.1

1. *$\hat{\beta}_\xi$ and $\hat{Y}_\xi$ are in general biased.*

2. *$\hat{\sigma}^2_\xi$ is in general positively biased, in other words, on average, the expectation of $\hat{\sigma}^2_\xi$ is equivalent to σ^2 plus a positive quantity.*

Bias estimation is difficult because $x'\beta$ is unknown. We next analyse the estimator variance in order to show that the bias and the variance evolve in the opposite way (see figure 7.2, p.78).

5.3.2 Estimators Variance

The dimension of the estimators vary with the size of the model. Nevertheless, using the formula for inverting by block, we can show that the estimators of the common components have smaller variances in the smaller model.

$$\mathrm{V}(\hat{\beta}_1) \leq \mathrm{V}([\hat{\beta}_{12}]_1) \leq \mathrm{V}([\hat{\beta}_{123}]_1).$$

where

$$\begin{array}{lrcl}
Y = X_1\beta_1 + \varepsilon & V(\hat{\beta}_1) & = & (X_1'X_1)^{-1}\sigma^2 \\
Y = X_{12}\beta_{12} + \varepsilon & V(\hat{\beta}_{12}) & = & \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1}\sigma^2 \\
Y = X_{123}\beta_{123} + \varepsilon & V(\hat{\beta}_{123}) & = & \begin{pmatrix} X_1'X_1 & X_1'X_2 & X_1'X_3 \\ X_2'X_1 & X_2'X_2 & X_2'X_3 \\ X_3'X_1 & X_3'X_2 & X_3'X_3 \end{pmatrix}^{-1}\sigma^2.
\end{array}$$

If we work with fitted values, we have the same phenomenon :

$$\begin{array}{lrcl}
Y = X_1\beta_1 + \varepsilon & V(\hat{Y}_1) & = & P_{X_1}\sigma^2 \\
Y = X_{12}\beta_{12} + \varepsilon & V(\hat{Y}_{12}) & = & P_{X_{12}}\sigma^2 = P_{X_1}\sigma^2 + P_{X_2 \cap X_1^\perp}\sigma^2 \\
Y = X_{123}\beta_{123} + \varepsilon & V(\hat{Y}_{123}) & = & P_{X_{123}}\sigma^2 = P_{X_1}\sigma^2 + P_{X_{23} \cap X_1^\perp}\sigma^2.
\end{array}$$

We can express this as a general result

Proposition 5.2

1. $\mathrm{V}([\hat{\beta}]_\xi) - \mathrm{V}(\hat{\beta}_\xi)$ *is a positive semi definite matrix, which means that the components common to the two models are better estimated (vary less) in the smaller model.*

2. *The variance of the fitted values in the smaller model is smaller than that of the fitted values in the larger model* $\mathrm{V}(\hat{Y}) \geq \mathrm{V}(\hat{Y}_\xi)$.

If the criterion of model choice is the variance, the user chooses models which have fewer parameters to estimate! (see figure 7.2, p.78) In general, it is best to obtain a model which provides a good estimate of the mean (small bias) and has a small variance. We have seen that a simple way to meet the first objective is to retain all the variables that are available while the second is met by eliminating many variables. The mean square error (MSE) helps us to meet these two objectives.

5.3.3 Mean Squared Error

The mean square error (MSE) of an estimator $\hat{\theta}$ of θ of dimension p is

$$\begin{aligned}\mathrm{EQM}(\hat{\theta}) &= \mathbb{E}((\theta-\hat{\theta})(\theta-\hat{\theta})') \\ &= \mathbb{E}(\theta-\hat{\theta})\mathbb{E}(\theta-\hat{\theta})' + V(\hat{\theta}),\end{aligned}$$

in other words the bias “squared” plus the variance. A biased estimator can be better than a unbiased one if its variance is smaller. We are going to use the MSE as a comparison measure. We can compare either estimators $\hat{\beta}_\xi \in \mathbb{R}^{|\xi|}$, fitted values $x'_\xi\hat{\beta}_\xi \in \mathbb{R}$, where x'_ξ corresponds to a row of the matrix X_ξ, or predicted values $x^{\star\prime}_\xi\hat{\beta}_\xi \in \mathbb{R}$, where $x^\star_\xi \in \mathbb{R}^{|\xi|}$ is a new observation. It is standard to treat the choice of variables *via* the analysis of the fitted value or predicted value and not *via* the estimators $\hat{\beta}_\xi$ whose dimensions vary, namely $|\xi|$. The following definitions introduces MSE and the prediction MSE.

Definition 5.1 (MSE)
We consider the regression model $Y = X\beta+\varepsilon$ where β, the unknown model parameter, can have null coordinates. Given $x \in \mathbb{R}^p$ an observation column vector, we have $x_\xi \in \mathbb{R}^{|\xi|}$ and $\hat{\beta}_\xi$ the least squares estimator obtained with these $|\xi|$ variables. The mean square error (MSE) is defined by

$$\mathrm{EQM}(\hat{y}_\xi) = \mathbb{E}((x'_\xi\hat{\beta}_\xi - x'\beta)^2) = \mathrm{V}(x'_\xi\hat{\beta}_\xi) + B^2(x'_\xi\hat{\beta}_\xi),$$

where $B(x'_\xi\hat{\beta}_\xi) = \mathbb{E}(x'_\xi\hat{\beta}_\xi) - x'\beta$ is the bias of $x'_\xi\hat{\beta}_\xi$.
If we have n observations x_ξ grouped in a matrix X_ξ and the least squares estimator $\hat{\beta}_\xi$ obtained using these $|\xi|$ variables, we define the trace of the MSE by

$$\mathrm{tr}(\mathrm{EQM}(\hat{Y}_\xi)) = \mathrm{tr}(\mathrm{V}(X_\xi\hat{\beta}_\xi)) + B(X_\xi\hat{\beta}_\xi)'B(X_\xi\hat{\beta}_\xi).$$

Let us calculate the decomposition of the MSE trace of $\hat{Y}_\xi$:

$$\begin{aligned}\mathrm{tr}(\mathrm{EQM}(\hat{Y}_\xi)) &= \mathrm{tr}(\mathrm{V}(X_\xi\hat{\beta}_\xi)) + B(X_\xi\hat{\beta}_\xi)'B(X_\xi\hat{\beta}_\xi) \\ &= \mathrm{tr}(\mathrm{V}(P_{X_\xi}Y)) + (\mathbb{E}(X_\xi\hat{\beta}_\xi) - X\beta)'(\mathbb{E}(X_\xi\hat{\beta}_\xi) - X\beta) \\ &= |\xi|\sigma^2 + \|(I-P_{X_\xi})X\beta\|^2. \qquad (5.1)\end{aligned}$$

In order to take out P_{X_ξ} from the brackets of the variance, P_{X_ξ} needs to be fixed and therefore *the model choice X_ξ does not depend on the data with which we* *evaluate the projection matrix otherwise the matrix is random.* If the choice of variables has been made on the same data set than the one used to estimate the parameters, we should consider adding another bias term called selection bias. Turning back to our example, we calculate the MSE of the 7 models

$$Y \quad = \quad \beta_1 X_1 + \beta_2 X_2 + \varepsilon = X_{12}\beta_{12} + \varepsilon.$$

We consider the model with one variable X_1, we have for the term tr(EQM), using $\mathcal{H}_2$ and some projectors properties (symmetry, impotency and trace):

$$\begin{aligned}
\operatorname{tr}(\mathrm{EQM}(X_1\hat{\beta}_1)) &= \operatorname{tr}(\mathrm{V}(X_1\hat{\beta}_1)) + B(X_1\hat{\beta}_1)'B(X_1\hat{\beta}_1) \\
&= \operatorname{tr}(\mathrm{V}(P_{X_1}Y)) + \|\mathbb{E}(X_1\hat{\beta}_1) - X_{12}\beta_{12}\|^2 \\
&= \sigma^2 \operatorname{tr}(P_{X_1}) + \|\mathbb{E}(P_{X_1}(X_{12}\beta_{12} + \varepsilon)) - X_{12}\beta_{12}\|^2 \\
&= \sigma^2 + \|P_{X_1^\perp} X_{12}\beta_{12}\|^2.
\end{aligned}$$

We have thus :

$$\begin{aligned}
\operatorname{tr}(\mathrm{EQM}(X_1\hat{\beta}_1)) &= \sigma^2 + \|P_{X_1^\perp} X_{12}\beta_{12}\|^2 \\
\operatorname{tr}(\mathrm{EQM}(X_2\hat{\beta}_2)) &= \sigma^2 + \|P_{X_2^\perp} X_{12}\beta_{12}\|^2 \\
\operatorname{tr}(\mathrm{EQM}(X_3\hat{\beta}_3)) &= \sigma^2 + \|P_{X_3^\perp} X_{12}\beta_{12}\|^2 \\
\operatorname{tr}(\mathrm{EQM}(X_{12}\hat{\beta}_{12})) &= 2\sigma^2 \\
\operatorname{tr}(\mathrm{EQM}(X_{13}\hat{\beta}_{13})) &= 2\sigma^2 + \|P_{X_{13}^\perp} X_{12}\beta_{12}\|^2 \\
\operatorname{tr}(\mathrm{EQM}(X_{13}\hat{\beta}_{13})) &= 2\sigma^2 + \|P_{X_{13}^\perp} X_{12}\beta_{12}\|^2
\end{aligned}$$

$$\begin{aligned}
\operatorname{tr}(\mathrm{EQM}(X_{23}\hat{\beta}_{23})) &= 2\sigma^2 + \|P_{X_{23}^\perp} X_{12}\beta_{12}\|^2 \\
\operatorname{tr}(\mathrm{EQM}(X_{123}\hat{\beta}_{123})) &= 3\sigma^2.
\end{aligned}$$

Choosing the model which has the smallest tr(EQM) among the seven initial models is equivalent to analysing the tr(EQM) of the following four models :

$$\operatorname{tr}(\mathrm{EQM}(X_1\hat{\beta}_1)), \quad \operatorname{tr}(\mathrm{EQM}(X_2\hat{\beta}_2)), \quad \operatorname{tr}(\mathrm{EQM}(X_3\hat{\beta}_3)) \quad \text{and} \quad \operatorname{tr}(\mathrm{EQM}(X_{12}\hat{\beta}_{12})).$$

In order to make further comments, we assume now that

- we know all the unknown quantities
- the smallest squared norm between $\|P_{X_1^\perp} X_{12}\beta_{12}\|^2$, $\|P_{X_2^\perp} X_{12}\beta_{12}\|^2$ and $\|P_{X_3^\perp} X_{12}\beta_{12}\|^2$ is $\|P_{X_1^\perp} X_{12}\beta_{12}\|^2$.

