Linear regression

Elise Arnaud elise.arnaud@univ-grenoble-alpes.fr

UGA Mosig

Table of contents

Understand the different types of statistical modeling

A statistical modeling consists in establishing a relation between variables, in the form of an equation, which is estimated on a set of observed data.

The challenge is to use this relationship, established and verified, on observations, for prediction purposes: we are here in the context of inference.

Different types of modeling

- Linear regression, which allows to explain a quantitative variable from quantitative explanatory variables (possibly also qualitative)
- The supervised classification, which allows to explain a qualitative variable from quantitative explanatory variables (possibly qualitative in addition). Attention, it must be distinguished from the unsupervised classification which is the clustering.
- Analysis of variance, to analyze the influence of one or two qualitative explanatory variables on a quantitative variable.

What is regression ?

Regression analysis is the most widely used statistical tool for understanding relationships among variables. Several possible objectives including:

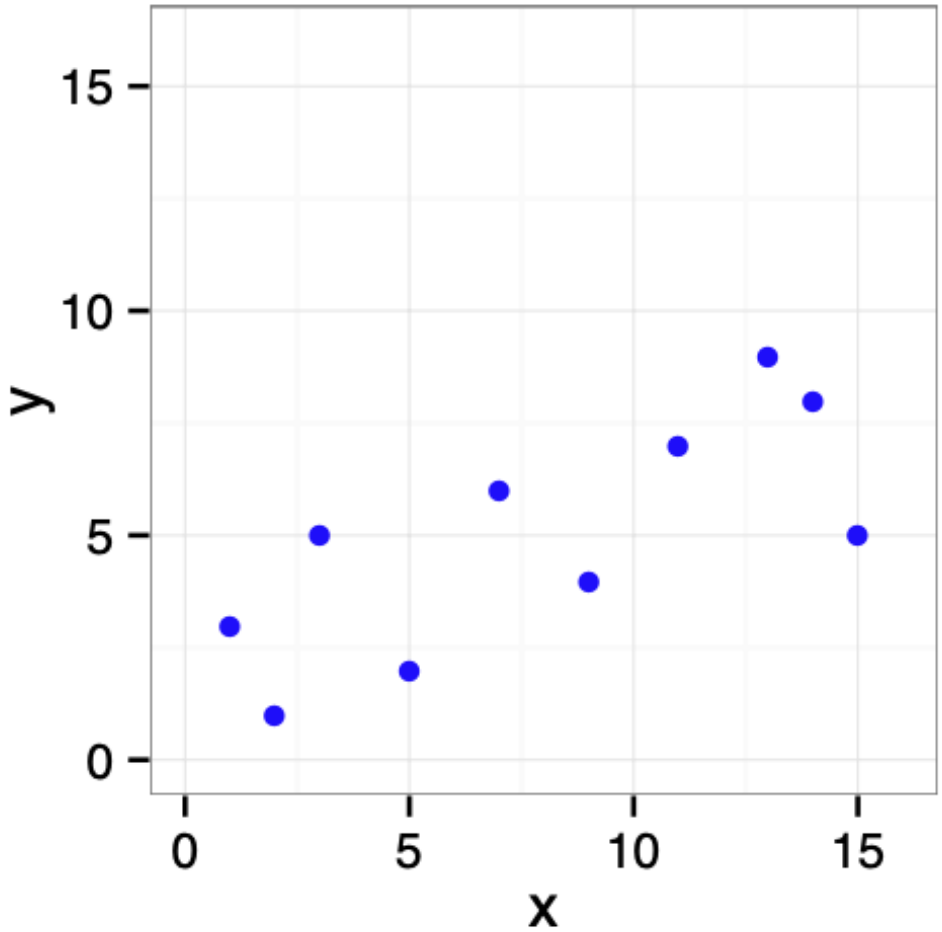
- Prediction of future observations. This includes extrapolation since we all like connecting points by lines when we expect things to be continuous
- Assessment of the effect of, or relationship between, explanatory variables on the response
- A general description of data structure (generally expressed in the form of an equation or a model connecting the response or dependent variable and one or more explanatory or predictor variable)
- Defining what you should "expect" as it allows you to define and detect what does not behave as expected

