Graduate J. Math. **1** (2016), 121 – 137

An overview on some regression models. Part I: Simple linear regression

GÉRARD GRÉGOIRE

Abstract

This article and its sequel form an introduction to the field of regression analysis. We start by presenting briefly a panorama of regression models: linear models, generalized linear models, nonparametric and semi-parametric regression models, nonlinear models. Then in the remainder of the text we focus on simple linear regression, which is the situation where the mean of a variable Y, i.e. the response variable, is depending linearly on another variable x_1 called the regressor. We are concerned with standard usual context, that is we assume, together with other conditions, that the error variable is gaussian. In this context we present estimates obtained by least squares methods and their basic properties. Then operational tools for statistical inference are designed: confidence and prediction intervals, significance tests and anova tables. The issue of diagnostic tests to detect observations which play particular roles in the regression is addressed in some detail: outliers, observations with high leverage effect and influent observations are concerned. Finally we indicate some ideas to deal with the situation where some of the model assumptions are not satisfied. It is worth emphasizing that all along the text we make a special effort to find a compromise between mathematical rigor and applied statistical concerns. We give precise proofs for specific issues that we consider of particular significance. Finally we illustrate both methods and results all along the text with one unique dataset analyzed by means of the R software.

MSC 2010. Primary 62J05; Secondary 62J10, 62J20

1 Introduction

The term regression gathers a set of statistical methods which aims at analyzing the relationship between a variable Y, the response variable, and independent variables, say $x_1, x_2, \ldots, x_p$. For simplicity, in this paper, we limit ourselves to the situation where Y is one-dimensional, but Y could as well be multidimensional. The response variable can be a quantitative variable as, for instance, it is the case with linear regression. Other regression models focus on qualitative Y: logistic regression is a classical example. The regression, together with classification, is one of the main tools for extracting information in datasets, and can be used in large datasets as in data mining or big data.

Regression methods are used in almost every domain: social sciences, life sciences including biology and medicine, environmental sciences, physical sciences and engineering. It is worth to note that the domain of modeling in economy has been a particularly fertile field for regression methods: the economists have not only extensively applied regression to their problems, but they also developed original methodological contributions to deal with particular situations (see for instance [8]).

Here we present some examples of applications in different fields, but numerous other ones could be outlined. Specialists of atmospheric pollution study by means of linear regression the relationship between the mean SO_2 concentration in a city and on one hand variables related to the size of the city and to the level of industrial activity, and on the other hand on meteorological variables such as wind speed, rainfall quantities, number of wet days, etc. Economists study by simple linear regression the relationship between the consumption and the income of a household, but design also more complicated regression models to model the behavior of the GNP of an economy as a linear function of investment, capital and labor amount. In marketing, logistic regression is used to segment the clientele, that is to define classes along the appetence of the customers for a product in relationship with variables such as age, marital situation, type of work, favorite pastime. In a similar way logistic regression is used in medical research for instance to estimate prostate cancer risk from variables such as PSA marker, results of biopsy, local examination and heredity.

Regression methods are designed to *explain* the behavior of Y as a function of input variables $x_1, \ldots, x_p$. They are intended to identify the input variables with a significant effect on Y, as well as the ones with no such effect, and to quantify the effect. They must be able also to *predict* the most probable value for Y given new values of x_i's, and to provide information on the accuracy of this prediction. This prediction ability is often used by engineers in order to control processes. For instance, in a paper mill bleaching of the pulp can

be achieved using chlorine dioxine. A regression analysis allows to relate the brightness grade to the quantity of chlorine dioxine used during the pulp cooking. Then the relationship can be used to control the brightness by choosing accurately the quantity of dioxine chlorine.

Finally let us mention that although regression methods are usually considered as general statistical methods, we can also find them as part of papers or books devoted to machine learning or statistical learning (see for instance [9], [10] and [11]). Loosely speaking the paradigm which is the basis of these learning works is that the experimenter is faced with a dataset (the training data) from which he has to learn relationships between some variables. We speak of unsupervised learning when the main variable of interest is unobserved, and supervised learning otherwise. Regression methods are supervised learning methods.

We start this paper by giving a global overview on various regression models that are commonly used by statisticians, depending on the type of data they are analyzing, and what questions they try to answer. Then in the second part of this introduction, we turn to the particular case of the *simple linear regression* model. We define precisely this model, list the statistical questions we are interested in, and present the dataset that we use in this paper to illustrate our methods and results.

1.1 An overview of classical regression models

Historians attribute to Legendre and Gauss in the beginning of the nineteenth century, the first mathematical work on linear regression. For the anecdote, the term "regression" goes back to Francis Galton who, in a study around 1875 comparing the sizes of fathers with the ones of their sons, observed that: *"... sons of tall fathers tend to be tall but not as tall as their fathers while sons of short fathers tend to be short but not as short as their fathers..."* and called this effect "a regression effect".

1.1.1 The simple linear regression model

We start with the simplest regression model, namely the *simple linear regression* which is the subject we develop after this introduction.
We are given a random variable Y whose mean depends in a linear way on the value of x_1, a nonrandom numeric variable:

$$\mathbb{E}(Y|\ x_1) = \beta_0 + \beta_1 x_1. \tag{1.1}$$

We write $\mu(x_1)$ the mean function $\mathbb{E}(Y|\ x_1)$. To be more precise we suppose that we can write

$$\begin{aligned} Y &= \mu(x_1) + \varepsilon && (1.2)\\ &= \beta_0 + \beta_1 x_1 + \varepsilon, && (1.3)\end{aligned}$$

with ε a random variable whose distribution is independent of x_1 and is centered, that is $\mathbb{E}(\varepsilon) = 0$. Note that this means in particular that the variance of Y doesn't depend on x_1. Thus we have $\mathbb{V}\mathrm{ar}(Y|x_1) = \mathbb{V}\mathrm{ar}(\varepsilon) = \sigma^2$(say).
Y is called the *response* variable (also dependent variable, explained variable).
x_1 is the *regressor* (also predictor, independent variable, explicative variable, input variable, covariable).
Given observations (x_{i1}, y_i), $i = 1, \ldots, n$, the statistician investigates the relationship, supposed to be linear, between the mean of Y and x_1.

We define the model in more details in the next section. At this stage, let's note that, even if the primary interest is clearly in the statistical inference about β_0 and β_1, σ^2, some other informations, such that for instance the accuracy of the model or the prediction intervals for Y, are worth investigating too.

1.1.2 The multiple linear regression model

The multiple linear regression is a generalization of the previous model: now the mean of Y is linearly dependent on $p \geq 1$ regressors $x_1, \ldots, x_p$:

$$\begin{aligned} \mu(x_1, \ldots, x_p) &= \mathbb{E}(Y|\ x_1, \ldots, x_p) \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \end{aligned}$$

More precisely we assume that we can write:

$$\begin{aligned} Y &= \mu(x_1, \ldots, x_p) + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \end{aligned} \tag{1.4}$$

where the distribution of ε doesn't depend on the regressors and $\mathbb{E}(\varepsilon) = 0$. Hence, as in the simple linear regression, the variance of Y, which is also the one of ε, doesn't depend on the regressors. We still denote σ^2 this variance.
Statisticians are often faced with situations where a response variable Y depends on several –possibly many– variables, and are interested in issues like the following ones. What are the variables with an effect on Y? How to model the effect of one variable when taking into account the effects of other ones? Is there any interaction between some variables? Are there variables which bring similar information about the response variable? Can we define a parcimonious model while preserving the information? etc. Multiple linear regression occupies a prominent place among the methods to tackle these multivariate issues.

Finally let us note that for sake of simplicity, we intentionnaly gave a restrictive definition of linear regression: (1.4) defines in fact a particular case of linear model. A polynomial model like the one defined by:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \varepsilon$$

is also a linear regression model, even though x_1 occurs nonlinearly. Indeed setting $\widetilde{x}_i = x_1^i$, $i = 1, \ldots, 3$ we get

$$Y = \beta_0 + \beta_1 \widetilde{x}_1 + \beta_2 \widetilde{x}_2 + \beta_3 \widetilde{x}_3 + \varepsilon.$$

In its most general form, a linear regression model is a model linear with respect to parameters β_i's. Hence in such a model the variable response Y may be related to the x_i's in a nonlinear way, but the relationship can be made linear with respect to other variables choosen in an adequate way.

1.1.3 Generalized linear model

What is called a generalized linear model, in short GLM, is a model where the dependence between the mean of the response variable Y and the regressors is defined through a link function. That is, it is not the mean that we model as a linear function of the regressors, but a function of the mean. Precisely we have:

$$g(\mu(x_1,\ldots,x_p)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

and we recall that $\mu(x_1,\ldots,x_p) = \mathbb{E}(Y|\ x_1,\ldots,x_p)$. We still have:

$$Y = \mu(x_1,\ldots,x_p) + \varepsilon$$

with $\mathbb{E}(\varepsilon) = 0$.
Let us observe that the variables ε don't play the same role as in linear regression. Their distributions may depend on the regressors and must belong to a particular class. Besides, while in linear regression the estimates are obtained by least squares method, it's the maximum likelihood method that is used in GLM models to get the estimates.
Two particular generalized linear models are popular: the logistic regression and the poissonian regression.

Logistic regression.
Let's suppose that the response variable is binary with

$$\begin{aligned} P(Y=1\,|\,x_1,\ldots,x_p) &= \pi(x_1,\ldots,x_p) \\ P(Y=0\,|\,x_1,\ldots,x_p) &= 1-\pi(x_1,\ldots,x_p). \end{aligned}$$

Hence we have:

$$\mu(x_1,\ldots,x_p) = \mathbb{E}(Y|\ x_1,\ldots,x_p) = \pi(x_1,\ldots,x_p).$$

The logistic model is defined by :

$$\log\frac{\mu(x_1,\ldots,x_p)}{1-\mu(x_1,\ldots,x_p)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \quad (1.5)$$

which means that the link function is given by:

$$g(u) = \log\frac{u}{1-u} = \text{logit}(u). \quad (1.6)$$

The logistic model is one of the models that are used to model the relationship between the probability of an event and a set of covariables.

Poissonian regression.
Let's consider the situation where, when the covariables are given by $x_1,\ldots,x_p$, Y is Poisson-distributed with mean $\mu(x_1,\ldots,x_p) = \exp(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)$. That is:

$$\log(\mu(x_1,\ldots,x_p)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

This means that the link function is given by:

$$g(u) = \log(u). \quad (1.7)$$

The poissonian regression is used to model the dependence between variables of counts and regressors.