We must therefore choose between

Model $\{1\}$: $\operatorname{tr}(\mathrm{EQM}(X_1\hat{\beta}_1)) = \sigma^2 + \|P_{X_1^\perp} X_{12}\beta_{12}\|^2$

Model $\{1, 2\}$: $\operatorname{tr}(\mathrm{EQM}(X_{12}\hat{\beta}_{12})) = 2\sigma^2$ (supposed to be the true one)

In order to choose the model which has the smallest tr(EQM), we need to compare σ^2 to $\|P_{X_1^\perp} X_{12}\beta_{12}\|^2$, all depends on the respective value of σ^2 and $\|P_{X_1^\perp} X_{12}\beta_{12}\|^2$. In the example of figure 5.3, we select model 2 (the true model) since in this case $\|P_{X_1^\perp} X_{12}\beta_{12}\|^2 > \sigma^2$. If on the other hand, $\|P_{X_1^\perp} X_{12}\beta_{12}\|^2 < \sigma^2$, we select model 1, in other words a model which is slightly wrong (bias term) but more precise (the variance is smaller) than the true model.

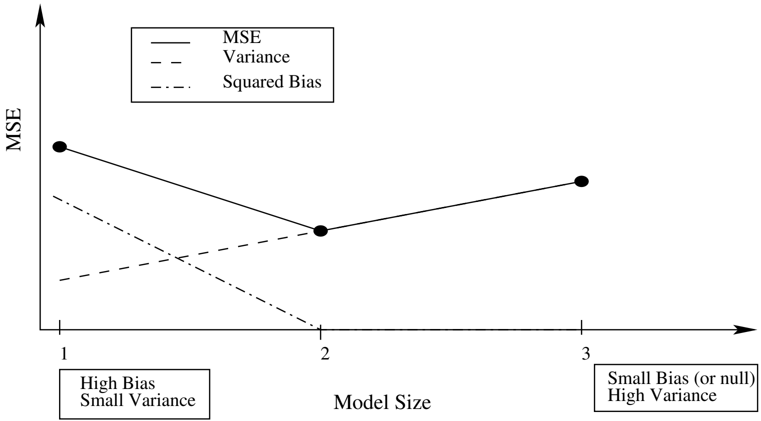


Figure 5.3 – Trade-off biass^2/variance in the case where $\operatorname{tr}\mathrm{EQM}(1) > 2\sigma^2$.

In general it is difficult to estimate the bias because the parameter value is unknown. On the other hand it is easier to estimate the variance. We will look at procedures for estimating MSE later on in this chapter but first we will introduce one more definition which is easier to handle: the prediction MSE or its trace.

5.3.4 Mean Squared Prediction Error

MSE or its trace is a standard criterion in statistics, but it doesn't involve the new observations $Y^\star$. If we want to evaluate the MSPE of these new observations $Y^\star$ we have the following definition :

Definition 5.2 (MSPE)
Consider $x^\star \in \mathbb{R}^p$, a new observation, and $x^\star_\xi$ its components corresponding to ξ. The MSPE is defined by

$$\mathrm{EQMP}(\hat{y}^p_\xi) = \mathbb{E}((x^{\star\prime}_\xi\hat\beta_\xi - y^\star)^2) = \mathrm{EQM}(x^{\star\prime}_\xi\hat\beta_\xi) + \sigma^2 - 2\mathbb{E}([x^{\star\prime}_\xi\hat\beta_\xi - x^{\star\prime}\beta]\varepsilon^\star).$$

If $\varepsilon^\star$ is not correlated with the ε (hypothesis $\mathcal{H}_2$), we then have

$$\mathrm{EQMP}(\hat{y}^p_\xi) = \mathrm{EQM}(x^\star_\xi\hat\beta_\xi) + \sigma^2.$$

If we possess $n^\star$ new observations $x^\star$ grouped in a matrix $X^\star$ we used trace of EQMP

$$\operatorname{tr}(\mathrm{EQMP}(\hat{Y}^p_\xi)) = \operatorname{tr}(\mathrm{EQM}(X^\star_\xi\hat\beta)) + n^\star\sigma^2 - 2\mathbb{E}((X^\star_\xi\hat\beta_\xi - X^\star\beta)'\varepsilon^\star).$$

If $\varepsilon^\star$ is not correlated with the ε, we then have

$$\mathrm{tr}(\mathrm{EQMP}(\hat{y}^p_\xi)) = \mathrm{tr}(\mathrm{EQM}(x^\star_\xi \hat{\beta}_\xi)) + n^\star \sigma^2.$$

Going back to our previous example

$$Y \quad = \quad \beta_1 X_1 + \beta_2 X_2 + \varepsilon = X_{12}\beta_{12} + \varepsilon$$

and assume that we have $n^\star$ new observations concatenated in the matrix $X^\star$. We thus have

$$\begin{aligned}
\mathrm{tr}(\mathrm{EQMP}(X^\star_1 \hat{\beta}_1)) &= (n^\star + 1)\sigma^2 + \|P_{X_1^\perp} X^\star_{12}\beta_{12}\|^2 \\
\mathrm{tr}(\mathrm{EQMP}(X^\star_2 \hat{\beta}_2)) &= (n^\star + 1)\sigma^2 + \|P_{X_2^\perp} X^\star_{12}\beta_{12}\|^2 \\
\mathrm{tr}(\mathrm{EQMP}(X^\star_3 \hat{\beta}_3)) &= (n^\star + 1)\sigma^2 + \|P_{X_3^\perp} X^\star_{12}\beta_{12}\|^2 \\
\mathrm{tr}(\mathrm{EQMP}(X^\star_{12} \hat{\beta}_{12})) &= (n^\star + 2)\sigma^2 \\
\mathrm{tr}(\mathrm{EQMP}(X^\star_{13} \hat{\beta}_{13})) &= (n^\star + 2)\sigma^2 + \|P_{X_{13}^\perp} X^\star_{12}\beta_{12}\|^2 \\
\mathrm{tr}(\mathrm{EQMP}(X^\star_{23} \hat{\beta}_{23})) &= (n^\star + 2)\sigma^2 + \|P_{X_{23}^\perp} X^\star_{12}\beta_{12}\|^2 \\
\mathrm{tr}(\mathrm{EQMP}(X^\star_{123} \hat{\beta}_{123})) &= (n^\star + 3)\sigma^2.
\end{aligned}$$

If we apply blindly the definition of MSPE (made for new observations) on the given estimation set X, Y we get for the first model:

$$\mathrm{tr}(\mathrm{EQMP}(\hat{Y}(X_1)) = \mathrm{tr}(\mathbb{E}((\hat{Y}(X_1) - Y)(\hat{Y}(X_1) - Y)')$$

that can be split in squared bias and variance leading to $\|P_{X_1^\perp} X\beta\|^2 + \sigma^2(n-1)$. We can do the same calculation for all the three models, and we obtain

$$\begin{aligned}
\mathrm{tr}(\mathrm{EQMP}(\hat{Y}(X_1)) &= \|P_{X_1^\perp} X\beta\|^2 + \sigma^2(n-1) \\
\mathrm{tr}(\mathrm{EQMP}(\hat{Y}(X_{12})) &= \sigma^2(n-2) \\
\mathrm{tr}(\mathrm{EQMP}(\hat{Y}(X_{123}))) &= \sigma^2(n-3).
\end{aligned}$$

pas clair

The tr(EQMP) wrongly applied on estimation set tells us that we need to select the model with the highest number of explanatory variables. In fact this criterion makes no sense when it is used with data which has been used to estimate the parameters. We just recover the same phenomenon as the one seen with the estimation error.

5.4 Standard Criteria for Models Selection

We will look now at the standard methods of model selection. The main selection criteria are AIC, BIC and their extensions; other exists such as C_p or $\mathrm{R}^2\,a$ but they won't be introduced here. On the other hand we have already seen that the F test between nested models allows us to compare models between each other using a classical/ standard test procedure and we will try to compare these three: AIC, BIC and F.

Let us begin by the F test.