Table of contents

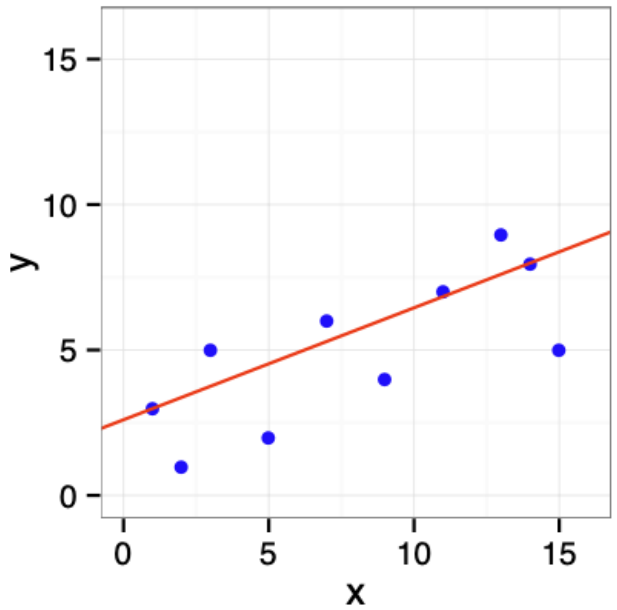
Starting With a Simple Data Set

How could we summarize the following data set ?

	x	y
1	1.00	3.00
2	2.00	1.00
3	3.00	5.00
4	5.00	2.00
5	7.00	6.00
6	9.00	4.00
7	11.00	7.00
8	13.00	9.00
9	14.00	8.00
10	15.00	5.00

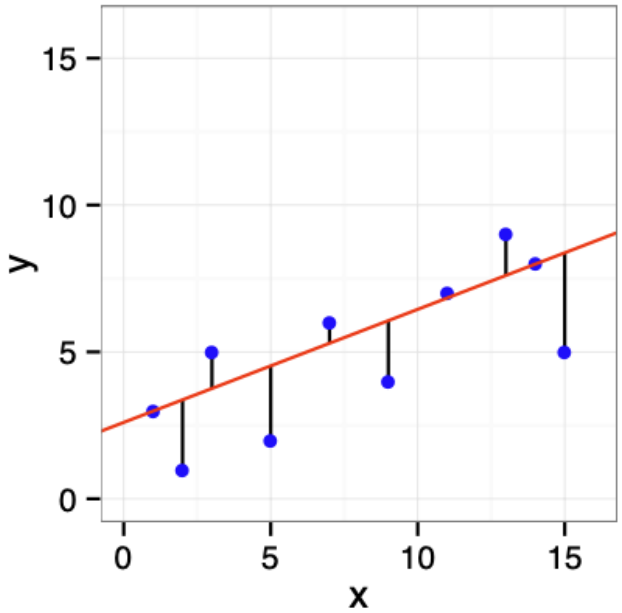


The "Eyeball" Method



- A straight line drawn through the maximum number of points on a scatter plot balancing about an equal number of points above and below the line
- Some points are rather far from the line. Maybe we should instead try to minimize some kind of distance to the line

Least Squares Line : What to minimize?

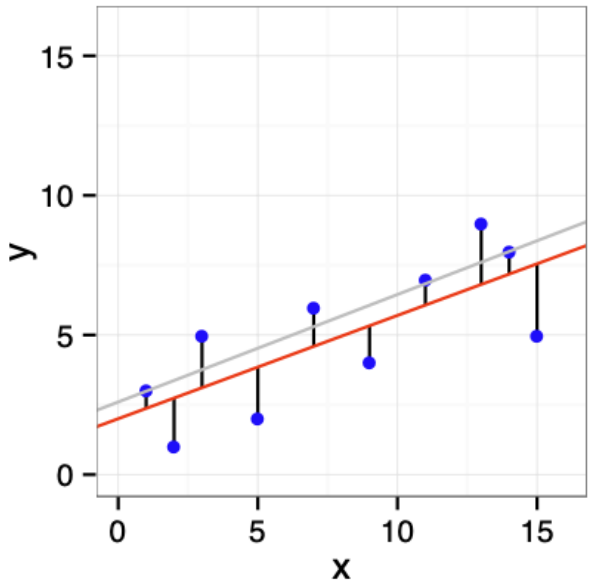


Intuitively, a large error is much more important than a small one. We could try to minimize the size of all residuals:

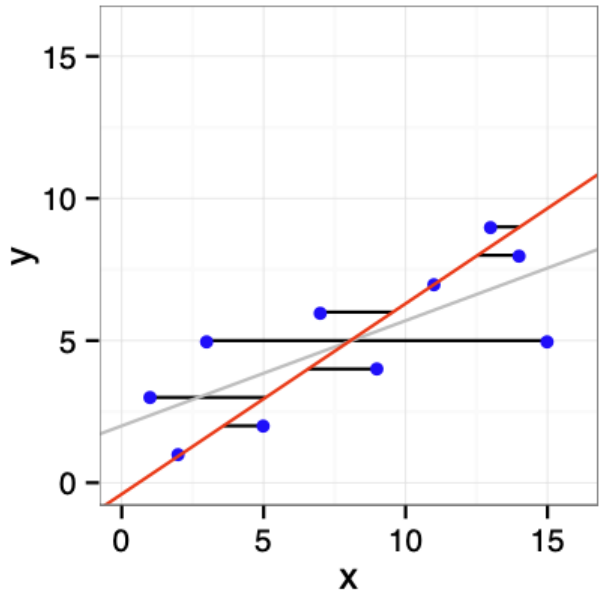
$$F(\alpha, \beta) = \sum_i e_i = \sum_i (y_i - \alpha - \beta x_i)^2$$

- If they were all zero we would have a perfect line
- Trade-off between moving closer to some points and at the same time moving away from other points

Least Squares Line : y as a function of x or the opposite?

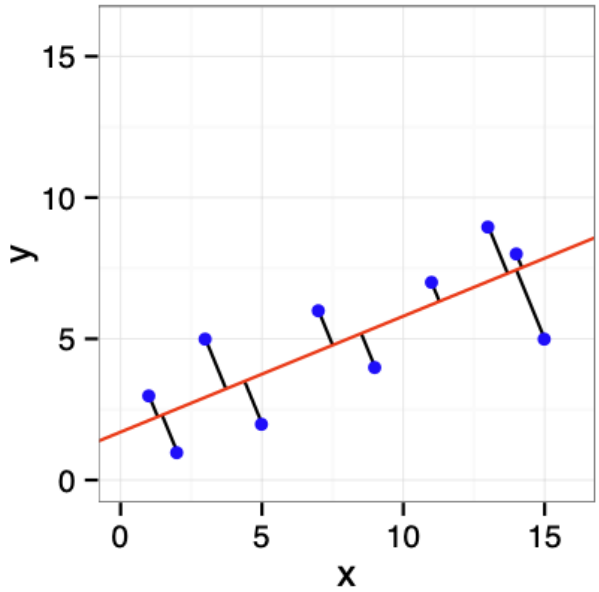


Least Squares Line : y as a function of x or the opposite?



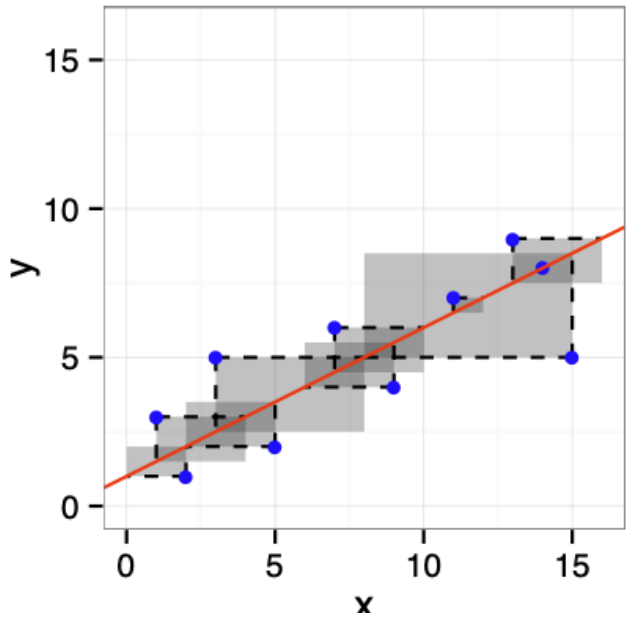
ok, do we have less asymetrical options ?