1.1.4 Miscellaneous regression topics

Nonparametric and semiparametric models.
So far the regression models presented were parametric. That is, in these models, the relationship between the distribution of Y and covariables goes through parameters $\beta_0,\beta_1,\ldots,\beta_p$. This is also the case for the nonlinear model, see the end of subsection 1.1.4.
Nonparametric models are used when there is no argument to base the relationship on a particular parametric function. A nonparametric regression model is defined by:

$$Y = f(x_1,\ldots,x_p) + \varepsilon \quad (1.8)$$

where f is unknown, $\mathbb{E}(\varepsilon) = 0$ and the distribution of ε doesn't depend on the regressors values.
Statistical inference about the unknown function f was a major research subject for the twenty five last years and numerous methods were developped: kernels, local polynomials, spline, orthogonal projections, wavelets, etc.
Semiparametric models are models with both a parametric and a nonparametric part. An example of such a model is the Cox model which is popular in survival data analysis. Cox model is defined through the hazard rate function. Let's recall that the (instantaneous) hazard function is defined by $h(t) = f(t)/(1-F(t))$, where f is the probability density function and F the cumulative density function. In the Cox model the hazard function of an individual with covariables $x_1,\ldots,x_p$ is defined by:

$$h(t\,;\,x_1,\ldots,x_p) = h_0(t)\times\exp(\beta_1 x_1+\cdots+\beta_p x_p). \quad (1.9)$$

where the nonparametric part $h_0(\cdot)$ is to be estimated as well as the parameters $\beta_1,\ldots,\beta_p$. This model is fairly flexible, since the shape of the hazard function can be adjusted via the baseline hazard function h_0 and effects of covariables are taken into account as a multiplicative factor.

Regression models involving other characteristics than the mean.
So far we focused essentially to modeling the mean of the response variable Y. Just above the Cox model, see (1.9), is an example of regression model where the link between the response variable Y and the regressors $x_1,\ldots,x_p$ involve other probabilistic characteristics than the mean.
Another example, again in the field of survival analysis, is the one of AFT models (Accelerated Failure Times models). In such a model, the lifetime of an individual with covariables $x_1,\ldots,x_p$, which we denote by $T(x_1,\ldots,x_p)$, is distributed as a multiple, depending on $x_1,\ldots,x_p$, of a baseline lifetime T_0:

$$T(x_1,\ldots,x_p) \sim a(x_1,\ldots,x_p)T_0 \quad (1.10)$$

Along the choice of the parametric distribution of T_0, different regression models arise out. It must be noted that, in Cox model as well as in AFT models, statistical inference is achieved via the maximum likelihood methodology.

Nonlinear regression.
A nonlinear regression model is a parametric model where the mean function $\mu(x_1, \ldots, x_p)$ of the response variable Y is nonlinear in the parameters $\beta_0, \beta_1, \ldots, \beta_p$. That is

$$\begin{aligned} Y &= \mu(x_1, \ldots, x_p) + \varepsilon \\ &= f(x_1, \ldots, x_p; \beta_0, \ldots, \beta_p) + \varepsilon \end{aligned}$$

where $\mathbb{E}(\varepsilon) = 0$. In the most simple version of this model we assume that the distribution of ε is independent of the regressors $x_1, \ldots, x_p$. The nonlinearity of f is to be understood in the following way: there exists at least one β_i for which the derivative of f with respect to β_i depends on at least one of the parameters. For instance, the Michaelis-Menten model for enzyme kinetics uses one regressor and is defined by:

$$f(x_1; \beta_0, \beta_1) = \frac{\beta_1 x_1}{\beta_0 + x_1}.$$

We can check that $\dfrac{\partial f}{\partial \beta_0} = -\dfrac{\beta_1 x_1}{(\beta_0 + x_1)^2}$ and $\dfrac{\partial f}{\partial \beta_1} = \dfrac{x_1}{\beta_0 + x_1}$, which is consistent with the given definition of nonlinearity. The solution goes through local linearization of the function and local least squares.

1.2 Relevant issues in simple linear regression

First let's recall what was considered above. A linear regression is a model where a variable Y is depending on x_1 in the following way:

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon. \tag{1.11}$$

where ε is a random variable with zero mean, and distributed independently of x_1. The random variable Y is called the response variable, x_1 is nonrandom and called the regressor. From (1.11), it comes that the mean function $\mu(\cdot)$ satisfies:

$$\mu(x_1) = \mathbb{E}(Y|\ x_1) = \beta_0 + \beta_1 x_1.$$

In a practical statistical setting, we are given observations (x_{i1}, y_i), $i = 1, \ldots, n$, from which we are to draw statistical inference about β_0 and β_1. In a first step we look for estimates of β_0 and β_1. A very popular method to get these estimates is the *Least Squares* method which consists in minimizing the square errors sum:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2 \tag{1.12}$$

as a function of β_0 and β_1. See Fig. 1: the least squares method looks after the line for which the sum of squared lengths of vertical segments between each data point and its projection onto the line is minimum.

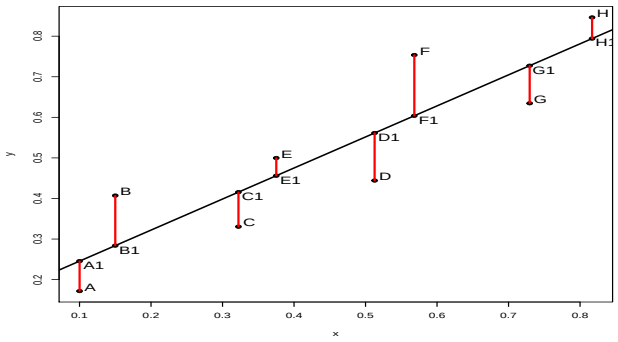

Fig. 1: The least squares methods minimizes the sum $AA_1^2 + BB_1^2 + \cdots + HH_1^2$.

By the least squares method we get an estimate for the model (1.11), that is $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and estimates for their precisions. With some conditions which are made precise in definition 2.1 below, this enables us to deal with the relevant statistical issues.

Some statistical issues that are of interest:

1. Is the regression of any interest? i.e. is $\widehat{\beta}_1$ significantly different from 0?
2. Can we give a confidence interval for β_1?
3. Can we give a measure for the extent to which the regression line explains the data?
4. Let x_1^0 be any new observation of x_1 and suppose that $Y^0 = Y(x_1^0)$ is not observed. Is it possible to give a confidence interval for $\mu(x_1^0) = \mathbb{E}(Y^0)$? How this confidence interval is related to the distance between x_1^0 and $(x_{i1})_{i=1,\ldots,n}$?
5. Is it possible to give a prediction interval for $Y^0 = Y(x_1^0)$?
6. Are there some data which greatly affect the estimated regression equation? Are there some data which are not well explained by the regression?

We provide answers to these issues in the following sections.

In this paper we rely on the R software to illustrate the methods and results. Readers who are not R users are invited to download the software (freely) from the cran website:

https://cran.r-project.org/

Further we use extensively the functions provided by the R library "car", which is an abbreviation for "Companion to Applied Regression". Note that a book by Fox and Weisberg with the same title [15] can help to use these functions, although most of them are easy to work with and don't need extensive documentation.
Through this paper we use the data set "trees". The R file contains variables Girth, Height and Volume for 31 felled black cherry trees. The file "trees" is available from the package "dataset" which can be downloaded when R is active. For illustration of simple linear regression we only consider Height and Volume and we provide below the 31 observations, sorted along the increasing values of Height.

	Height	Volume		Height	Volume
1	63	10.2	17	77	42.6
2	64	24.9	18	78	34.5
3	65	10.3	19	79	24.2
4	66	15.6	20	80	22.6
5	69	21.3	21	80	31.7
6	70	10.3	22	80	58.3
7	71	25.7	23	80	51.5
8	72	16.4	24	80	51.0
9	72	38.3	25	81	18.8
10	74	22.2	26	81	55.4
11	74	36.3	27	82	55.7
12	75	18.2	28	83	19.7
13	75	19.9	29	85	33.8
14	75	19.1	30	86	27.4
15	76	21.0	31	87	77.0
16	76	21.4			

Tab. 1: Observations of (Height,Volume) for 31 black cherry trees

This dataset is used all along the text to illustrate results and methods: all the given R outputs concern these data except in subsection 6.2.

We are to estimate the relationship between the response variable "Volume" and the regressor "Height" assuming this relationship is linear.

The R commands:

```
************************************
trees3<- trees[,2:3]
o<- order(trees3$Height)            # to  sort
trees3$Height<- trees3$Height[o]    # the file trees3
trees3$Volume<- trees3$Volume[o]    # along Height
plot(trees3,xlab="Height (ft)",ylab="Volume
(cubic ft)",xlim=c(60,90),ylim=c(5,85))
************************************
```

sort trees3 along increasing values of "Height" and provide Fig. 2.

2 Model, assumptions and Least Squares method

Here we set more precisely the model and assumptions.

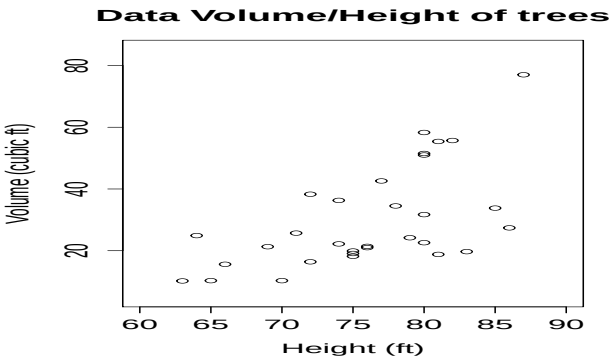

Fig. 2: Black cherry trees data

Definition 2.1. The Simple Linear Regression (SLR) model.
The simple linear regression model is defined in the following way.
We assume that, for $i = 1, 2, \ldots, n$, *we are given* x_{i1} *and we observe* y_i *a value of* Y_i *satisfying:*

$$Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i \tag{2.1}$$

where

- x_{i1}, $i = 1, \ldots, n$, *are deterministic values of* x_1,
- ε_i *are centered i.i.d. (independent and identically distributed) gaussian variables.*

In other words $(\varepsilon_i)_{i=1,\ldots,n}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$-distributed variables. It follows that the variables Y_i are independent and $\mathcal{N}(\beta_0 + \beta_1 x_{i1}, \sigma^2)$-distributed.
Let us emphasize that we have in fact three parameters to investigate, β_0, β_1 and σ^2.

Definition 2.2. Least squares estimates.
$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)'$ *is a Least Squares estimate (LSE) of* $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ *when for any* $(\tilde{\beta}_0, \tilde{\beta}_1)$*:*

$$S(\widehat{\beta}_0, \widehat{\beta}_1) \leq S(\tilde{\beta}_0, \tilde{\beta}_1),$$

where S *is defined by (1.12).*

Hereafter we note:
$\mathbf{1} = (1, 1, \ldots, 1)' \in \mathbb{R}^n$
$\boldsymbol{x}_1 = (x_{11}, x_{21}, \ldots, x_{n1})'$
$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$
$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$.