5.4.1 Tests between Nested Models

If the competing models are nested within each other, it is then possible to use a test procedure. We write the model ξ with $|\xi|$ variables and the model ξ_{+1} corresponding to the model ξ to which we have added an additional variable. In order to choose between these two nested models, we have the following test statistic

$$F = \frac{\mathrm{SSR}(\xi) - \mathrm{SSR}(\xi_{+1})}{\hat{\sigma}^2}.$$

For F to be Fisher distributed, the $\hat{\sigma}^2$ estimate must have χ^2 distribution independent from the numerator. There are two different manners to obtain the σ^2 estimate:

1. The σ^2 estimator is derived using $\mathrm{SSR}(\xi_{+1})/(n - |\xi| - 1)$.
   The σ^2 estimator is obtained from the "larger" model, in other words the (ξ_{+1}) model. This solution is in general used by statistical packages ;

2. The σ^2 estimator is derived using $\mathrm{SSR}(p)/(n - p)$.
   The estimator is obtained from the full model.

We have therefore the following theorem.

Theorem 5.1 (Tests between nested models)
Given two models, the ξ and ξ_{+1} models. The test statistic which enables us to test the hypothesis $\mathrm{H}_0 : \quad \mathbb{E}Y \in \Im(X_\xi)$ *against the alternative hypothesis* $\mathrm{H}_1 : \quad \mathbb{E}Y \in \Im(X_{\xi_{+1}})$*, is*

1. *The variance σ^2 is estimated using* $\mathrm{SSR}(\xi_{+1})/(n - |\xi| - 1)$. *If*

$$F_1 = \frac{\mathrm{SSR}(\xi) - \mathrm{SSR}(\xi_{+1})}{\mathrm{SSR}(\xi_{+1})} \times (n - |\xi| - 1) > f_{1,n-|\xi|-1}(1-\alpha)$$

   then the ξ model is rejected at the α test level, in favour of the (ξ_{+1}) model, a variable is added to the model.

2. *The variance σ^2 is estimated by* $\mathrm{SSR}(p)/(n - p)$. *If*

$$F_2 = \frac{\mathrm{SSR}(\xi) - \mathrm{SSR}(\xi_{+1})}{\mathrm{SSR}(p)} \times (n - p) > f_{1,n-p}(1-\alpha).$$

   the the ξ model is rejected at the α test level, in favour of the (ξ_{+1}) model, a variable is added to the model.

It is difficult to compare these two manners of proceeding since σ^2 is not estimated in the same way.

5.4.2 Criteria using Likelihood

Under the normality assumption for the residuals, the log-likelihood of the sample is

$$\mathcal{L}(Y,\beta,\sigma^2) \quad = \quad -\frac{1}{2\sigma^2}\|Y - X\beta\|^2 - \frac{n}{2}(\log\sigma^2 + \log 2\pi).$$

The log-likelihood (evaluated at the maximum likelihood estimator) for the model with $|\xi|$ variables is therefore equivalent to

$$\mathcal{L}(Y,\beta,\sigma^2_\xi) \quad = \quad -\frac{n}{2}\log\frac{\mathrm{SSR}(\xi)}{n} - \frac{n}{2}(1 + \log 2\pi).$$

Selecting a model using maximum likelihood is equivalent to selecting a model having the smallest SSR . We need therefore to introduce a penalty. In order to minimise a criterion, we proceed inversely to the log-likelihood and the criteria are in general written

$$\min_{\beta,\xi,\sigma^2.} -2\mathcal{L}(Y,\beta,\sigma^2,\xi) + 2|\xi| f(n),$$

where $f(n)$ is a penalty function depending n.

Akaike Information Criterion (AIC)

This criterion introduced by Akaike in 1973 is defined by a model containing the indexed/ subscripted variables by ξ :

$$\mathrm{AIC}(\xi) = -2\log\mathcal{L}(\xi) + 2|\xi|.$$

By definition $f(n)$ is equal to 1. AIC penalises the log-likelihood by twice the number of parameters $|\xi|$. We obtain an equivalent definition

$$\mathrm{AIC}(\xi) = cte + n\log\frac{\mathrm{SSR}(\xi)}{n} + 2|\xi|$$

Using this criterion is easy : we only need to calculate it for all the competing ξ models and choose the one which has the smallest AIC.

Bayesian Information Criterion (BIC)

BIC defined by Schwarz in 1978 is defined as

$$\mathrm{BIC}(\xi) = -2\log\mathcal{L}(\xi) + |\xi|\log n = cte + n\log\frac{\mathrm{SSR}(\xi)}{n} + |\xi|\log n.$$

I also consists in penalising the log-likelihood by the number of $|\xi|$ parameters times a function of the observations (and no longer 2). By definition, $f(n)$ is equal to $\log n/2$. So, as the number of observations n increases the penalty increases

and in general larger than 2 (as soon as $n > 7$) and therefore *BIC tends to select smaller models than AIC.*
Depending on the number of individuals n and the number of selected variables, we can summarise the criteria and the size of the model in the following table:

Standard Criteria	$\lvert\xi\rvert$ size of model
TEST or BIC	small
AIC	bigger

Table 5.2 – Comparison of the $\lvert\xi\rvert$ sizes of the selected models with $n > 7$.

5.5 Selection Procedure

Model selection can be seen as the search for an optimal model, in the sense of the chosen criterion, among all the other possibilities. This can be seen as an optimisation of an objective function (the criterion). Consequently, and similarly to the options in optimisation, we can either do an exhaustive search since the number of models is finite, or from a starting point use a method of optimising the objective function.
We note in general that finding a global minimum of the objective function is not guaranteed in the step by step procedure and only a local optimum depending on the starting point will be found. If the explanatory variables are orthogonal, then the optimum will always be a global optimum.

5.5.1 Exhaustive Search

When all the models with p variables are possible, there are $2^p - 1$ possibilities and therefore we cannot envisage to use this method when p is large. Some algorithmic procedures however minimise the number of computations allowing us to consider this possibility in the case of numbers of moderate size.
Note that this type of search makes no sense when we want to use tests since this procedure only compare two models nested within each other.

5.5.2 Sequential Methods

This type of search is necessary for tests since we can only test nested models. On the other hand, it does not allow in general to find a local optimum. It is best to repeat this procedure from different starting points. Concerning other criteria, this selection procedure can only be advised on when an exhaustive search is impossible (n large, p large, etc.).

Forward Selection

At each step, a variable is added to the model.

- If the forward selection method uses a F test, we add the variable X_i whose (*p-value*) associated with the partial Fisher test statistic which compares two models is smallest. We stop when all the variables have been added or when the p-value is larger than a given threshold value.

- If the forward selection method uses a choice criterion, we add the variable X_i whose addition to the model leads to an optimisation of the choice criterion. We stop when either all variables have been added or when none of the variables allows the optimisation of the criterion of choice. (see also fig. 5.4).

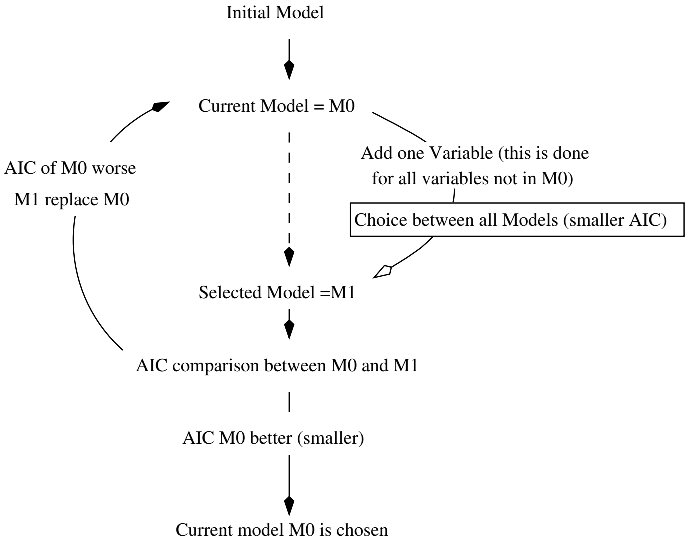

Initial Model
Current Model = M0
Add one Variable (this is done for all variables not in M0)
Choice between all Models (smaller AIC)
AIC of M0 worse
M1 replace M0
Selected Model =M1
AIC comparison between M0 and M1
AIC M0 better (smaller)
Current model M0 is chosen

Figure 5.4 – Forward procedure using AIC.

Backward Selection

First, all the variables are added to the model.

- If backward elimination uses a F test, we remove he variable X_i whose *p-value*, associated to the partial Fisher test statistic is largest. We stop when all the variables have been removed from the model or when the *p-value* is smaller than a given threshold value.

- If the backward elimination procedure uses a choice criterion, we remove the variable X_i which lead the biggest value of the criterion. We stop when all the variables have been removed or when none of the variables allows an increase of the choice criterion.

Stepwise selection

It is the same principle than for the forward selection procedure , except that we can remove variables already introduced. Indeed, it can happen that some of the variables first introduced are no longer significant after the introduction of new variables.

Intercept

We note that in general the “constant”, made of 1 and associated to the “intercept”, is in general treated separately and is always present in the model. This variable is the “general mean” of Y the variable to be explained. The other variables are seen to be refinement to explain better.