Least Distances Line (a.k.a. Deming Regression)



- Note that somehow, this makes sense only if we have a square plot, i.e., if x and y have the same units

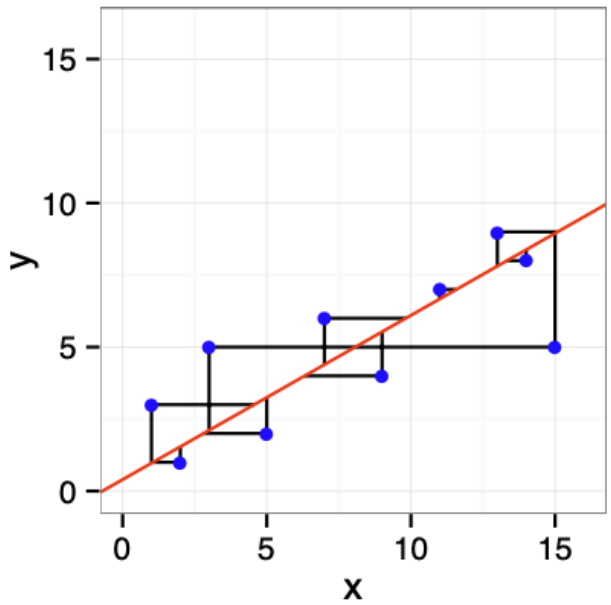
Least Rectangles Line



- Minimize

$$E(\alpha, \beta) = \sum_i |x_i - \frac{y_i - \alpha}{\beta}|.|y_i - \alpha - \beta x_i|$$