Usual empirical moments for $\boldsymbol{x}_1$ and $\boldsymbol{y}$ are involved throughout the text, they are defined by:

$$\bar{\boldsymbol{x}}_1 = 1/n \sum x_{i1} \text{ and } \bar{\boldsymbol{y}} = 1/n \sum y_i$$
$$var(\boldsymbol{x}_1) = 1/n \sum (x_{i1} - \bar{\boldsymbol{x}}_1)^2$$
$$cov(\boldsymbol{y}, \boldsymbol{x}_1) = \frac{1/n \sum (y_i - \bar{\boldsymbol{y}})(x_{i1} - \bar{\boldsymbol{x}}_1)}{1/n \sum (x_{i1} - \bar{\boldsymbol{x}}_1)^2},$$

where the sums are understood to be over i varying from 1 to n.

Theoretical mean, variance and covariance will be denoted by $\mathbb{E}$, $\mathbb{V}$ar and $\mathbb{C}$ov.

In the standard setting where the values of x_1 are not all the same, the Least Squares problem has a unique solution.

Theorem 2.3. *Suppose that $\boldsymbol{x}_1 = (x_{11}, x_{21}, \ldots, x_{n1})'$ differs from $c \cdot \mathbf{1}$, then there exists a Least Squares estimate. The estimate is unique and given by:*

$$\begin{cases} \widehat{\beta}_1 = \dfrac{cov(\boldsymbol{y}, \boldsymbol{x}_1)}{var(\boldsymbol{x}_1)} \\ \widehat{\beta}_0 = \bar{\boldsymbol{y}} - \widehat{\beta}_1 \bar{\boldsymbol{x}}_1. \end{cases} \tag{2.2}$$

Proof. The proof is straightforward. The function S, as defined in (1.12) is $\mathcal{C}_2$ and the Hessian matrix is given by:

$$\mathcal{H} = 2n \begin{bmatrix} 1 & \bar{\boldsymbol{x}}_1 \\ \bar{\boldsymbol{x}}_1 & \dfrac{1}{n}\sum x_{i1}^2 \end{bmatrix}. \tag{2.3}$$

Due to the assumption $\boldsymbol{x}_1 \neq c \cdot \mathbf{1}$ (i.e. $var(\boldsymbol{x}1) \neq 0$), $\mathcal{H}$ is positive definite and it follows that S is a convex function. Further $\partial S/\partial \beta_0 = 0$ yields $\bar{\boldsymbol{y}} = \beta_1 \bar{\boldsymbol{x}}_1 + \beta_0$ and $\partial S/\partial \beta_1 = 0$ leads to $cov(\boldsymbol{y}, \boldsymbol{x}_1) - \beta_1 var(\boldsymbol{x}_1) = 0$. The result follows. □

In the rest of the paper we use:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} \tag{2.4}$$

$$e_i = y_i - \widehat{y}_i. \tag{2.5}$$

The terms $\widehat{y}_i$ are called *fitted values* (*fitted responses*, *adjusted values*,...), and the e_i are the *residuals*. We can write

$$\begin{aligned} \varepsilon_i &= y_i - (\beta_0 + \beta_1 x_{i1}) \\ e_i &= y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1}). \end{aligned}$$

This suggests that we can see the residual e_i as an estimate of the error ε_i.
It is convenient also to consider vectors:

$$\boldsymbol{e} = (e_1, e_2, \ldots, e_n)' \tag{2.6}$$

$$\widehat{\boldsymbol{y}} = (\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_n)'. \tag{2.7}$$

Clearly we have $\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$.
It can be shown that the vector $\widehat{\boldsymbol{y}}$ is the projection of $\boldsymbol{y}$, in $\mathbb{R}^n$, onto the subspace generated by the vectors $\mathbf{1}$ and $\boldsymbol{x}_1$. More details on these geometrical aspects will be given in part II. A consequence is that $\sum e_i = 0$ and we can estimate σ^2 by:

$$\widehat{\sigma^2} = \frac{\sum e_i^2}{n-2} \tag{2.8}$$

where the term $(n-2)$ in the denominator of (2.8) follows from properties given the next section.

3 Basic properties of the Least Squares Estimate (LSE)

Theorem 3.1. *For the model 2.1 $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)'$ is gaussian and:*

$$\widehat{\beta}_0 \sim \mathcal{N}(\beta_0, \frac{\sigma^2}{n}\big[1 + \frac{\bar{\boldsymbol{x}}_1^2}{var(\boldsymbol{x}_1)}\big]) \tag{3.1}$$

$$\widehat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{n}\frac{1}{var(\boldsymbol{x}_1)}) \tag{3.2}$$

and $\mathbb{C}ov(\widehat{\beta}_0, \widehat{\beta}_1) = -\dfrac{\sigma^2}{n}\dfrac{\bar{\boldsymbol{x}}_1^2}{var(\boldsymbol{x}_1)}$.
Further $(n-2)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$ and $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ are independently distributed.

Proof. From (2.2) we can see that $\widehat{\boldsymbol{\beta}}$ is a linear function of $\boldsymbol{y}$, and consequently of $\boldsymbol{\varepsilon}$. Hence the first part of the theorem follows from elementary calculations.
For the second part we use a variant of Cochran's theorem. The vector $\boldsymbol{e}/\sigma$ is the orthogonal projection $P_{V_1^\perp}\, \boldsymbol{\varepsilon}/\sigma$ of $\boldsymbol{\varepsilon}/\sigma$ on the subspace $V_1^\perp$ orthogonal complementary of the subspace $V_1 = \mathcal{V}\{\mathbf{1}, \boldsymbol{x}_1\}$ generated by $\mathbf{1}$ and $\boldsymbol{x}_1$. Applying Cochran's theorem, this implies that $\dfrac{1}{\sigma^2}||\boldsymbol{e}||^2$ is χ^2_{n-2}-distributed. Besides $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ can be written $L\boldsymbol{\varepsilon}/\sigma$ and L satisfies $LP_{V_1^\perp} = 0$. It then follows that $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{e}$ are uncorrelated and thus also independent. □

Remark 3.2. It follows from Theorem 3.1 that $\widehat{\boldsymbol{\beta}}$ is unbiased and the same holds true for $\widehat{\sigma}^2$. We have also $\mathbb{V}\text{ar}(\widehat{\sigma}^2) = 2\sigma^4/(n-2)$.
The least squares estimate $\widehat{\boldsymbol{\beta}}$ is also a maximum likelihood estimate. This follows from the fact that, up to an additive constant, the log-likelihood function can be written:

$$-\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i - \beta_0 - \beta_1 x_{i1})^2.$$

Proposition 3.3. *Assume that $n \cdot var(\boldsymbol{x}_1) \to \infty$ and that $\bar{\boldsymbol{x}}_1$ is bounded, then $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ converge in probability respectively to $\boldsymbol{\beta}$ and σ^2.*

Proof. From Theorem 3.1 and Remark 3.2 the result is clear for $\widehat{\sigma}^2$. By the assumptions the variances of $\widehat{\boldsymbol{\beta}}$ are going to 0. Since the estimates are unbiased the result follows. □

Note that the given conditions are only sufficient. The sequence of experimental designs defined by $\boldsymbol{x}_1(n) = (1, 2, \ldots, n)'$ results in convergent estimates although $\bar{\boldsymbol{x}}_1(n)$ is unbounded. An example of nonconvergence case is given by $\boldsymbol{x}_1(n) = (1, 1/2, \ldots, 1/n)'$: in this case $var(\boldsymbol{x}_1(n))$ is vanishing way too fast. But clearly, this example is not of much interest from a statistical point of view and in standard realistic situations the convergence holds true.

Theorem 3.4. *In the class of unbiased estimates of $\boldsymbol{\beta}$, the LSE $\widehat{\boldsymbol{\beta}}$ has the minimum variance. Moroever $\widehat{\boldsymbol{\beta}}$ is asymptotically efficient, i.e. achieves asymptotically the best possible variance.*

We postpone the proof of this result to part II devoted to multiple linear regression.
The first part of this result is often rephrased by saying that $\widehat{\boldsymbol{\beta}}$ is BUE (Best Unbiased Estimate).

For our dataset the R commands:

```
**************************************
reg_trees3<- lm(Volume~Height,data=trees3)
summary(reg_trees3)
**************************************
```

provides:

```
**************************************
Call:
lm(formula = Volume ~ Height, data = trees3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.1236    29.2731  -2.976 0.005835
Height        1.5433     0.3839   4.021 0.000378
**************************************
```

Thus $\widehat{\beta}_0 = -87.1236$ and $\widehat{\beta}_1 = 1.5433$. We will see in the end part of section 4.1 the interpretation of the other informations given in the table above.
To plot the regression line together with the data, we only run:

```
**************************************
plot(trees3,xlab="Height (ft)",ylab="Volume
(cubic ft)",xlim=c(60,90),ylim=c(5,85))
abline(reg_trees3)
**************************************
```

The result is given by Fig. 3.

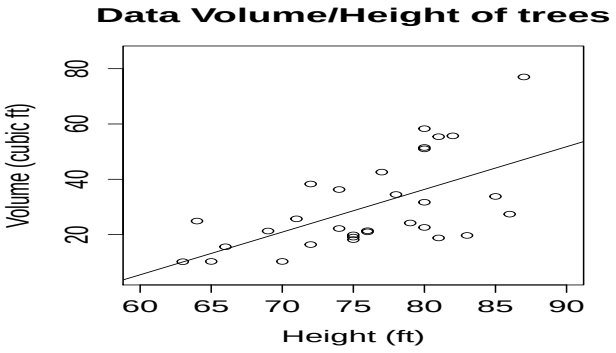

Fig. 3: Black cherry trees data. Observations and regression line for (Height,Volume)

It is clear that we can't use directly the above results to perform statistical inference on a dataset since σ^2 is usually unknown. Hence we substitute $\widehat{\sigma^2}$ to σ^2 in (3.1) and (3.2) to get natural estimates s_0^2 and s_1^2 for the variances of $\widehat{\beta}_0$ and $\widehat{\beta}_1$. The standard deviation estimates s_0 and s_1 are given by:

$$s_0 = \sqrt{\frac{\widehat{\sigma^2}}{n}\left[1+\frac{\bar{\boldsymbol{x}}_1^2}{var(\boldsymbol{x}_1)}\right]} \tag{3.4}$$

$$s_1 = \sqrt{\frac{\widehat{\sigma^2}}{n}\frac{1}{var(\boldsymbol{x}_1)}}. \tag{3.5}$$

By use of Theorem 3.1, this gives us a modification of (3.1) and (3.2) which can be used for inference purposes:

Corollary 3.5. *With an appropriate rescaling, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are Student distributed with $n-2$ degrees of freedom. More precisely:*

$$\frac{\widehat{\beta}_0-\beta_0}{s_0} \sim T_{n-2} \tag{3.6}$$

$$\frac{\widehat{\beta}_1-\beta_1}{s_1} \sim T_{n-2}. \tag{3.7}$$

Proof. $(\widehat{\beta}_0-\beta_0)/s_0$ can be rewritten:

$$\frac{\widehat{\beta}_0-\beta_0}{\sqrt{\frac{\sigma^2}{n}\left[1+\frac{\bar{\boldsymbol{x}}_1^2}{var(\boldsymbol{x}_1)}\right]}}\Bigg/\sqrt{\frac{1}{n-2}\left(\frac{(n-2)\widehat{\sigma^2}}{\sigma^2}.\right)} \tag{3.8}$$

From Theorem 3.1 it comes that the left part of this expression is $\mathcal{N}(0,1)$-distributed and independent from the right part. Since $\dfrac{(n-2)\,\widehat{\sigma^2}}{\sigma^2}$ is χ^2_{n-2}-distributed, we get the result for $\widehat{\beta}_0$. The proof for $\widehat{\beta}_1$ is similar. □

Test procedures, confidence intervals (CI) and prediction intervals (PI) will be defined in the next section via this result.