Chapter 6

Ridge, Lasso and Elastic net

6.1 Ridge regression

The least squares problem consists in finding the coefficient vector $\hat{\beta}$ that minimizes the ordinary least squares, i.e.:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|^2. \tag{6.1}$$

The assumption $\mathcal{H}_1$ (the matrix X is of full rank) then allows us to find a unique solution to the problem posed, namely $\hat{\beta} = (X'X)^{-1}X'Y$. When $rank(X) < p$, the matrix $(X'X)$ is not invertible and the relationship giving $\hat{\beta}$ no longer makes sense. We can still project Y onto $\Im(X)$, but $\hat{Y}$ no longer admits a unique decomposition on the columns of X, the model is then unidentifiable.
Finally, note that when X is full rank, the variance of $\hat{\beta}$ is

$$\mathrm{V}(\hat{\beta}) \;\; = \;\; \sigma^2(X'X)^{-1}.$$

and depends directly on the rank of X. So, even when X is of full rank but $X_j \approx \sum_{i \neq j} \alpha_i X_i$ (we often say the (explanatory) variables are highly correlated empirically), the variance of the estimators will be high and the precision will decrease. It is therefore important to use methods adapted to rank deficiency. A fairly old method to make a matrix invertible is to modify its diagonal.

6.1.1 A historical solution

This method was proposed in the 70 and consists in replacing $(X'X)^{-1}$ by $(X'X + \lambda\mathbb{I})^{-1}$. We obtain the ridge estimator

$$\hat{\beta}_{\text{ridge}}(\lambda) = (X'X + \lambda\mathbb{I})^{-1}X'Y, \tag{6.2}$$

where λ is a positive constant to be determined. The choice of λ is very important for the method's performance. This is because,

- $\hat{\beta}_{\text{ridge}}(\lambda) \approx 0$ for high λ values;
- $\hat{\beta}_{\text{ridge}}(\lambda) \approx \hat{\beta}$ for low values of λ and in the case where $\hat{\beta}$ exists.

6.1.2 Minimizing penalized LS

We saw in the previous section that the presence of colinearities between the columns of tends to increase the variance of the estimators. One approach to reducing this variance is to penalize the LS criterion by the norm of the parameters. This involves minimizing

$$\|Y - X\beta\|^2 + \lambda\|\beta\|^2. \tag{6.3}$$

where $\lambda \geq 0$ is a parameter to be calibrated. Other penalties are possible and will be discussed later. To obtain the solution to this problem, we derive with respect to β

$$2(-X')(Y - X\beta) + 2\lambda\beta.$$

And cancelling the derivative, we obtain the solution

$$\hat{\beta}_{\text{ridge}}(\lambda) = (X'X + \lambda\mathbb{I})^{-1}X'Y.$$

We note that the solution to the problem (6.3) is the ridge estimator.

6.1.3 Equivalence with a constraint on the norm of coefficients

Another way of looking at the ridge method is to minimize the LS criterion under a constraint on the parameter norm. Consider the estimator $\tilde{\beta}$ defined by

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^p, \|\beta\|^2 \leq \delta}{\text{argmin}} \|Y - X\beta\|^2, \tag{6.4}$$

where $\delta \geq 0$ is a parameter to be calibrated. We compute the Lagrangian to obtain the solution of the problem

$$\|Y - X\beta\|^2 + \mu(\|\beta\|^2 - \delta).$$

A necessary optimum condition is given by the cancellation of its partial derivatives derivatives at the optimum point $(\tilde{\beta}, \tilde{\mu})$

$$\begin{aligned} -2X'(Y - X\tilde{\beta}) + 2\tilde{\mu}\tilde{\beta} &= 0 \\ \|\tilde{\beta}\|^2 - \delta &= 0. \end{aligned} \tag{6.5}$$

The first equation shows that the solution to this problem is the ridge estimator

$$\tilde{\beta} = \hat{\beta}_{\text{ridge}} = (X'X + \tilde{\mu}\mathbb{I})^{-1}X'Y.$$

dépend de $\tilde{\mu}$!!

In order to calculate the value of $\tilde{\mu}$, pre-multiply (6.5) on the left by $\hat{\beta}'_{\text{ridge}}$, we get $\tilde{\mu} = (\hat{\beta}_{\text{ridge}}X'Y - \hat{\beta}'_{\text{ridge}}X'X\hat{\beta}_{\text{ridge}})/\|\hat{\beta}_{\text{ridge}}\|^2$. We can also check that this pair is indeed a minimum of the function by noting that the hessian is a symmetrical matrix of the form $A'A$, i.e. positive semi-definite.
Geometrically, ridge regression amounts to searching in a ball of $\mathbb{R}^p$ of radius δ, the coefficient $\hat{\beta}_{\text{ridge}}$ closest in the least squares sense. Placing ourselves now in the space of observations $\mathbb{R}^n$, the image of the constraint ball by X is an ellipsoid constraint. Since the ellipsoid is included in $\Im(X)$, in the case where δ is small, the optimum $\hat{\beta}_{\text{ridge}}$ is such that $X\hat{\beta}_{\text{ridge}}$ is the projection of $X\hat{\beta}$ onto this ellipsoid constraint (see fig. 6.1). In the opposite case, where $\|\hat{\beta}|^2 \leq \delta$, $\hat{\beta}$ is in or on the the ellipsoid and is indeed the solution.

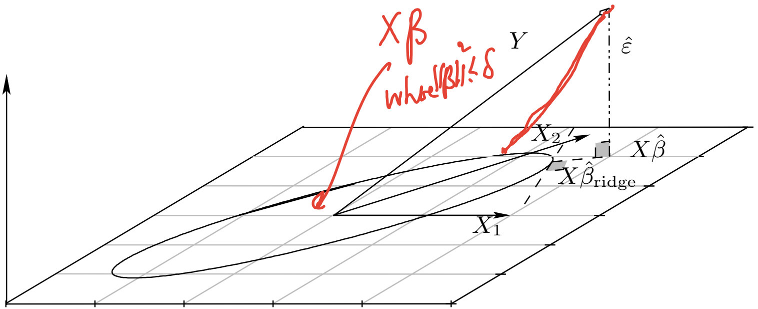

$X\beta$
where $\|\beta\|^2 \leq \delta$
Y
$\hat{\varepsilon}$
X_2
$X\hat{\beta}$
$X\hat{\beta}_{\text{ridge}}$
X_1

Figure 6.1 – Coefficients constraint and ridge regression : $\hat{\beta}_{\text{ridge}}$ represents the ridge estimator and $\hat{\beta}$ represents the the LS estimator.

6.1.4 Statistical properties of the ridge estimator $\hat{\beta}_{\text{ridge}}$

Let's recall the main properties of the OLS estimator (non-penalized)

1. *Existence and uniqueness*: the fit (via LS) $\hat{Y} = P_X Y + \bar{y}\mathbb{1}$ exists and is unique, and we can always find one (and especially several) $\hat{\beta}$ that satisfy $\hat{Y} = X\hat{\beta} + \bar{y}\mathbb{1}$. If hypothesis $\mathcal{H}_1$ is verified, then the vector $\hat{\beta}$ is unique and we have the explicit formula (*closed form*) $\hat{\beta} = (X'X)^{-1}X'Y$.

2. *Nullity of all coefficients of* $\hat{\beta}$: for LS, except in special cases (such as $X'Y = 0$, all explanatory variables X are uncorrelated to Y), there's no reason to get all coordinates of $\hat{\beta}$ equal to zero.

   ou plutôt : $Y \perp \Im(X)$.

3. *Bias*: the OLS estimator $\hat{\beta}$ is unbiased (when the regression model is true and $\mathbb{E}(\varepsilon) = 0$)

4. *Variance*: we've already calculated the variance as $\sigma^2(X'X)^{-1}$.

Let's consider these four items for ridge estimator.

1. *Existence and uniqueness*: the *closed form* formula for ridge estimator

   $$\hat{\beta}_{\text{ridge}} \quad = \quad (X'X + \lambda\mathbb{I})^{-1}X'Y.$$

   shows us that, as soon as $\lambda > 0$ fixed, existence and uniqueness of the ridge estimator (the result of a function of λ), without having to use assumption $\mathcal{H}_1$. However $\hat{Y}_{\text{ridge}}(\lambda) = X\hat{\beta}_{\text{ridge}}(\lambda) + \bar{y}\mathbb{1}$ is not a projection.

2. *Nullity of all coefficients of* $\hat{\beta}_{\text{ridge}}$: The same as LS estimator applies to the ridge estimator: as soon as $\lambda > 0$ is finite, we have $\hat{\beta}_{\text{ridge}}(\lambda) \neq 0$, except in the very special case where $X'Y = 0$.

3. *Bias*: let's return to the definition of the LS estimator

   $$\hat{\beta} = (X'X)^{-1}X'Y$$

   and pre-multiplying it on the left by $X'X$, we have $\overset{\prime}{\cancel{X}}X\hat{\beta} = X'Y$ that we can put inside the *closed form* formula for ridge estimator giving us

   $$\hat{\beta}_{\text{ridge}} \quad = \quad (X'X + \lambda\mathbb{I})^{-1}X'X\hat{\beta}. \tag{6.6}$$

   This formula makes it easy to calculate the bias and variance of the ridge estimator. Calculating the the expectation of the ridge estimator gives

   $$\begin{aligned}
   \mathbb{E}(\hat{\beta}_{\text{ridge}}) &= (X'X + \lambda\mathbb{I})^{-1}(X'X)\mathbb{E}(\hat{\beta}) \\
   &= (X'X + \lambda\mathbb{I})^{-1}(X'X)\beta \\
   &= (X'X + \lambda\mathbb{I})^{-1}(X'X + \lambda\mathbb{I} - \lambda\mathbb{I})\beta \\
   &= \beta - \lambda(X'X + \lambda\mathbb{I})^{-1}\beta.
   \end{aligned}$$

   The bias of the ridge estimator is therefore

   $$B(\hat{\beta}_{\text{ridge}}) \quad = \quad -\lambda(X'X + \lambda\mathbb{I})^{-1}\beta. \tag{6.7}$$

   In general, this quantity is non-zero. ridge estimator is biased and the regression is said to be biased. This bias can be thought as a handicap compared to the the LS estimator.