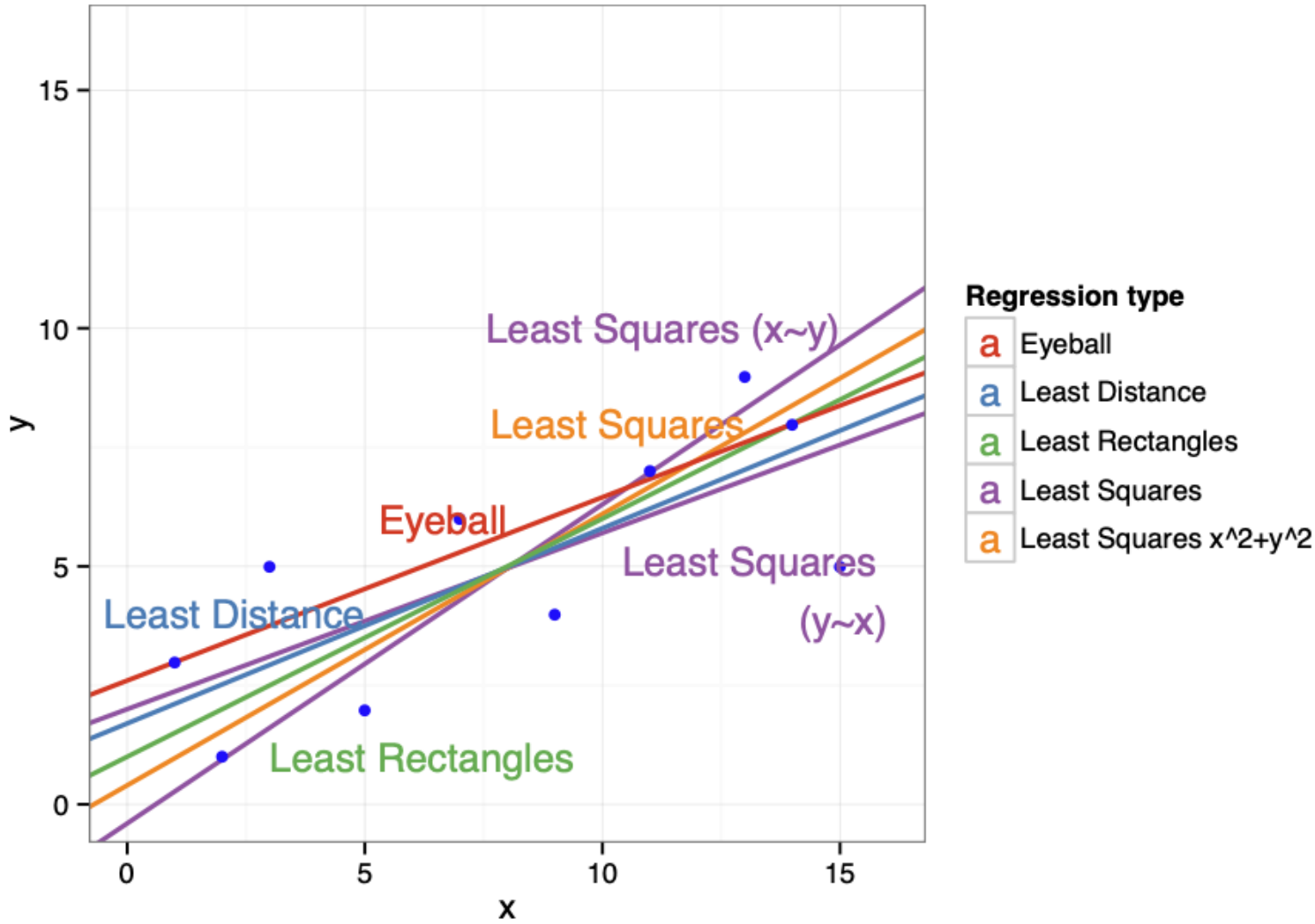
Least Rectangles Line



- Minimize

$$D(\alpha, \beta) = \sum_i [(x_i - \frac{y_i - \alpha}{\beta})^2 + (y_i - \alpha - \beta x_i)^2]$$

Which line to choose?



Which line to choose?

- Eyeball: AFAIK nothing
- Least Squares: classical linear regression $y\tilde{x}$
- Least Squares in both directions: I don't know
- Deming: equivalent to Principal Component Analysis
- Rectangles: may be used when one variable is not "explained" by the other, but are inter-dependent

This is not just a geometric problem. You need a model of to decide which one to use

Table of contents

The model

The simple linear regression model assumes, as its name suggests, that there is a linear relationship between the variable to be explained Y and the explanatory variable X:

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

- Y is a random variable, observable:
- X is a deterministic variable (not random), observable
- β_1 and β_2 are unknown parameters (not observable)
- ε is a centred random variable (around O) of unknown variance σ^2 (it is also a parameter of the model). It refers to the modelling error made.

The objective is to determine and estimate the parameters of the regression line: the intercept β_1 and the directing coefficient β_2.

The data

we have n observations $(x_i, y_i)_{i\in\{1,\dots,n\}}$ of an i.i.d. sample of (X, Y)
Thus, according to the regression model posed above, we have

$$\forall i \in \{1, \dots, n\} : y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

The errors $(\varepsilon_i)_{i\in\{1,\dots,n\}}$ verify for $(i,j)i \in \{1, \dots, n\}^2$:

- $\mathbb{E}(\varepsilon_i) = 0$ (they are centred around 0)
- $\mathrm{Var}(\varepsilon_i) = \sigma^2$ (their variance, unknown, is constant and equal to σ^2)
- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ if $i \neq j$ (they have no linear dependence).