4 Statistical inference

4.1 Tests procedures for the regression coefficients

Building tests procedures for hypotheses on β_0 or β_1 follows from corollary 3.5 in a straightforward way.
In this paragraph as well as the rest of the paper we denote $t_n(\alpha)$ the $100\times(1-\alpha)\%$-quantile of the Student distribution with n degrees of freedom, that is:

$$P(T_n > t_n(\alpha)) = 1-\alpha.$$

Proposition 4.1 (**Two-tailed t-test.**). *The test of H_0 : $\beta_1=\beta_1^0$ against H_1 : $\beta_1\neq\beta_1^0$ is defined by the critical region:*

$$|\widehat{\beta}_1-\beta_1^0| > t_{n-2}(\alpha/2)\cdot s_1$$

and if $t=\dfrac{\widehat{\beta}_1-\beta_1^0}{s_1}$ the $p-$value is given by $P(|T_{n-2}|>|t|) = 2\,P(T_{n-2}>|t|)$.

Proposition 4.2 (**One-tailed t-test.**). *The test of* $H_0 : \beta_1 = \beta_1^0$ *against* $H_1 : \beta_1 > \beta_1^0$ *is defined by the critical region:*

$$\widehat{\beta}_1 > \beta_1^0 + t_{n-2}(\alpha) \cdot s_1$$

and the $p-$*value is given by* $P(T_{n-2} > t)$.

For the alternative hypothesis $H_1 : \beta_1 < \beta_1^0$ the critical region is given by $\widehat{\beta}_1 < \beta_1^0 - t_{n-2}(\alpha) \cdot s_1$ and the $p-$value is $P(T_{n-2} < t)$.

The tests can be performed on our data set using the information given by the R command "summary": see the R output just after the statement of Theorem 3.4. We obtained:

$s_0 = 29.2731 \quad \widehat{\beta}_0/s_0 = -2.976 \quad P(|T_{29}| > 2.976) = 0.005835,$
$s_1 = 0.3839 \quad \widehat{\beta}_1/s_1 = 4.021 \quad P(|T_{29}| > 4.021) = 0.000378.$

Thus the p-value for the two-tailed test of $\beta_0 = 0$ (resp. $\beta_1 = 0$) is equal to 0.005835 (resp. 0.000378). This means that we reject the null-hypothesis for β_0 as well as for β_1.
For the one-tail test of $\beta_1 = 0$ against $\beta_1 > 0$ the p-value is 0.000189 and we again reject H_0.
We will see in the next section another formulation of this test using anova analysis.

4.2 Confidence intervals (CI) for regression coefficients

Using corollary 3.5 with s_0 and s_1 as defined by (3.4) and (3.5) we get $100 \times (1-\alpha)\%$ confidence intervals for β_0 and β_1.

Proposition 4.3. *The intervals given by:*

$$\left[\widehat{\beta}_0 - t_{n-2}(\alpha/2) \cdot s_0,\ \widehat{\beta}_0 + t_{n-2}(\alpha/2) \cdot s_0\right]$$
$$\left[\widehat{\beta}_1 - t_{n-2}(\alpha/2) \cdot s_1,\ \widehat{\beta}_1 + t_{n-2}(\alpha/2) \cdot s_1\right]$$

are $100 \times (1-\alpha)\%$ *confidence intervals for* β_0 *and* β_1.

Note that we write as well $\widehat{\beta}_0 \pm t_{n-2}(\alpha/2)s_0$ for the β_0 CI, and there is an analogous expression for β_1.

Confidence intervals for regression coefficients are provided by R using `confint` command. For our dataset we get:

```
**************************************
confint(reg_trees3,level=0.90)
                    5 %       95 %
(Intercept) -136.8623664 -37.384861
Height         0.8911071   2.195592
**************************************
```

Thus we see that the 90% CI for β_0 is $[-136.86,\ -37.38]$ and the one for β_1 is $[0.89,\ 2.20]$.

The level argument is optional: when omitted, the default level 95% is applied.

```
**************************************
 confint(reg_trees3)
                    2.5 %     97.5 %
(Intercept) -146.993871 -27.253357
Height         0.758249   2.328451
**************************************
```

4.3 Confidence intervals (CI) for the mean

Suppose that from observations (x_{i1}, y_i), $i = 1, \ldots, n$, we obtained an estimation for the linear model $Y = \beta_0 + \beta_1 x_1 + \varepsilon$ with usual assumptions and let x_1^0 be any new value for the variable x_1. We are to design a $100 \times (1-\alpha)\%$-CI for the mean $\mu(x_1^0) = \beta_0 + \beta_1 x_1^0$.
From the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ we obtained using the observations at hand, it is natural to set as pointwise estimate for the mean: $\widehat{\mu(x_1^0)} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1^0$.
This estimate is unbiased and, using results of Theorem 3.1, some elementary calculations show that its variance is given by:

$$\frac{\sigma^2}{n}\left[1 + \frac{(x_1^0 - \bar{\boldsymbol{x}}_1)^2}{var(\boldsymbol{x}_1)}\right]$$

which we estimate substituting again $\widehat{\sigma^2} = \sum e_i^2/(n-2)$ to σ^2.
We then get a $100 \times (1-\alpha)\%$-CI for $\mu(x_1^0)$.

Proposition 4.4. *The interval given by:*

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_1^0 \pm t_{n-2}(\alpha/2)\sqrt{\frac{\widehat{\sigma^2}}{n}\left[1 + \frac{(x_1^0 - \bar{\boldsymbol{x}}_1)^2}{var(\boldsymbol{x}_1)}\right]} \quad (4.1)$$

is a $100 \times (1-\alpha)\%$ *confidence interval for* $\mu(x_1^0)$.

We note that as x_1^0 is going far from $\bar{\boldsymbol{x}}_1$, the CI is becoming increasingly large.

4.4 Prediction intervals (PI)

Now, given x_1^0 a new value of x_1 as done above, we are to build an interval which contains $Y^0 = Y(x_1^0)$ with probability $1-\alpha$. The best pointwise prediction of Y^0 is its mean which can be estimated by:

$$\widehat{Y^0} = \widehat{\mu(x_1^0)} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1^0.$$

The $100 \times (1-\alpha)\%$ prediction interval (PI) will be in the form $\widehat{Y^0} \pm \zeta$ where ζ satisfies

$$\begin{aligned} P(\widehat{Y^0} - \zeta \le Y^0 \le \widehat{Y^0} + \zeta) &= 1-\alpha \qquad \text{i.e. :} \\ P(|\widehat{Y^0} - Y^0| \le \zeta) &= 1-\alpha. \end{aligned}$$

$\widehat{Y^0}$ and Y^0 are independant gaussian variables with same means. Consequently $\widehat{Y^0} - Y^0$ is gaussian with zero mean and its variance satisfies:

$$\begin{aligned} \mathbb{V}\text{ar}(\widehat{Y^0} - Y^0) &= \mathbb{V}\text{ar}(\widehat{Y^0}) + \mathbb{V}\text{ar}(Y^0) \\ &= \frac{\sigma^2}{n}\left[1 + \frac{(x_1^0 - \bar{\boldsymbol{x}}_1)^2}{var(\boldsymbol{x}_1)}\right] + \sigma^2. \end{aligned}$$

Finally we replace σ^2 by $\widehat{\sigma^2}$ and we get the $100 \times (1-\alpha)\%$-PI for Y^0.

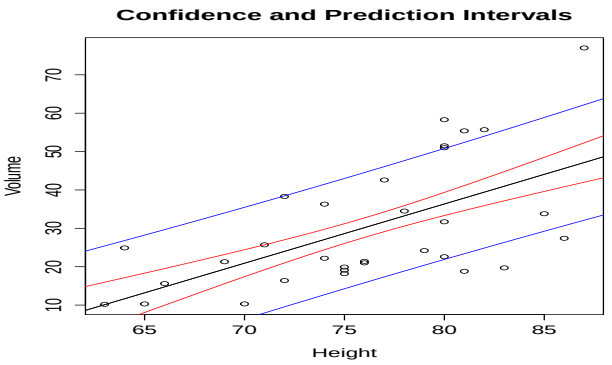

Fig. 4: Black cherry trees data. 70% confidence intervals for the mean, and 70% prediction intervals in the regression of Volume on Height. The black color line is the estimated mean function. The narrow band (red color) is related to CI intervals for the mean, and the large band (blue color) is related to PI for the variable "Volume".

Proposition 4.5. *The interval given by:*

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_1^0 \pm t_{n-2}(\alpha/2)\sqrt{\frac{\widehat{\sigma^2}}{n}[1 + n + \frac{(x_1^0 - \bar{\boldsymbol{x}}_1)^2}{var(\boldsymbol{x}_1)}]} \tag{4.2}$$

is a $100\times(1-\alpha)\,\%$ *prediction interval for* $Y^0 = Y(x_1^0)$.

Notice that the shape of CI and PI intervals should be roughly similar, but the bands defined by the endpoints of the intervals are wider for PI-intervals and show less curvature than CI-intervals. See below these curves estimated for our dataset.