4. *Variance*: it can be calculated as follows

   $$\begin{aligned}
   V(\hat{\beta}_{\text{ridge}}) &= \mathrm{V}((X'X + \lambda\mathbb{I})^{-1}X'Y) \\
   &= (X'X + \lambda\mathbb{I})^{-1}X'\,\mathrm{V}(Y)X(X'X + \lambda\mathbb{I})^{-1} \\
   &= \sigma^2(X'X + \lambda\mathbb{I})^{-1}X'X(X'X + \lambda\mathbb{I})^{-1}. \qquad (6.8)
   \end{aligned}$$

   Compared to LS variance, the ridge estimator variance involves $(X'X+\lambda\mathbb{I})^{-1}$ and not $(X'X)^{-1}$. As $\lambda\mathbb{I}$ increases the eigenvalues of $(X'X + \lambda\mathbb{I})$ it reduces the variance.

6.1.5 Comparison of ridge and LS estimators through MSE

The comparison of ridge and LS estimators is usually done using EQM. From the expressions of the bias and variance (see previous subsection), we have

$$\begin{aligned}
\mathrm{EQM}(\hat{\beta}) &= \sigma^2(X'X)^{-1} \\
\mathrm{EQM}(\hat{\beta}_{\text{ridge}}) &= \mathbb{E}(\hat{\beta}_{\text{ridge}}-\beta)\mathbb{E}(\hat{\beta}_{\text{ridge}}-\beta)'+V(\hat{\beta}_{\text{ridge}}) \\
&= \lambda^2(X'X+\lambda\mathbb{I})^{-1}\beta\beta'(X'X+\lambda\mathbb{I})^{-1}+\sigma^2(X'X+\lambda\mathbb{I})^{-1}X'X(X'X+\lambda\mathbb{I})^{-1} \\
&= (X'X+\lambda\mathbb{I})^{-1}\left[\lambda^2\beta\beta'+\sigma^2(X'X)\right](X'X+\lambda\mathbb{I})^{-1}.
\end{aligned}$$

It's difficult to compare two matrices, so consider the trace of the MSE matrix, we have

$$\mathrm{tr}[\mathrm{EQM}(\hat{\beta})] = \sigma^2\,\mathrm{tr}((X'X)^{-1}) = \sigma^2\left(\sum_{j=1}^{p}\frac{1}{\lambda_j}\right),$$

where $\{\lambda_j\}_{j=1}^{p}$ are the eigenvalues of $X'X$. Since some of these eigenvalues are zero or almost zero, the trace of the MSE is infinite or very large. We can show that the trace of the MSE matrix of the ridge estimator is equal to

$$\mathrm{tr}[\mathrm{EQM}(\hat{\beta}_{\text{ridge}})] = \sum_{i=1}^{r}\frac{\sigma^2\lambda_i+\lambda^2[P'\beta]_i^2}{(\lambda_i+\lambda)^2} \quad \text{where} \quad X'X = P\,\mathrm{diag}(\lambda_i)P'.$$

This last equation gives the form of the MSE as a function of the regression parameter λ. We can find a sufficient condition find a sufficient condition on λ, a condition that is independent of the explanatory variables,

$$\lambda \le \frac{2\sigma^2}{\beta'\beta},$$

which allows us to know that the MSE trace of the ridge estimator is smaller than that of the LS estimator. In other words, when $\lambda \le 2\sigma^2/\beta'\beta$, the ridge regression is more accurate (in parameter estimation) than the LS estimator in the sense of the MSE trace. However, this condition depends on unknown parameters β and σ^2 and it cannot be used to select a value for λ.
Note that if X is orthogonal, then by definition $X'X = \mathbb{I}$ and so the definition of ridge regression amounts to dividing $\hat{\beta} = X'Y$ the LS estimator by $(1+\lambda)$ and thus shrink the coefficients of the same amount.

6.2 Problem of centering-scaling variables

Up to now, we've always considered the general model

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

and assumed that one of the explanatory variables could be the constant $\mathbb{1}$. This variable had the same role as the other potential explanatory variables. However, it is usual not to include it in the penalisation. For this reason, we will now consider the following model

$$Y = \mu\mathbb{1} + X\beta + \varepsilon.$$

The value of a coefficient β_j depends on the scale of the measurements of the associated explanatory variable X_j : for example, β_j will be different if the variable is measured in grams or in kilograms. In order to avoid penalizing or favoring one coefficient over another (and by the way one variable against an other), it is desirable that each coefficient be assigned in a similar way. A classical way of achieving it is to reduce all the explanatory variables. A centered-scaled variable $\widetilde{X}_j$ from the variable X_j can be written as

$$\widetilde{X}_j = (X_j - \bar{x}_j\mathbb{1})/\hat{\sigma}_{X_j},$$

where $\bar{x}_j$ is the empirical mean of X_j (*i.e.* $\sum_{i=1}^n x_{ij}/n$) and $\hat{\sigma}^2_{X_j}$ is an estimate of the empirical variance (for example $\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2/n$). *The matrix* $\widetilde{X}$ *will therefore contain centered-scaled variables.* The model (6.9) then becomes

$$Y = \tilde{\mu}\mathbb{1} + \widetilde{X}\widetilde{\beta} + \varepsilon. \tag{6.9}$$

The variables $\tilde{X}$ are centered (and scaled), so they are all orthogonal to the variable $\mathbb{1}$. Since variable $\mathbb{1}$ is excluded from the constraint and is orthogonal to the others, its coefficient, estimated by a regression is simply the empirical mean of the observations of Y: $\widehat{\tilde{\mu}} = \bar{y}$. After estimating the parameters with $\mathbb{1}$ and $\widetilde{X}$ and Y, it is possible to predict a new value $x'_{n+1} = (x_{n+1,1}, \cdots, x_{n+1,p})$ using the following formula:

$$\hat{y}^p_{n+1} = \widehat{\tilde{\mu}} + \sum_{j=1}^p \left(\frac{x_{n+1,j} - \bar{x}_j}{\hat{\sigma}_{X_j}}\right)\widehat{\tilde{\beta}}_j.$$

Note that this forecast can be written as a linear combination of the initial variables

$$\hat{y}^p_{n+1} = \left(\widehat{\tilde{\mu}} - \sum_{j=1}^p \bar{x}_j \frac{\hat{\tilde{\beta}}_j}{\hat{\sigma}_{X_j}}\right) + \sum_{j=1}^p x_{n+1,j}\frac{\hat{\tilde{\beta}}_j}{\hat{\sigma}_{X_j}}.$$

From now on, in order to have the simplest possible notations notations, we'll assume that the variables in the X matrix are centered and scaled, and that the model is written as follows (remove the tildes):

$$Y = \mu\mathbb{1} + X\beta + \varepsilon.$$

6.3 Penalties: ridge, lasso, elasticnet...

Most of the penalties used to regularize the least squares criterion are based on the l_2 and l_1 norms leading to the ridge and lasso estimators. For a given $\lambda > 0$, let's consider the following minimization problems

$$(\hat{\mu}, \hat{\beta}_{\text{ridge}}(\lambda)) = \underset{\mu\in\mathbb{R},\beta\in\mathbb{R}^p}{\text{argmin}} \|Y - \mu\mathbb{1} - X\beta\|^2 + \lambda\|\beta\|_2^2$$

and

$$(\hat{\mu}, \hat{\beta}_{\text{lasso}}(\lambda)) = \underset{\mu\in\mathbb{R},\beta\in\mathbb{R}^p}{\text{argmin}} \|Y - \mu\mathbb{1} - X\beta|^2 + \lambda\|\beta\|_1.$$

Since the variables X_j are centered (scaled), they are orthogonal vector $\mathbb{1}$, the estimator of μ does not depend on λ. Both problem is equivalent to the problem of minimizing LS under constraints:

$$\min_{\mu\in\mathbb{R},\beta\in\mathbb{R}^p:\|\beta\|\leq\delta} \|Y - \mu\mathbb{1} - X\beta\|^2.$$

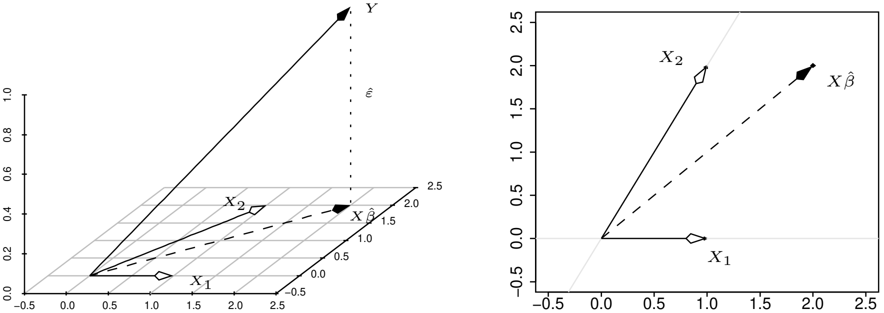
Let's illustrate the choice of the norm graphically via the formulation of LS under constraints. Let's choose a model with two explanatory variables X_1 and X_2 and therefore two coefficients β_1 and β_2:

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

The data in the table 6.1 are used to calculate $\hat{\beta}$ which here is $(1,1)'$ and we can also obtain the graphical representation of the figure 6.2a. The plane of $\Im(X)$ (seen from above) is given in figure 6.2b.