To go further: matrix writing

We can write:

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where:

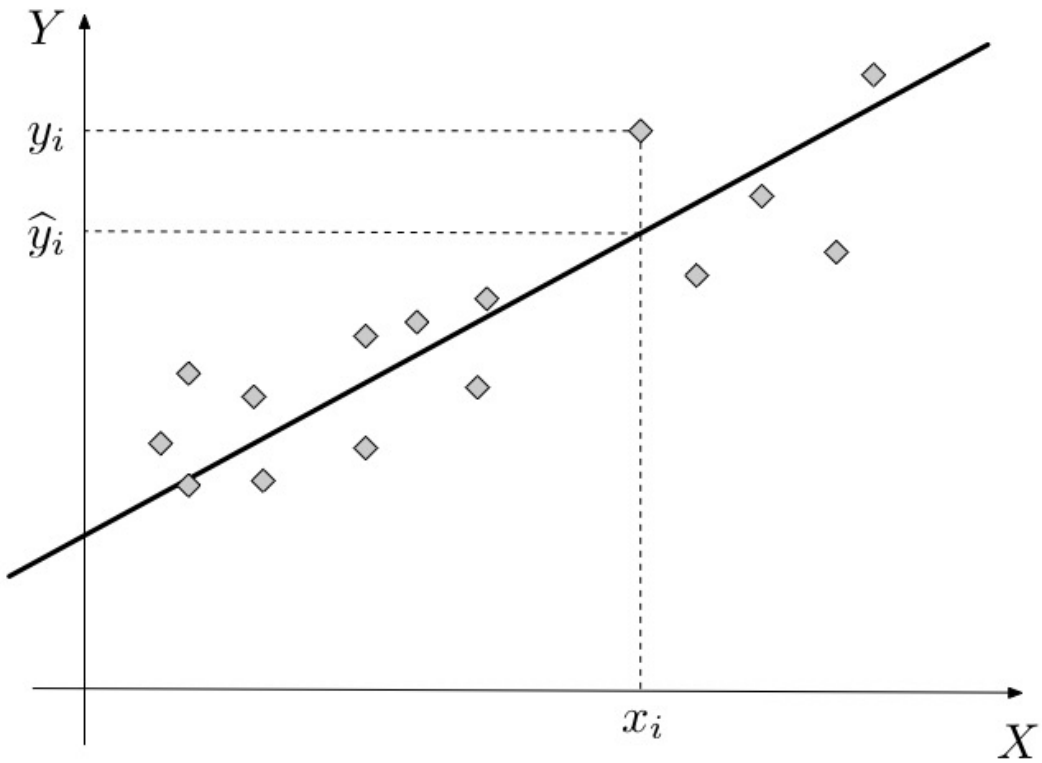
$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \mathbb{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

This notation will be used a lot in the case where we have not one, but several explanatory variables.

The Ordinary Least Squares estimator

We call the OLS estimator of β_1 and β_2 the $\widehat{\beta}_1$ and $\widehat{\beta}_2$ values minimising the sum of the squares of the residuals:

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$



The Ordinary Least Squares estimator

In the case where at least one of the x_i differs from the others (which is always the case in practice), the OLS estimators of (β_1, β_2) are:

$$\widehat{\beta}_2 = \frac{s_{XY}}{s_X^2}$$

$$\widehat{\beta}_1 = \bar{y} - \widehat{\beta}_2 \bar{x}$$

where

$$\bar{x} = \frac{1}{n}\sum_i x_i \qquad \bar{y} = \frac{1}{n}\sum_i y_i$$

$$s_X^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 \qquad s_{XY} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Let us note that the directing coefficient of the line $\widehat{\beta}_2$ is proportional to the empirical covariance between X and Y, which is a measure of the linear dependence between the variables.

The regression line

The equation of the regression line is:

$$y = \widehat{\beta}_1 + \widehat{\beta}_2 x$$

It can be shown that this line passes through the barycentre of the scatter plot $(\overline{x}, \overline{y})$.