4.5 Plots of CI and PI intervals using R for black cherries trees data

We use the function `ci.plot()` of the library RcmdrPlugin.HH to get the plot given by Fig. 4. For example the higher curve is the graphical representation of the dots $(Height, \widehat{Y}(Height))$. The interpretation of other curves is analogous except the central line which is the regression line. The set between the lower and higher curves is named *prediction band*, and the one defined by the two other curves *confidence bands*. Let's emphasize that confidence and prediction bands are not true bands. The correct interpretation is only pointwise and no information is given by these curves on the joint distributions.

```
**************************************
library(RcmdrPlugin.HH)
ci.plot(reg_trees3,main="95% confidence
  and prediction intervals for reg_trees3")
**************************************
```

It is possible, but a bit less straightforward, to use the standard R function `predict()`:

```
**************************************
> new<- data.frame(Height<-seq(60,90,0.5))
> pred.w.plim <- predict(reg_trees3, new,
    interval = "prediction",level=0.70)
> pred.w.clim <- predict(reg_trees3, new,
    interval = "confidence",level=0.70)
> plot(trees3,xlab="Height",ylab="Volume",
  main="Confidence and Prediction Intervals")
> matplot(new$Height, cbind(pred.w.clim,
    pred.w.plim), lty =c(1,1,1,1,1,1),
col=c("black","red","red","black","blue",
"blue"),type="l",ylab="Volume",add=TRUE)
**************************************
```

5 Variance decomposition. ANOVA table and R^2

5.1 Anova table

Let us remind that $\widehat{\boldsymbol{y}}$ is the orthogonal projection of $\boldsymbol{y}$ on the subspace $V_1 = \mathcal{V}\{\mathbf{1}, \boldsymbol{x}_1\}$ generated by $\mathbf{1}$ and $\boldsymbol{x}_1$, and $\bar{\boldsymbol{y}}$ the projection of $\boldsymbol{y}$ on $\mathcal{V}\{\mathbf{1}\}$. Besides, from $\sum e_i = 0$ it follows that $\widehat{\bar{\boldsymbol{y}}} = \bar{\boldsymbol{y}}$. So we can write:

$$\sum y_i^2 = \sum \widehat{y}_i^2 + \sum e_i^2 \tag{5.1}$$

and:

$$\sum (y_i - \bar{y})^2 = \sum (\widehat{y}_i - \bar{y})^2 + \sum e_i^2. \tag{5.2}$$

We define the following notation:

$$\begin{aligned} SS_{Tot} &= \sum (y_i - \bar{y})^2, \\ SS_{reg} &= \sum (\widehat{y}_i - \bar{y})^2, \\ SS_{res} &= \sum e_i^2. \end{aligned}$$

These quantities are called sums of squares. Thus (5.2) can be rewritten:

$$SS_{Tot} = SS_{reg} + SS_{res}, \tag{5.3}$$

equation which is called the *Sum of Squares decomposition*.
Dividing by n turns (5.3) into:

$$var(\boldsymbol{y}) = var(\widehat{\boldsymbol{y}}) + var(e) \tag{5.4}$$

which is called the *variance decomposition formula*.

Proposition 5.1. SS_{res} *and* SS_{reg} *are independently distributed with* $SS_{res}/\sigma^2 \sim \chi^2_{n-2}$, *and under* H_0 : $\beta_1 = 0$, $SS_{reg}/\sigma^2 \sim \chi^2_1$.

Proof. We already know that $(n-2)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$. As $(n-2)\widehat{\sigma}^2/\sigma^2 = SS_{res}/\sigma^2$ we have a part of the result.
Besides we use again the Cochran's theorem. Let

$$\widetilde{\boldsymbol{y}} = \frac{1}{\sigma}(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{x}_1))$$

where
$\boldsymbol{\mu}(\boldsymbol{x}_1) = (\mu(x_{11}), \mu(x_{21}), \ldots, \mu(x_{n1}))'$.
It is clear that $\widetilde{\boldsymbol{y}} \sim \mathcal{N}(0, \mathbb{1}_n)$. Hence it follows that

the projection $P_{V_1}(\widetilde{\boldsymbol{y}})$ is $\mathcal{N}(0, P_{V_1})$distributed, where $V_1 = \mathcal{V}\{\mathbf{1}, \boldsymbol{x}_1\}$ is the subspace generated by $\mathbf{1}$ and $\boldsymbol{x}_1$. Since $\boldsymbol{\mu}(\boldsymbol{x}_1) = \beta_0\mathbf{1} + \beta_1\boldsymbol{x}_1$, we have $P_{V_1}(\widetilde{\boldsymbol{y}}) = \frac{1}{\sigma}(\widehat{\boldsymbol{y}} - \boldsymbol{\mu}(\boldsymbol{x}_1))$.
The projection $P_{\mathbf{1}}(\widetilde{\boldsymbol{y}})$ on $\mathcal{V}\{\mathbf{1}\}$ is given by $\frac{1}{\sigma}(\bar{\boldsymbol{y}} - \mu(\bar{\boldsymbol{x}}_1)\mathbf{1})$. This allows to write:

$$P_{V_1}(\widetilde{\boldsymbol{y}}) - P_{\mathbf{1}}(\widetilde{\boldsymbol{y}}) = \frac{1}{\sigma}(\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\mathbf{1}) - \frac{1}{\sigma}(\boldsymbol{\mu}(\boldsymbol{x}_1) - \mu(\bar{\boldsymbol{x}}_1)\mathbf{1}).$$

Since $P_{V_1}P_{\mathbf{1}} = P_{\mathbf{1}}P_{V_1} = P_{\mathbf{1}}$, $P_{V_1} - P_{\mathbf{1}}$ is the projection onto the subspace orthogonal to $\mathbf{1}$ inside V_1,and its rank is one. In addition $(P_{V_1} - P_{\mathbf{1}})(\widetilde{\boldsymbol{y}})$ is $\mathcal{N}(0, P_{V_1} - P_{\mathbf{1}})$ distributed, and thus the distribution of $\frac{1}{\sigma}(\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\mathbf{1})$ is $\mathcal{N}(\frac{1}{\sigma}(\boldsymbol{\mu}(\boldsymbol{x}_1) - \mu(\bar{\boldsymbol{x}}_1)\mathbf{1}), P_{V_1} - P_{\mathbf{1}})$.
Finally note that, under $H_0 : \{\beta_1 = 0\}$, $\boldsymbol{\mu}(\boldsymbol{x}_1) = \mu(\bar{\boldsymbol{x}}_1)\mathbf{1}$, and consequently $\frac{1}{\sigma}(\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\mathbf{1}) = (P_{V_1} - P_{\mathbf{1}})(\widetilde{\boldsymbol{y}})$ and $\|\frac{1}{\sigma}(\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\mathbf{1})\|^2 = \|(P_{V_1} - P_{\mathbf{1}})(\widetilde{\boldsymbol{y}})\|^2 \sim \chi^2_1$. Thus we have $\frac{1}{\sigma^2}\sum(\widehat{y}_i - \bar{\boldsymbol{y}})^2 = SS_{reg}/\sigma^2 \sim \chi^2_1$. Moreover, by orthogonality of $\boldsymbol{e}$ and $(\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\mathbf{1})$, SS_{res} and SS_{reg} are independent. □

In the ANOVA table MS (Mean Square) are SS (Sum of squares as defined just after (5.2)) normalized by their degrees of freedom and we use notation:

$$\begin{aligned} MS_{reg} &= SS_{reg}/1 \\ MS_{res} &= SS_{res}/(n-2). \end{aligned}$$

Corollary 5.2. *Under $H_0 : \beta_1 = 0$, the rate $F = MS_{reg}/MS_{res}$ is $F_{1,n-2}$ distributed, where $F_{1,n-2}$ denotes the Fisher distribution with 1 and $n-2$ degrees of freedom.*

Remark 5.3. The test based on F and the one using $\widehat{\beta}_1/s_1$ are identical: it can be proven that $F = (\widehat{\beta}_1/s_1)^2$. Recall also that $F_{1,\nu}$ may be seen as the distribution of the square of a T_ν-distributed variable. This test is displayed by the ANOVA table given below.

Source	SS	d.f.	MS	F	p-value
Regression	SS_{reg}	1	MS_{reg}	$F = \frac{MS_{reg}}{MS_{res}}$	$P(F_{1,n-2} > F)$
Error	SS_{res}	n-2	MS_{res}		
Total	SS_{Tot}	n-1			

where **SS** stands for **Sum of Squares**, **MS** for **Mean Squares**, and we recall that $SS_{reg} = \sum(\widehat{y}_i - \bar{y})^2$, $MS_{reg} = \frac{SS_{reg}}{1}$, $SS_{res} = \sum e_i^2$, $MS_{res} = \frac{SS_{res}}{n-2}$ and $SS_{Tot} = \sum(y_i - \bar{y})^2$.

The Anova table provided by R for our dataset trees3 is given by:

```
**************************************
anova(reg_trees3)
Analysis of Variance Table
Response: Volume
          Df Sum Sq Mean Sq F value    Pr(>F)
Height     1 2901.2 2901.19  16.165 0.0003784
Residuals 29 5204.9  179.48
**************************************
```

Hence we get that $SS_{reg} = 2901.2 = MS_{reg}$ and $SS_{res} = 5204.9$ with 29 degrees of freedom. Consequently $MS_{res} = 5204.9/29 = 179.48$. Recall that we get above t-value for Height $\frac{\widehat{\beta}_1}{s_1} = 4.021$. It can be checked that the values of $\left(\frac{\widehat{\beta}_1}{s_1}\right)^2$ and $\frac{MS_{reg}}{MS_{res}}$ up to computational approximations are actually the same.

5.2 The R-squared and adjusted R-squared coefficients

The R-squared and adjusted R-squared are indices which measure the accuracy of fitting the data by linear regression.

Definition 5.4. The R-squared coefficient is defined by:

$$R^2 = \frac{var(\widehat{\boldsymbol{y}})}{var(\boldsymbol{y})} \tag{5.5}$$

and is usually called *coefficient of determination*.

Thus the R^2 coefficient gives the variance part explained by the regression. R^2 provides an information about the extent to which the points of the scatterplot are close to the regression line. The larger the value of R^2, the closer to the regression line are the points. Some properties:

- $0 \leq R^2 \leq 1$. The larger the value of R^2, the better the goodness of fit.
- $R^2 = 0$ means that $\widehat{\beta}_1 = 0$, that is the estimated regression line is horizontal. In other words the simple linear regression doesn't bring any information about the variations of Y.
- $R^2 = 1$ means that $e_i = 0$ for $i = 1, \ldots, n$. That is all the points are aligned and the line is the regression line.
- It is straightforward to show that $R^2 = \rho^2(\boldsymbol{y}, \boldsymbol{x}_1) = \rho^2(\boldsymbol{y}, \widehat{\boldsymbol{y}})$. This follows from:

$$\begin{aligned} var(\widehat{\boldsymbol{y}}) &= var(\widehat{\beta}_0\mathbf{1} + \widehat{\beta}_1\boldsymbol{x}_1) \\ &= \widehat{\beta}_1^2\, var(\boldsymbol{x}_1) \\ &= \rho^2(\boldsymbol{y}, \boldsymbol{x}_1)\frac{var(\boldsymbol{y})}{var(\boldsymbol{x}_1)}var(\boldsymbol{x}_1), \end{aligned}$$

and:

$$\begin{aligned} \rho^2(\boldsymbol{y}, \widehat{\boldsymbol{y}}) &= \rho^2(\boldsymbol{y}, \widehat{\beta}_0\mathbf{1} + \widehat{\beta}_1\boldsymbol{x}_1) \\ &= (sign(\widehat{\beta}_1))^2\rho^2(\boldsymbol{y}, \boldsymbol{x}_1). \end{aligned}$$

- We can easily see that

$$F = \frac{MS_{reg}}{MS_{res}} = (n-2)\,\frac{R^2}{1-R^2}$$

which gives a third formulation for the test of $\beta_1 = 0$.