Y	X_1	X_2
2	1	1
2	0	2
1	0	0

Table 6.1 – Data for 3D graphic illustration

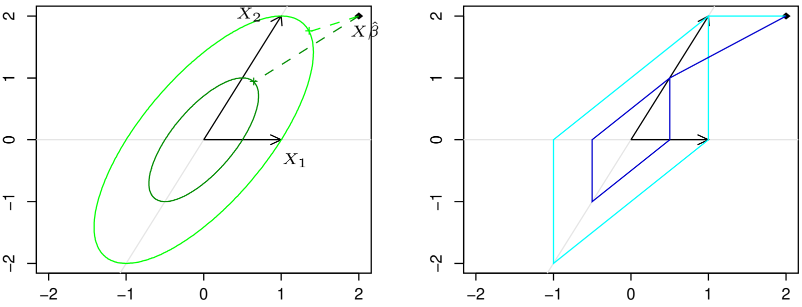


(a) Representation of Y, P_XY and $\hat{\varepsilon}$

(b) Representation of $\Im(X)$.

Figure 6.2 – Graphical presentation of table data 6.1.

We will consider for the l^2 norm squared (*i.e.* $\|\beta\|^2 = \sum_{i=1}^p \beta_i^2$) which corresponds to the ridge regression and a l^1 norm $\|\beta|_1 = \sum_{i=1}^p |\beta_i|$) which corresponds to the lasso regression. Let's represent with a norm 1 and 0.5 in each case in the subspace $\Im(X)$ the constrained subspace in $\Im(X)$. In this example, we'll assume that LS $\hat\beta$ doesn't satisfy the constraint, otherwise the solution to the problem is $\hat\beta$.



(a) $J(\beta)$ is a l^2 norm squared **(b)** $J(\beta)$ is a standard l^1

Figure 6.3 – Representation of subspace $\Im(X)$ with constraints of norm .5 and 1.

In the case of ridge regression (fig. 6.3a) we see that among the points satisfying the constraint (*i.e. the* $(\beta_1, \beta_2)'$ *inside or on the edges of the ellipse*), the closest to $X\hat\beta$ (this is the ridge estimator) has neither of its two coordinates zero, for both the 1 constraint and the 0.25 constraint. On the other hand, in the case of lasso regression (fig. 6.3b), we see that among the points that the constraint (*i.e.* inside or on the edges of the diamond), the closest to $X\hat\beta$ (this is the lasso estimator) has coordinates in the X_1, X_2 coordinate system $(0, 1)'$ or $(0, 0.5)'$ depending on the constraint under consideration. In both cases, the variable variable X_1 has coefficient 0, so it is not selected. Only a constraint greater than 1 would allow to be non-zero. The most frequently used regularization functions penalize vectors that have too many coordinates. They are based on norms or mixtures of norms:

- Ridge regression: $\|\beta\|^2 = \sum_{j=1}^p \beta_j^2$, penalizes β vectors with strong coordinates.

- Lasso regression: $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$, penalizes β vectors with strong coordinates, and also leads to variable selection (see discussion below).

- Regression elasticnet: $\alpha \sum_{j=1}^p |\beta_j| + (1-\alpha) \sum_{j=1}^p \beta_j^2$, which provides a compromise between the two above solutions, at the cost of an additional coefficient to choose α which in general is .5.

- Regression group lasso: the coefficients are naturally into K distinct groups, and we wish to retain or eliminate as a whole, but you don't want to eliminate just one variable in a group. The penalty is then $\sum_{k=1}^{K} \lambda_k \|\beta^{(k)}\|_2$ where $\beta^{(k)}$ is the sub-vector of the coefficients corresponding to the group k.

- Regression fused lasso $\alpha\|\beta\|_1^2 + (1-\alpha)\sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j|$: enables you to select a certain number of variables but also to limit the variations between coefficients of consecutive variables.

6.4 Statistical properties of lasso

For a given $\lambda > 0$, let's consider the following minimization problem

$$(\hat{\mu}, \hat{\beta}_{\text{lasso}}(\lambda)) = \underset{\mu\in\mathbb{R},\beta\in\mathbb{R}^p}{\operatorname{argmin}} \|Y - \mu\mathbb{1} - X\beta\|^2 + \lambda\|\beta\|_1.$$

Since the variables X_j are centered (scaled), they are orthogonal vector $\mathbb{1}$, the estimator of μ does not depend on λ. Let's look at point the following three points for lasso estimator (see section 6.1.4, p.63 for ridge and LS).

1. *Existence and uniqueness*: The function

$$h(\beta) = \|Y - \mu\mathbb{1} - X\beta\|^2 + \lambda\|\beta\|_1 = h_1(\beta) + h_2(\beta)$$

is the sum of two convex functions (since $\lambda > 0$ is fixed) and is therefore convex on $\mathbb{R}^p$. We deduce that it has a minimum. The set of points at which this minimum is attained, set denoted $\{\hat{\beta}_{\text{lasso}}(\lambda)\}$, is non-empty, ensuring existence.
As in ridge and LS we have that $X\hat{\beta}_{\text{lasso}}(\lambda)$ is unique (see exercices). To fulfill uniqueness of the vector $\hat{\beta}_{\text{lasso}}(\lambda)$ we need to add the assumption $\mathcal{H}_1'$: X_ξ is of full rank, where X_ξ is the matrix matrix X restricted to columns $j \in \{1,\ldots,p\}$ for which where $|z_j| = \lambda$, with $z \in \partial h_2$ satisfying the equation (6.10).

It would be nice (and practical) to have a formula that directly gives $\hat{\beta}_{\text{lasso}}$, like the one (6.2) for $\hat{\beta}_{\text{ridge}}$ (or the one for $\hat{\beta}$). For lasso, the function h is not differentiable for any λ (take $0_p \in \mathbb{R}^p$ as an example). This problem can be circumvented by taking the sub-differential, but the equation obtained does not allow us to find $\hat{\beta}_{\text{lasso}}$ with an explicit formula (except in the orthogonal orthogonal case, see section 6.4.2) and an iterative algorithm must be implemented (see section 6.4.2).

2. *Nullity of all coefficients of the estimator vector*: the lasso estimator is the argument of the minimum of the function. For a convex function $x \mapsto f(x)$, we have the following equivalence:

$$\hat{x} \in \underset{x\in\mathbb{R}^p}{\operatorname{argmin}} f(x) \Leftrightarrow 0 \in \partial f(\hat{x})$$

which, applied to our case, tells us that for lasso estimators $\hat{\beta}_{\text{lasso}}(\lambda)$, the sub-differential of vector $0 \in \mathbb{R}^p$. Using the fact that the sub-differential of a sum of two convex functions is the sum of the two sub-differentials and that the sub-differential of a differentiable function (namely $h_1(\beta)$) is the singleton limited to the gradient (so here $\partial h_1(\beta) = \{\nabla h_1(\beta)\} = \{-2X'(Y - \mu\mathbb{1} - X\beta)\}$), we then have

$$0 \in \{-2X'(Y - \mu\mathbb{1} - X\hat{\beta}_{\text{lasso}}(\lambda))\} + \partial h_2(\hat{\beta}_{\text{lasso}}(\lambda)).$$

Finally, using the fact that the sub-differential of $\|.\|_1$ norm is known:

$$z \in \partial h_2(\beta) \Leftrightarrow \begin{cases} z_j = \lambda \operatorname{sign}(\beta_j) = \lambda \frac{\beta_j}{|\beta_j|} & \text{if } \beta_j \neq 0, \\ z_j \in [-\lambda, \lambda] & \text{if } \beta_j = 0, \end{cases}$$

we therefore have a necessary and sufficient condition for $\hat{\beta}_{\text{lasso}}(\lambda)$ to be an argument of the minimum of h:

$$-2X'(Y - \mu\mathbb{1} - X\hat{\beta}_{\text{lasso}}(\lambda)) + z \quad = \quad 0 \tag{6.10}$$

with $z \in \partial h_2(\beta)$. Unfortunately, and as announced, this equation doesn't give a closed form formula for $\hat{\beta}_{\text{lasso}}(\lambda)$. However, we can deduce an interesting fact about the lasso. Recall that X is centered thus $X_j \perp \mathbb{1}\ \forall j$ and $X'\mathbb{1} = 0$. From the equation (6.10), we have, with $z \in \partial h_2(\beta)$:

$$2X'X\hat{\beta}_{\text{lasso}}(\lambda) = 2X'Y - z,$$

which, pre-multiplying by $\hat{\beta}'_{\text{lasso}}(\lambda)$ gives

$$0 \leq 2\hat{\beta}'_{\text{lasso}}(\lambda)X'X\hat{\beta}_{\text{lasso}}(\lambda) = \hat{\beta}'_{\text{lasso}}(\lambda)(2X'Y - z).$$

If we call ξ the set of explanatory variables for which the the coefficient of $\hat{\beta}'_{\text{lasso}}(\lambda)$ is non-zero, then we have by replacing z by its value:

$$0 \leq \sum_{j \in \xi} [\hat{\beta}_{\text{lasso}}(\lambda)]_j (2[X'Y]_j - \lambda \operatorname{sign}([\hat{\beta}_{\text{lasso}}(\lambda)]_j)).$$

In order for this necessary condition to be met the following conditions are needed:

- if $[\hat{\beta}_{\text{lasso}}(\lambda)]_j$ is positive then $2[X'Y]_j > \lambda \operatorname{sign}([\hat{\beta}_{\text{lasso}}(\lambda)]_j) \geq 0$;
- if $[\hat{\beta}_{\text{lasso}}(\lambda)]_j$ is negative then $2[X'Y]_j < \lambda \operatorname{sign}([\hat{\beta}_{\text{lasso}}(\lambda)]_j) \leq 0$.

Taking the largest absolute element of vector $X'Y$ (denoted $\|X'Y\|_\infty = \max_j |[X'Y]_j|$), we have that if $\lambda \geq 2\|X'Y\|_\infty$ then neither of the above two points can be verified, meaning that for $\lambda \geq 2\|X'Y\|_\infty$ we have $\hat{\beta}_{\text{lasso}}(\lambda) = 0$.

In conclusion, the counterpart of point 2) for the lasso is as follows: *if $\lambda \geq 2\|X'Y\|_\infty$ then the vector $\hat{\beta}_{\text{lasso}}(\lambda)$ has all its coordinates zero, no variable is selected.*

As soon as the value of λ falls below this threshold, the first variable, the one whose index corresponds to $\|X'Y\|_\infty$, is added to the model. Recall that if the X and Y variables are centered, $X'Y$ represents, to the nearest $1/n$, the correlation between each X variable and the Y variable. In this case, this corresponds to the explanatory variable most correlated explanatory variable with Y, i.e. the same variable as in a bottom-up selection from a model with just the constant.

3. *Bias and variance*: as we have no explicit formula for the lasso estimator (except in the orthogonal case), it is more difficult to obtain them.

6.4.1 Special case: X orthogonal

When the matrix X is orthogonal (thus $X'X = \mathbb{I}_p$), the LS and ridge estimators ridge estimators are simplified:

$$\begin{aligned}\hat{\beta} =& (X'X)^{-1}X'Y = X'Y\\ \hat{\beta}_{\text{ridge}}(\lambda) =& (X'X + \lambda\mathbb{I})^{-1}X'Y = \frac{X'Y}{1+\lambda}.\end{aligned}$$

The ridge estimator is a contracted version of the OLS estimator: the j^{th} component of the ridge estimator is equal to $\hat{\beta}_j/(1+\lambda)$ where $\hat{\beta}_j$ is the j^{th} component of the LS estimator and therefore each of its coordinates has been divided by $1+\lambda > 1$ (as soon as $\lambda > 0$).
We denote by $\hat{\beta}_{\text{lasso}}(\lambda)$ the lasso estimator obtained by

$$\hat{\beta}_{\text{lasso}}(\lambda) = \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2 + \lambda\|\beta\|_1.$$

We have assumed that X is an orthogonal matrix; this assumption provides an explicit formula for the the lasso estimator. The equation (6.10) becomes

$$2\hat{\beta}_{\text{lasso}}(\lambda) = 2X'Y - z$$

which becomes for each coordinate j

$$[\hat{\beta}_{\text{lasso}}(\lambda)]_j = [X'Y]_j - \frac{z_j}{2}.$$

If the coordinate $[\hat{\beta}_{\text{lasso}}(\lambda)]_j$ is not zero, we have that $z_j = \lambda\,\text{sign}([\hat{\beta}_{\text{lasso}}(\lambda)]_j)$, which gives us for each non-zero coordinate:

$$[\hat{\beta}_{\text{lasso}}(\lambda)]_j = [X'Y]_j - \frac{\lambda\,\text{sign}([\hat{\beta}_{\text{lasso}}(\lambda)]_j)}{2} = [X'Y]_j - \frac{\lambda\,\text{sign}([X'Y]_j)}{2}.$$

If the coordinate is positive, we deduce that $[X'Y]_j > 0$ and furthermore that that $[X'Y]_j > \lambda/2$. If the coordinate is negative, we deduce that $[X'Y]_j < 0$ and that

$[X'Y]_j < -\lambda/2$. We therefore have that $[X'Y]_j$ has the same sign as $[\hat{\beta}_{\text{lasso}}(\lambda)]_j$, which allows us to replace $sign([\hat{\beta}_{\text{lasso}}(\lambda)]_j)$ with the sign of $[X'Y]_j$, or $\frac{[X'Y]_j}{|[X'Y]_j|}$:

$$[\hat{\beta}_{\text{lasso}}(\lambda)]_j = [X'Y]_j(1 - \frac{\lambda}{2|[X'Y]_j|})_+$$

and we keep the positive part of the factor furthest to the right to signs are identical.

Recall that the OLS estimator in this case is $\hat{\beta}_j = [X'Y]_j$. Thus the j^{th} component of the estimator lasso estimator is

$$\text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda/2)_+$$

where $(x)_+ = \max(x, 0)$. The lasso sets to 0 the components for which for which the OLS estimator is smaller in absolute value than 2. The other components are simply simply the components of the corresponding OLS estimator shrunk to 0 by $\lambda/2$.

The figure 6.4 represents the behavior of the ridge and lasso estimators as a function of the value of the LS estimator. We speak of soft tresholding for ridge and hard tresholding for lasso. for lasso.

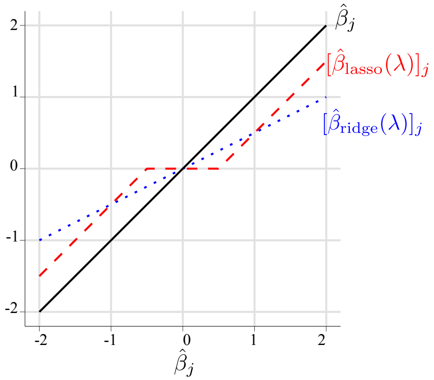


Figure 6.4 – Ridge (dotted), lasso (dashed) and OLS (solid) estimators as a function of the OLS estimator with $\lambda = 1$

.

6.4.2 Lasso algorithm

The most popular algorithm, introduced in 1998, is a coordinate-by-coordinate descent. As we shall see, if we fix all β coordinates except j^{th}, we can easily find an explicit way of writing this coordinate $[\hat{\beta}_{\text{lasso}}(\lambda)]_j$ as a function of the data and the other coordinates. The idea of this algorithm is to perform this minimization each of the j coordinates iteratively, and stop when the algorithm is no longer progressing.

Let's show that, if we assume that all β coordinates except j^{th} are fixed, then we have an explicit formula for it. Note that $\beta_j \mapsto h(\beta)$ is convex and, moreover, derivable as soon as $\beta_j \neq 0$. We then have, leaving aside the notation $\hat{\beta}_{\text{lasso}}(\lambda)$ for a lighter $\hat{\beta}$, that the derivative (with respect to β_j) is zero in $\hat{\beta}_j$:

$$-2X_j'(Y - \bar{y}\mathbb{1} - \sum_{k\neq j}\hat{\beta}_k X_k) + 2\hat{\beta}_j X_j' X_j + \lambda\frac{\hat{\beta}_j}{|\hat{\beta}_j|} = 0 \quad \text{si} \quad \hat{\beta}_j \neq 0.$$

Noting $R_j = X_j'(Y - \bar{y}\mathbb{1} - \sum_{k\neq j}\hat{\beta}_k X_k)$, we get

$$2R_j = 2\hat{\beta}_j X_j' X_j + \lambda\frac{\hat{\beta}_j}{|\hat{\beta}_j|} \quad \text{si} \quad \hat{\beta}_j \neq 0.$$

Since $\lambda > 0$ and $X_j'X_j > 0$, we deduce that the sign of R_j is the same as that of $\hat{\beta}_j$ and we can therefore replace replace $\frac{\hat{\beta}_j}{|\hat{\beta}_j|}$ by $\frac{R_j}{|R_j|}$, which gives us

$$\hat{\beta}_j = \frac{R_j}{X_j'X_j}(1 - \lambda\frac{1}{2|R_j|}) \quad \text{si} \quad \hat{\beta}_j \neq 0.$$

Since $\hat{\beta}_j$ is of the same sign as R_j, in order to guarantee this condition in the previous equation, we need to consider only the positive part of the right-most factor:

$$\hat{\beta}_j = \frac{R_j}{X_j'X_j}(1 - \lambda\frac{1}{2|R_j|})_+$$

We thus obtain the algorithm 1.