Note that we can write $R^2 = 1 - \dfrac{\sum e_i^2}{\sum (y_i - \bar{\boldsymbol{y}})^2}$. Correcting for the degrees of freedom gives the so-called adjusted R^2.

Definition 5.5. *The adjusted R-squared coefficient is defined by:*

$$\begin{aligned} R_a^2 &= 1 - \frac{\sum e_i^2/(n-2)}{\sum (y_i - \bar{\boldsymbol{y}})^2/(n-1)} \\ &= 1 - \frac{n-1}{n-2}(1-R^2). \end{aligned}$$

The adjusted R-squared coefficient is interesting mainly in multiple regression (replacing in the just above definition $(n-2)$ by $(n-p-1)$ with p the number of variables) for the comparison of models with different numbers of variables.

Using R software, significance tests, estimates of standard deviations, R^2 and R_a^2 values are simply provided by the command `summary`. For the regression of Volume on Height we get:

```
**************************************
summary(reg_trees3)

Call:
lm(formula = Volume ~ Height, data = trees3)

Residuals:
    Min      1Q  Median      3Q     Max
-21.274  -9.894  -2.894  12.068  29.852

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.1236    29.2731  -2.976 0.005835
Height        1.5433     0.3839   4.021 0.000378

Residual standard error: 13.4 on 29 degrees of freedom
Multiple R-squared:  0.3579,Adjusted R-squared:  0.3358
F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784
**************************************
```

Let us notice that, although the link between Height and Volume is statistically significant, R^2=0.3579 is a rather weak value and thus the relationship between the two variables is not very tight.

6 Residuals and diagnostic elements

We discuss in this section several types of residuals. For this we introduce h-values and present tools to investigate the influence of each observation on the final estimation. We are also interested in the extent to which each observation is either well explained or not by the estimated model.

6.1 Residuals and h-values

Let's recall that

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \varepsilon_i \qquad (6.1) \\ \widehat{y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + e_i. \qquad (6.2) \end{aligned}$$

So we can see the residuals e_i are "estimates" of the errors ε_i. Nevertheless, although they are gaussian, the residuals do not exhibit exactly the same behaviour as the errors: in particular they are neither identically distributed
nor independent as stated by the following proposition.

Proposition 6.1. *The vector* $\boldsymbol{e} = (e_1, \ldots, e_n)'$ *is gaussian and*

$$e_i \sim \mathcal{N}(0, \sigma^2(1-h_{ii})) \qquad (6.3)$$

and for $i \neq j$, $cov(e_i, e_j) = -\sigma^2 h_{ij}$.
For any i *and* j, h_{ij} *is defined by:*

$$h_{ij} = \frac{1}{n}\Big[1 + \frac{(x_{i1} - \bar{\boldsymbol{x}}_1)(x_{j1} - \bar{\boldsymbol{x}}_1)}{var(\boldsymbol{x}_1)}\Big]. \qquad (6.4)$$

Note that for $i = j$ we get

$$h_{ii} = \frac{1}{n}\Big[1 + \frac{(x_{i1} - \bar{\boldsymbol{x}}_1)^2}{var(\boldsymbol{x}_1)}\Big] \qquad (6.5)$$

and we use to write h_i instead of h_{ii}. The values h_i are often called *h-values* or *hat-values* and the matrix $H = (h_{ij})$ is the *hat-matrix*. It can be checked that $\widehat{\boldsymbol{y}} = H\boldsymbol{y}$, and that H is symetric and idempotent: H is in fact the projection matrix on $V_1 = \mathcal{V}\{\mathbf{1}, \boldsymbol{x}_1\}$ which is the subspace generated by $\mathbf{1}$ and $\boldsymbol{x}_1$. We will go in more details about the hat-matrix when dealing with the multiple linear regression.
To overcome the defects of the raw residuals, several modified residuals are proposed in the literature and implemented in usual statistical softwares. In particular we have:

- **Standardized** residuals:

$$t_i = \frac{e_i}{\widehat{\sigma}\sqrt{1-h_i}}.$$

- **Studentized** residuals or **cross-validated** residuals or **rstudents** residuals:

$$t_i^* = \frac{e_i}{\widehat{\sigma}_{(i)}\sqrt{1-h_i}}.$$

where $\widehat{\sigma} = \sqrt{\widehat{\sigma^2}}$ and $\widehat{\sigma}_{(i)}$ is the same but calculated in the model fitted without the i-th observation (x_{i1}, y_i).

Given that the model is true, it can be proven that t_i^* is T_{n-3}−distributed. This result would appear surprising: indeed, since e_i is correlated with e_j for $j \neq i$, it is reasonable to suspect that e_i is correlated with $\widehat{\sigma}_{(i)}$. The proof will be outlined in the multiple regression part.

In the standard case where, as n goes to ∞, $\bar{\boldsymbol{x}}_1$ and $var(\boldsymbol{x}_1)$ are bounded, asymptotically the residuals e_i are i.i.d. and $\mathcal{N}(0, \sigma^2)$-distributed. Likewise, the standardised residuals are also asymptotically i.i.d. and $\mathcal{N}(0, 1)$-distributed.

Remark 6.2. Whenever faced with a real dataset, it is advisable to plot data and residuals. Different datasets can result in the same regression line and the same R^2 but dramatically different fitting (see Fig. 8). Plotting residuals may give more information about fitting or reveal some structure and lead to question the model. Residuals also possibly show that some observations play a particular role in the fitting (influent observations), or are badly explained by the regression line (outliers). We detail these points in the following paragraph.

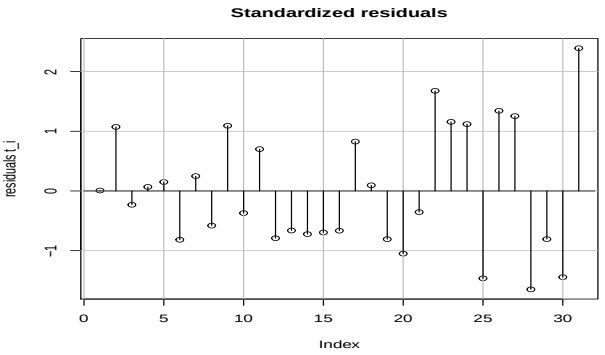

Fig. 5: Standardized residuals for the regression of Volume on Height

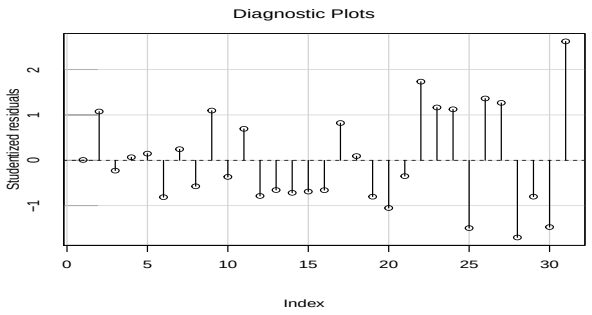

Fig. 6: Studentized residuals for the regression of Volume on Height

Fig. 5 and 6 give plots of standardized residuals and studentized residuals in the regression of Volume on Height for the black cherry trees dataset. When scrutinizing the differences (see Fig. 7), we note that they are significantly marked for observations numbered 22, 25, 28, 30, 31, and the difference is larger for the latter observation. It appears that, except for the fact that they are globally increasing, the residuals don't exhibit any clear structure. Let's observe also that $t^*_{31} = 2.622$ and that $P(|T_{28}| > 2.622) \simeq 0.007$, but we will see that this is not enough to conclude that observation 31 is not consistent with the model.

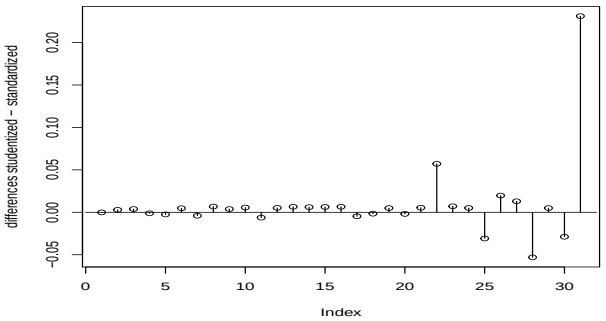

Fig. 7: Differences studentized-standardized residuals

6.2 Four datasets which result in the same regression characteristics

Tomassone and al. [14] build several datasets where the regression lines are the same, as well as the residuals standard errors and the R^2 coefficients: see Fig. 8. Since $\widehat{\sigma}^2$ is the same for the four regressions, it follows also that the estimated covariance matrices of $\widehat{\beta}$ are identical. Using R we plotted the datasets and the regression lines, see fig. 8. The lm procedure of R on the first dataset $(\boldsymbol{x}_1, \boldsymbol{y}_a)$ gives the following results:

```
****************************************
Call:
lm(formula = ya ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5245 -1.6734 -0.2616  2.3119  5.4835

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5248     2.6662   0.197  0.84678
x1            0.8082     0.1701   4.750  0.00031
---

Residual standard error: 3.223 on 14 degrees of freedom
Multiple R-squared:  0.6171, Adjusted R-squared:  0.5898
F-statistic: 22.57 on 1 and 14 DF   p-value: 0.0003102
****************************************
```

Except for the quantiles of residuals, we get precisely the same results for the estimated regressions calculated on $(\boldsymbol{x}_1, \boldsymbol{y}_b)$, $(\boldsymbol{x}_1, \boldsymbol{y}_c)$ and $(\boldsymbol{x}_1, \boldsymbol{y}_d)$.
It would appear surprising that these four rather different datasets lead to the same R^2 value. It must be noted that R^2 is of interest as long as the usual assumptions of a linear regression are satisfied. Looking at the plots gives evidence that dataset (a) is consistent with these assumptions and that it is not the case for (b), (c) and (d). The mean function seems nonlinear in (b) while in (c) there is an observation which plays a particular role and which is unconsistent with the other ones, and in (d) the variance of the error

seems to be increasing as x_1 is growing. All this emphasizes the importance of a graphical representation of data.