Algorithm 1: Lasso regression by coordinate-by-coordinate descent (with X variables centered scaled)

```
$\beta^0 \in \mathbb{R}^p$
$k \leftarrow 0$ repeat
    for $j \leftarrow 1$ to $p$ do
        $\beta_j^{k+1} \leftarrow \frac{R_j}{X_j'X_j}(1 - \frac{\lambda}{2|R_j|})_+$ with $R_j = X_j'(Y - \bar{y}\mathbb{1} - \sum_{k\neq j}\beta_k^k X_k)$
    end
    $k \leftarrow k + 1$
until $\beta^{k+1} \approx \beta^k$
```

6.5 Choice of λ by cross validation

The choice of λ is crucial. A small λ and the estimator will be very close to the S estimator and a large λ and the estimator will be close to the origin. Classically the value for λ is obtained via K fold cross validation : the data are split in K groups, we use $K-1$ groups to estimate the parameters and then use the last group for prediction. This is done K times. Let us illustrate that procedure with $K = 4$ (usual choice for K is $K = 10$)

• First define *a grid for* λ.

This is clearly a choice left to the user but the classical grid choice for lasso is the following:

1. Calculate the maximal value for λ:

$$\lambda_0 = 2\|X'Y\|_\infty$$

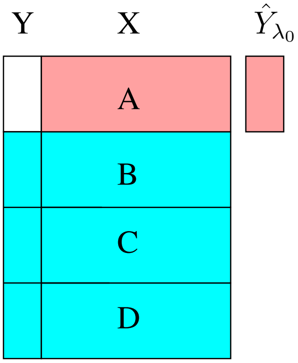
2. Choose a grid length T, usually $T = 100$

3. the grid is:

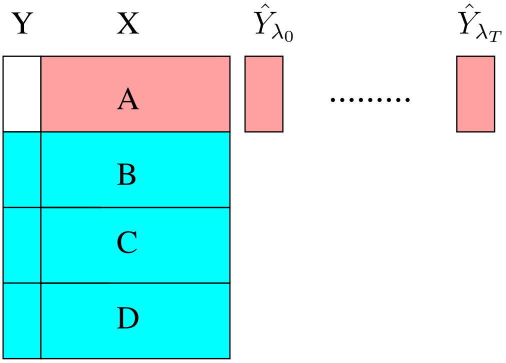
$$\lambda_j = \lambda_0 10^{\frac{-4j}{T-1}}, \quad 0 \leq j \leq 99$$

For elasticnet divide the grid values by α and for ridge one can divide by 0.01 or 0.001.

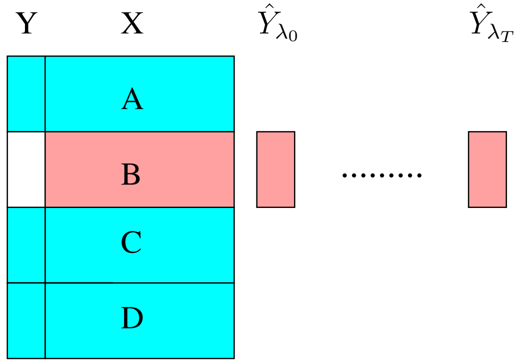
• Then *split the data in K groups.* Use $K-1$ groups to estimate for the first value of the grid and predict the last group to obtain $\hat{Y}$ for the remaining group denoted $\hat{Y}_{\lambda_0}$.



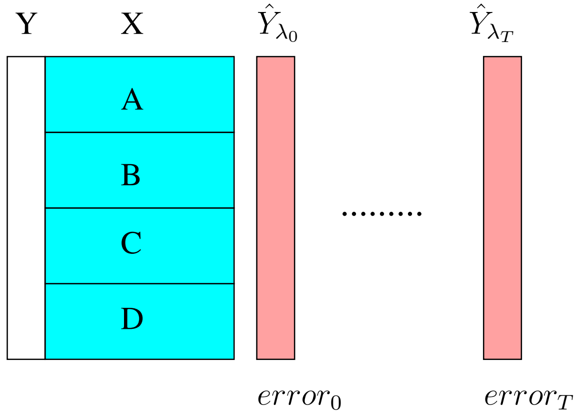
Do the same for the other values of λ.



Change the group you want to predict



• At the end, every one has been predicted for all the different λ from the grid. It is possible to evaluate the predicted error for all the values of the grid using $\|Y - \hat{Y}_{\lambda_j}\|^2$ (or other loss function)



and then just choose the best λ and use all the data to estimate the chosen estimator !

Chapter 7

Models comparison

7.1 Introduction

We have collected data $(X_i, Y_i)_{1,\cdots,n}$ where Y is a real an we have p potentially explanatory variables to predict Y. So far, we have at our disposal 5 algorithms to predict Y

- Least square
- Least square with variables selection (different methods are available)
- Ridge estimator
- Lasso estimator
- Elastic net estimator

Which one suits the best to our data ? It is important to remenber the graph where the estimation error decreases with the complexity (here the number of variables).

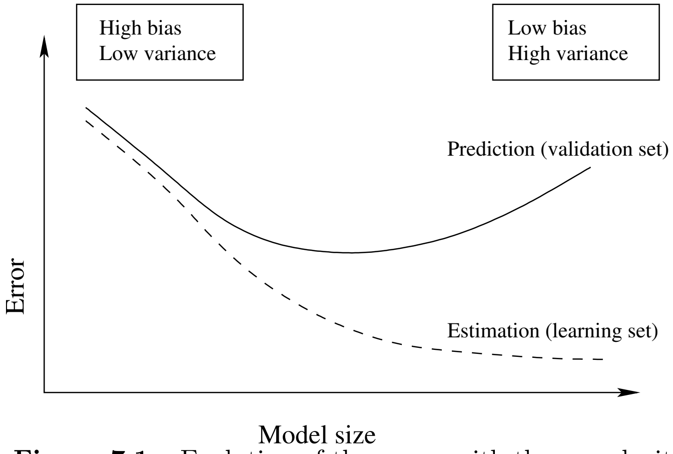


Figure 7.1 – Evolution of the errors with the complexity.

In our case, the LS estimator is the one having more variables and so the highest complexity compare to the 4 other algorithms. So on the estimation error, the LS will be on the rigth part of the graph. But do we overfit ?

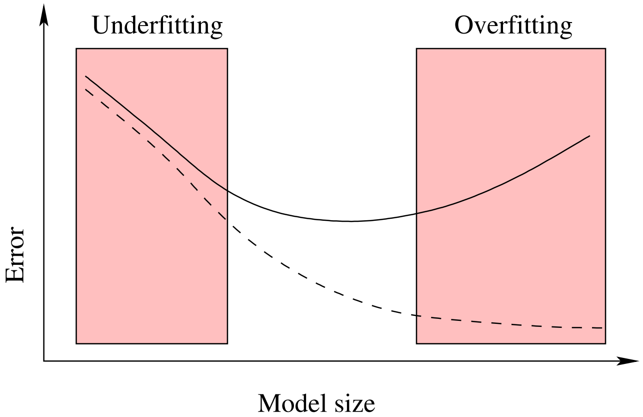
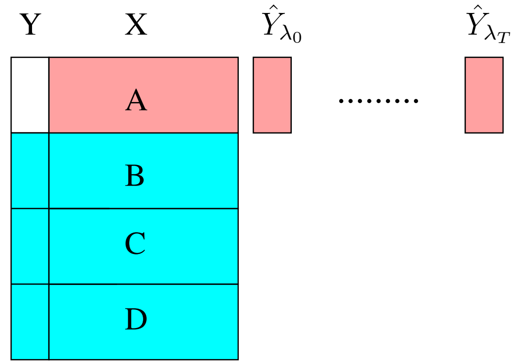


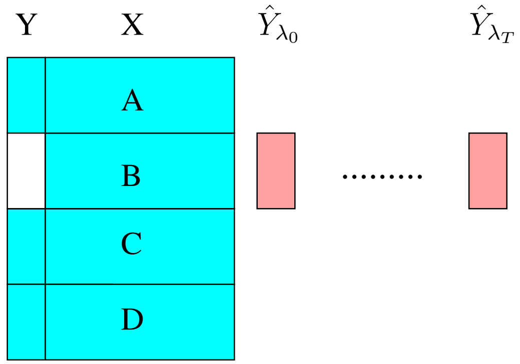
Figure 7.2 – Evolution of the errors with the complexity.

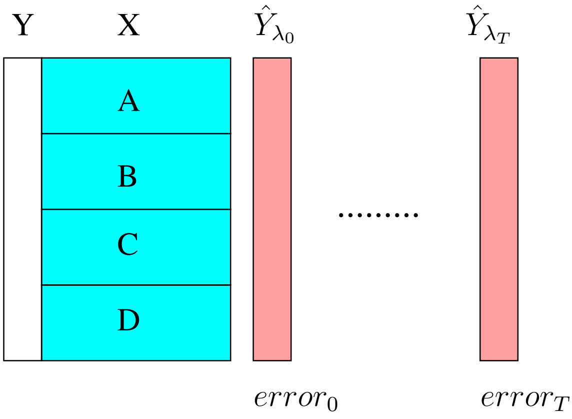
In order to answer that question, we need to estimate the prediction error and we are going to use again $K-$fold cross validation.

7.2 Cross validation

The data are splitted in K groups, we use $K-1$ groups to estimate the different algorithms and then use the last group for prediction. This is done K times.







We choose the method with the smallest error and estimate the corresponding parameters with all the data. It is interesting to note here that using penalised regression needs to select λ which is usually done in Cross Validation so in fact we usually perform a cross validation for selecting the parameters in a cross validation for selecting the best method.

7.3 Moving forward: feature ingeniering

you maay want to add variables by transforming the one you have such as taking the square or some polynom transformation (usually up to degree 3), you could use interaction by multiplying variables together $X_{12} = X_1 * X_2 ....$