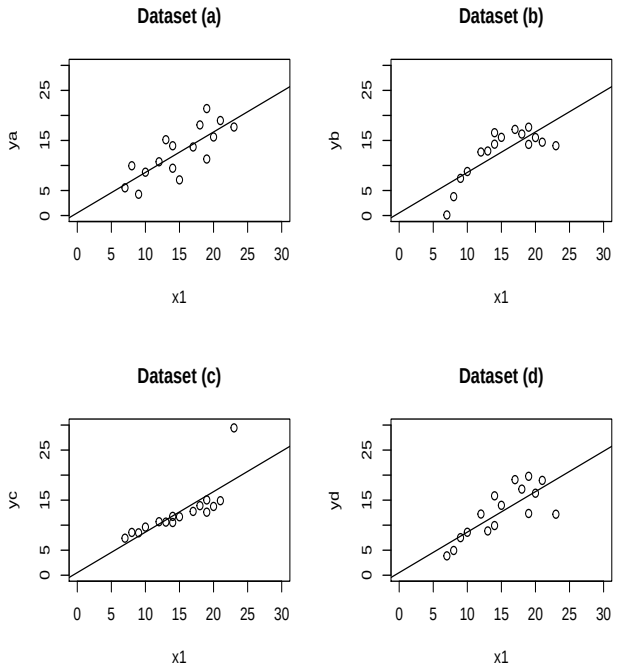

Fig. 8: Four datasets resulting in identical regression lines, residuals standard errors and R-squared (from Tomassone et al. [14]).

6.3 Diagnostic elements, outliers, observations with high leverage effect, influent observations

Once a linear regression has been fitted to data, an important task to be carried out is what is called diagnostic's analysis. It consists in the analysis of residuals to detect particular observations:

- Observations which are badly explained by the model (outliers): their residuals are too high and inconsistent with the model.
- Observations with high leverage effect. This means that moving such an observation results in a markedly different model.
- Influent observations. An observation is said influent if dropping out this observation leads to a fitted model significantly different from the model fitted with the whole dataset.

For each of the three types of observations, indicators are defined in the literature to help detection.

- **Outliers**. The rstudent residuals are usually the main tool to detect outliers. Each rstudent t_i^* is $T_{n-3}-$distributed when the model (1.11) is correct. This will be made clearer when studying multiple regression where it is shown that in the model $y = \beta_0 + \beta_1 x_1 + \delta\, I\{x_1 = x_{i1}\}$ the test statistic of $H_0 : \{\delta = 0\}$ is $T_{n-3}-$distributed under H_0. In practice we consider the highest t_i^*, say t_{max}^*, and the corresponding observation is declared outlier when $|t_{max}^*|$ is higher than the $100 \times (1 - \alpha/2n)$-quantile of the distribution of $max(|t_1|, |t_2|, \ldots, |t_n|,)$ under H_0 :{ the model (1.11) is correct}. In fact we can't compute the quantile, but using the Bonferroni inequality we get an approximation for this quantile given by the $100 \times (1 - \alpha/2n)$-quantile of the T_{n-3} distribution.

- **Observations with high leverage effect**. The way to detect this type of observations is to use hat-values h_i. By the idempotence of H we get

$$\begin{aligned} h_i &= \sum_j h_{ij}^2 \\ &= h_i^2 + \sum_{j \neq i} h_{ij}^2, \end{aligned}$$

from which, using also (6.5), it comes that $1/n \leq h_i \leq 1$.
Now from $\widehat{\boldsymbol{y}} = H\boldsymbol{y}$ it follows

$$\begin{aligned} \widehat{y}_i &= \sum_j h_{ij} y_j \\ &= h_i y_i + \sum_{j \neq i} h_{ij} y_j. \end{aligned}$$

This latter equality shows that when h_i is close to 1, which happens when x_{i1} is far from $\bar{\boldsymbol{x}}_1$, $\widehat{y}_i$ is close to y_i. That is observations remote from the mean $\bar{\boldsymbol{x}}_1$ show high h_i and attract the regression line. A consequence is that when such an observation is moved around its location, the regression line moves in a similar way.
As $\sum h_i = Tr(H) = 2$, the mean value of the h_i values is $2/n$, and an observation is said to have a high leverage effect when $h_i \geq 2\,\dfrac{2}{n}$.

- **Influent observations**. As said in the beginning of this subsection, an observation is said to be influent if the fitted model without this observation is significantly different from the model fitted with the whole set of observations. A number of indicators of this difference were defined: among these are the Cook's distance, the dfitts, dfbeta... These indicators focus on the variation of different characteristics of the fitting. Cook's distance and dfitts focus on the gap between the values of $(\widehat{y}_i, i = 1, 2, \ldots, n)$ in the two models, while dfbeta compares $\widehat{\beta}$, the estimated and $\widehat{\beta}_{(i)}$, where the subscript (i) stands for an estimate obtained without the $i-$th observation.
Below we use only the Cook's distance D_i. A computation shows that D_i can be written in a very simple way:

$$D_i = \frac{t_i^2}{2} \frac{h_i}{1 - h_i}.$$

Then when $D_i > 4/(n-2)$ the i-th observation is declared influent.

6.4 Diagnostic analysis on a toy dataset

The Fig. 9 illustrates the ideas of diagnostic analysis on an artificial toy dataset. It can be shown that:

- The highest rstudent is $t^*_{13} = 4.349$. The Bonferroni approximation to the 5% critical value is given by $c^* = 3.556$. Hence observation 13 is considered as an outlier. The following rstudent is $t^*_{18} = 2.188$ and can't be declared as outlier.
- The highest h-value is $h_{19} = 0.519$, the following one is $h_{18} = 0.176$. The threshold is 0.211. Therefore we see that there is only one observation, observation 19, with high leverage effect.
- The threshold for the Cook's distance is 0.235, and only observations 13 and 18 exhibit distance over the threshold: $d_{18} = 0.433$ and $d_{13} = 0.317$. So in view of this criterium they are the only ones which are considered as influent. Using dffits coefficients leads to the same conclusion. It is worthwhile to note that 19 is not an influent observation.

In summary:
- 13 is both an outlier and influent observation, but has not a high leverage effect.
- 18 is an influent observation without high leverage effect and is not an outlier.
- 19 is an observation with high leverage but is neither influent nor an outlier.

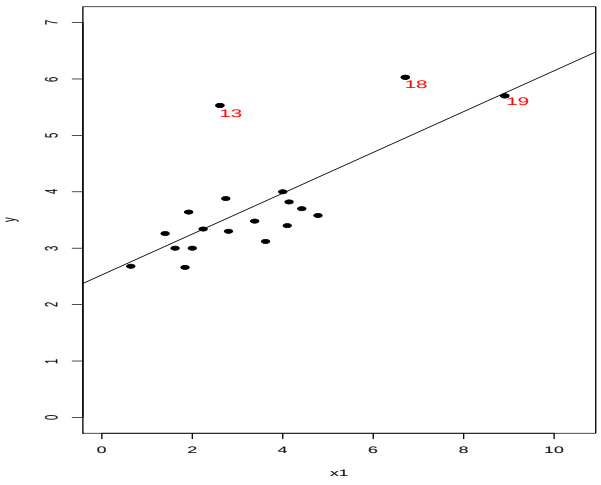

Fig. 9: Diagnostics: an artificial dataset with outliers, influent data and data with high leverage effects

Valuable references about diagnostics analysis include, among others, R. Cook and S. Weisberg [4] and D. A. Besley, E. Kuh and R. E. Welsch [3].

6.5 Diagnostic analysis for black cherries trees data with R software

R provides all the functions needed for diagnostics analysis and associated graphical procedures. For instance for the regression of Volume on Heigth in the black cherries trees data, we get h-values, fitted values, standardized residuals and studentized residuals by:

```
**************************************
hatvalues(reg_trees3)
fitted(reg_trees3)
rstandard(reg_trees)
rstudent(reg_trees3)
**************************************
```

and indicators needed to detect influent observations are provided by:

```
**************************************
dffits(reg_trees)
dfbeta(reg_trees3)
cooks.distance(reg_trees3)
**************************************
```

or simply:

```
**************************************
influence.measures(reg_trees3)
**************************************
```

The library "car" allows synthesized plots of the relevant information:

```
**************************************
influenceIndexPlot(reg_trees3,vars=
    c("Studentized","Cook","hat"),id.n=4)
**************************************
```

see Fig. 10.
The reader is invited to check that, for the regression of Volume on Heigth, observations 1,2,3 and 31 have high leverage effect (threshold 0.129). There is no outlier: Bonferroni threshold 3.491, with $\alpha = 5\%$, for the highest absolute value of rstudents is 3.491. The Cook's distance considers observations 30 and 31 as influent observations (threshold 0.138).

7 Departures from the basic model assumptions

Statisticians know that in "real life" the model assumptions are rarely totally satisfied. Very often they are faced with departures from model assumptions. This point is the subject of lengthy developments: how to deal with the non-linearity of the mean function $\mu(x_1)$, non-normality of the errors ε_i, heteroscedasticity (non-constant variance), random regressor? These issues are fundamental when faced with real data, and are sometimes neglected. Since this paper is intended above all to be a motivating introduction to a particular field of statistics, we seize the opportunity to briefly indicate and outline directions that deal with departures from the model assumptions.

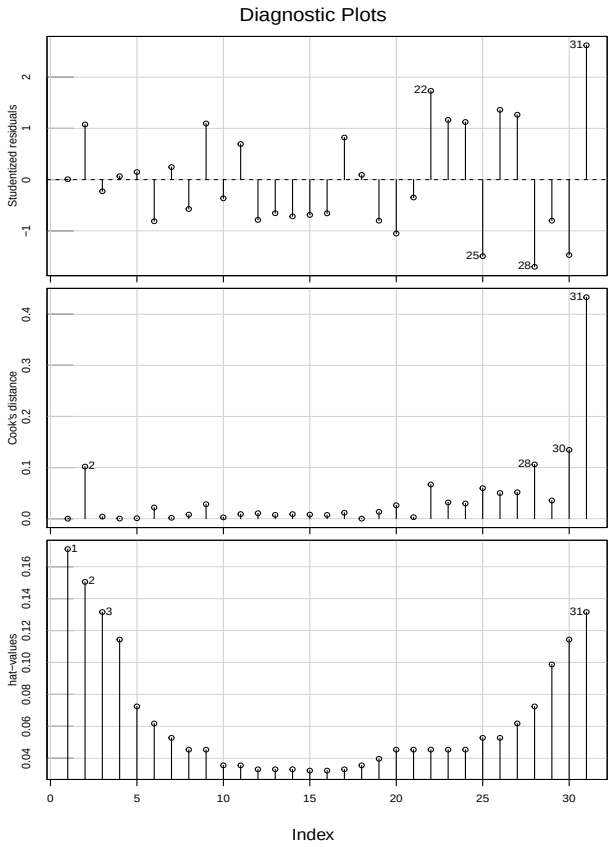

Fig. 10: Studentized residuals, Cook's distance and hat-values for Volume regression on Height

7.1 Non-linearity

When the relationship between the mean of y and x_1 in non-linear a possible strategy is to apply a transformation to the response variable y, or to the regressor x_1, or to both y and x_1. Usual transformations are $Log_e(x)$, $1/x$, $\sqrt{x}$. It is also possible to choose a transformation in the family of powers of y. It is what does the Box-Cox procedure. The family is defined by:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y) & \text{if } \lambda = 0. \end{cases}$$

The parameter λ can be estimated together with β using the method of maximum likelihood. This amounts to choose the λ value which minimizes the sum of squared residuals $SS_{res}(\lambda)=\sum e_i^2(\lambda)$ where the $e_i(\lambda)$ are the residuals coming from the fitting of the λ-model. That is we select the value of λ for which the fit is the best. Note that the transformations are not always defined when $y < 0$, but this difficulty can be circumvented by adding a positive constant to each of the y values.
It can happen that the usual transformations are unable to linearize the mean function. In this case a sensitive strategy is to rely on multiple linear regression using an approximation of the mean function: this can be for instance an approximation by a constant piecewise function, or a polynomial function or a spline function.

7.2 Non-normality

When the errors are non-gaussian, we use nevertheless the methodology developed above, and the main consequence is that $\widehat{\beta}$ is non-gaussian and is not asymptotically efficient. But $\widehat{\beta}$ is still unbiased, and is BLUE; that is exhibits the best variance in the class of linear unbiased estimates.
Thus when the sample size n is small, the test procedures, confidence intervals and prediction intervals discussed above can't be used since the T distributions are no longer relevant.
When on the contrary n is large, $\widehat{\beta}$ is approximately gaussian, precisely $(\widehat{\beta}_i - \beta_i)/s_i \approx \mathcal{N}(0,1)$, $i = 0, 1$. As the T_n distribution and $\mathcal{N}(0,1)$ become close, when n is large, the tests and CI procedures used in the gaussian case remain valid. Note that this is not the case for prediction intervals.
Let's remark also that we can at first apply transformations such as those presented in the previous paragraph. Such transformations can reduce substantially the non-normality.

7.3 Heteroscedasticity

The basic model (1.11) assumes homoscedasticity; that is $\mathbb{V}\text{ar}(\epsilon_i)$ doesn't depend on x_{i1}. When it is not the case, that is $\mathbb{V}\text{ar}(\varepsilon_i) = \sigma^2(x_{i1})$, the developments above based on a unique σ^2 don't make sense.
Once again it is possible that a transformation of y makes the variance more homogeneous and overcome the difficulty. In some special situations we can find out an adhoc way to return to the basic homoscedastic model. Another method consists in using weighted linear regression, that is minimizing:

$$S_W = \sum w_i (y_i - \beta_0 - \beta_1 x_{i1})^2$$

where $w_i = 1/\sigma_i^2$ and $\sigma_i = \sqrt{\mathbb{V}\text{ar}(\varepsilon_i)}$. Of course if the variances are unknown we must substitute estimates $\widehat{\sigma}_i^2$ (to be defined) to σ_i^2.
The weighted linear regression is in fact a particular case of generalized linear regression, i.e. GLS, which deals with the more general situation where the variance-covariance matrix of $\boldsymbol{\varepsilon}$ is different from $\sigma^2 \boldsymbol{I}_n$, that is we allow non-equal variances as well as autocorrelation of errors.

7.4 Random regressor

We can be faced with situations where the x_1- values are not fixed deterministic values: the set $\{x_{11}, \ldots, x_{n1}\}$ can be a $n-$sample for the random regressor which we now denote by X_1. In this context the standard regression linear model is defined in the following way. We have:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon,$$

and we assume that:

$$\mathbb{E}(\varepsilon \,|\, X_1) = 0 \quad \text{a.s.}, \tag{A.1}$$

$$\text{and} \quad \mathbb{V}\text{ar}(\varepsilon \,|\, X_1) = \sigma^2 \quad \text{a.s.} \tag{A.2}$$

It can be easily checked that (A.1) implies $\mathbb{E}(\varepsilon) = 0$ and that under (A.1) and (A.2) we have $\mathbb{C}\text{ov}(\varepsilon, X_1) = 0$.
Furthermore we assume that the conditional distribution of ε given X_1 is gaussian, which in turn can be precisely written $\mathcal{L}(\varepsilon \,|\, X_1) = \mathcal{N}(0, \sigma^2)$. From this it follows that the unconditional distribution of ε is also $\mathcal{N}(0, \sigma^2)$.
The reader can note that the model with random regressor is closely mimicking the model with fixed design.

The observations are given by (x_{i1}, y_i), $i = 1, \ldots, n$, and are values of n i.i.d. couples of variables (X_{i1}, Y_i), distributed as (X_1, Y). In particular, this assumption excludes any autocorrelation in $(\varepsilon_i)_{i=1,\ldots,n}$ as well as in $(X_{i1})_{i=1,\ldots,n}$ and ensures that, for $i \neq j$, X_{i1} and ε_j are independent.
We can't go into too many details on the random regressor issue. Thus we merely outline below some important points.

- **Conditional statistical inference.**
  When we perform statistical inference conditionally upon the observed values $\{x_{11}, \ldots, x_{n1}\}$ of the random regressor X_1, it is clear that all the results presented above for the least squares method hold true. The estimate $\widehat{\boldsymbol{\beta}}$ is unbiased, gaussian and optimal in the class of unbiased estimates. Moreover $\widehat{\sigma^2}$ is unbiased, $(n-2)\widehat{\sigma^2}/\sigma^2$ is χ^2_{n-2}-distributed. $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma^2}$ are independent and consequently the studentized $\widehat{\beta}_i$, $i = 0, 1$, are $T_{n-2}-$distributed. The CI and PI intervals designed above thus remain valid.

- **Unconditional statistical inference.**
  When we are concerned with unconditional inference, we have to be careful, even if the most part of the least squared methodology remains valid.
  It must be noted that the unconditional distribution of $\widehat{\boldsymbol{\beta}}$ is not necessarily gaussian, $(n-2)\widehat{\sigma^2}/\sigma^2$ is χ^2_{n-2}-distributed but the independence between $\widehat{\sigma^2}$ and $\widehat{\boldsymbol{\beta}}$ is not ensured. Nevertheless, in spite of these points, the conditional properties of the least squared estimates imply that the studentized parameter estimates $(\widehat{\beta}_i - \beta_i)/s_i$, $i = 0, 1$, are still $T_{n-2}-$distributed. Hence the tests and CI for β_i, $i = 0, 1$, are still valid. A similar argument leads to the same conclusion when we are concerned with the CI of $\mu(x_1^0)$, where x_1^0 is a new observation of X_1, or the PI of $Y(x_1^0)$, or even the one of $Y^0 = \beta_0 + \beta_1 X_1^0 + \varepsilon$.

- **When ε and X_1 are correlated.**
  We saw above that (A.1) and (A.2) entail $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{C}\text{ov}(\varepsilon, X_1) = 0$. Thus we could be interested in substituting to (A.1) the assumption:

$$\mathbb{C}\text{ov}(\varepsilon, X_1) = 0. \tag{A.1b}$$

  Under (A.1b) it is clear that $\mathbb{E}(\varepsilon|X_1) \neq 0$ but $\mathbb{E}(\varepsilon) = 0$ can possibly be preserved.
  It is worthwhile to note that, while (A.1b) is not much weaker than (A.1), we are not ensured under (A.1b) that $\widehat{\beta}$ is unbiased. Nor are we ensured that the usual finite sample properties hold true.
  It is essential also to note that when ε and X_1 are correlated, then the least squared estimates are biased and unconsistent. As a consequence the usual test procedures and confidence intervals are not valid.
  It is not that easy to detect a correlation between ε and X_1. And when such a correlation is suspected, to overcome this shortcoming is not always possible. A method (instrumental variables method) to do that has been developped by econometricians: it is based on the existence of other variables $Z_1, \ldots, Z_q$, $q \geq 1$, which are strongly correlated with X_1 and uncorrelated with ε. As it would be out of the scope of this paper to go into details about this method, we refer readers interested in this topic, for instance, to [8] or [12].

References

[1] A. Antoniadis, J. Berruyer and R. Carmona. Regression non linéaire et applications. Economica, 1992.

[2] J.-M. Azaïs and J.-M. Bardet. Le modèle linéaire par l'exemple. Dunod, coll. Sciences Sup, 2005.

[3] D. A. Besley, E. Kuh and R. E. Welsch. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, New York,1980.

[4] R. Cook and S. Weisberg. Residuals and Influence in Regression. Chapman and Hall, 1982.

[5] G. Grégoire. Regression Models: a brief introduction. Statistics for Astrophysics. Methods and Applications of the Regression. Eds D. Fraix-Burnet and Valls-Gabaud. EAS Publications Series Vol. 66. EDP Sciences 2014. p. 3-9.

[6] G. Grégoire. Simple Linear Regression. Ibid. p. 19-40.

[7] G. Grégoire. Multiple Linear Regression. Ibid. p. 45-72.

[8] W. Greene. Econometric Analysis. Macmillan Publishing Company, 6th edition, 2007.

[9] G. James, D. Witten, T. Hastie. & R. Tibshirani. An Introduction to Statistical Learning. Springer Texts in Statistics, 2013.

[10] T. Hastie, R. Tibshirani & J. Friedman. The Elements of Statistical Learning. 2nd edition. Springer Series in Statistics. Springer, 2008.

[11] C. Bishop. Pattern recognition and machine learning. Springer, 2006.

[12] R. C. Hill, W. E. Griffiths, G. C. Lim. Principes of Econometrics. Wiley, Fourth edition, 2011.

[13] X. Guyon. Statistique et économétrie. Du modèle linéaire...au modèle non-linéaire. Ellipses universités. Paris, 2001.

[14] R. Tomassone, E. Lesquoy de Turckheim and C. Millier. La régression. Nouveaux regards sur une ancienne méthode statistique. Masson 1983.

[15] J. Fox and S. Weisberg. An R Companion to Applied Regression, Second Edition, Sage, 2011.

[16] D. Montgomery and E. Peck. Introduction to linear regression analysis, Second Edition, Wiley, 2008.

[17] C. R. Rao. Linear Statistical Inference and Its applications. Wiley, 1965.

Gérard GRÉGOIRE
E-mail address: gerard.gregoire@imag.fr

Laboratoire Jean Kuntzmann
Bâtiment IMAG
Université Grenoble Alpes
700 Avenue Centrale
38401 Domaine Universitaire de Saint-Martin-d'Hères, France