The Ordinary Least Squares estimator

Note that the minimized distance with OLS is $e_i = y_i - \hat{y}_i$ (green), not the distance of the point to the regression line (red):

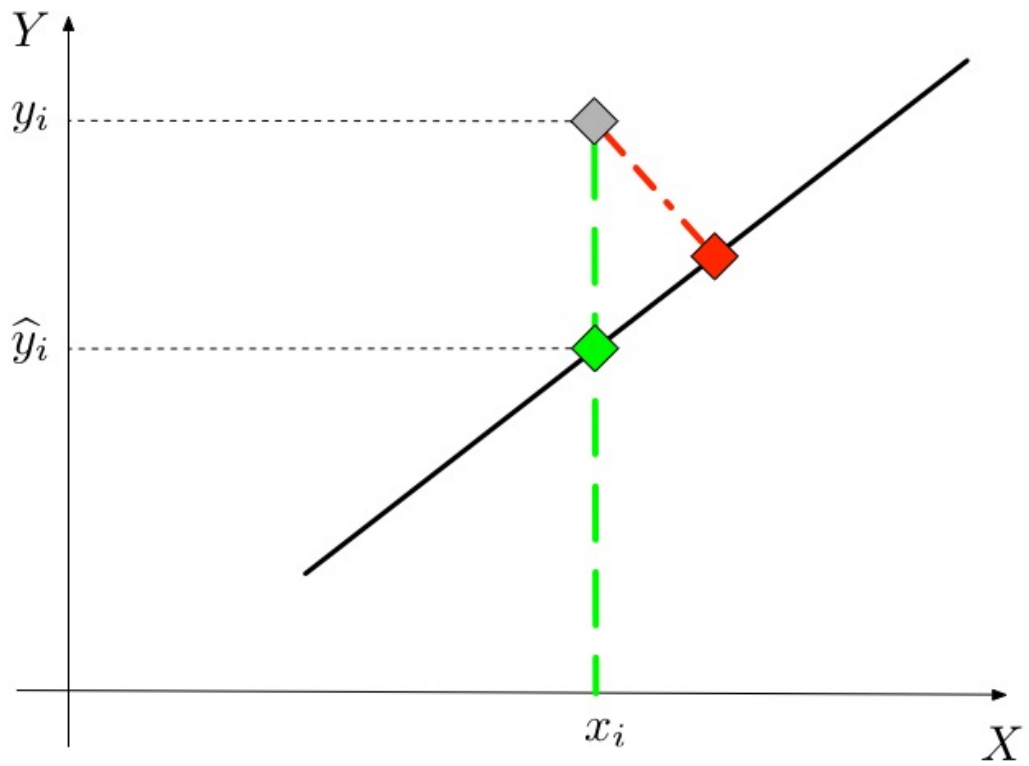


Figure: Least squares distance

Adjusted values and residuals

For observation i, the quantity is called the fitted value (or estimated value):

$$\hat{y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i$$

The residual is the difference between the observed value of the variable to be explained and its estimate. It represents the part unexplained by the model. The residual, for individual i , is thus:

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

The residuals, depending on the estimated parameters, are calculable, unlike the noise which depends on the unknown parameters:

$$\varepsilon_i = y_i - \beta_1 - \beta_2 x_i$$

The residual e_i is an estimate of the noise $\varepsilon_\mathbf{i}$. It represents the part not explained by the model for individual i.

To go further: the statistical properties of the parameters

It can be shown that $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are unbiased estimators of β_1 and β_2:

$$\forall j \in \{1, 2\} : \mathbb{E}\left(\widehat{\beta}_j\right) = \beta_j$$

We can be all the more confident in the quality of these estimators as they are said to be BLUE Best Linear Unbiased Estimators: among all the linear and unbiased estimators of β_1, and β_2, the estimators of the OLS of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are of minimal variance

The residual variance

The residual variance is :

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$$

This is an unbiased estimator of σ^2.

ANOVA

The interest of a linear regression model lies in its capacity to explain part of the variations of the variable Y by the variations of the variable X.

The variation of a variable y_i is obtained by considering the differences between the observed values y_i and their mean y_i. Now we have :

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + y_i - \hat{y}_i$$

where

- $\hat{y}_i - \bar{y}$ is the variation explained (or restored) by the model
- while $y_i - \hat{y}_i$ is the variation not explained by the model.

ANOVA

We can establish the formula for the decomposition of the variance (ANOVA: ANalysis Of VAriance):

SST	= SSM	+SSE
$\sum_{i=1}^{n} (y_i - \bar{y})^2$	$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

- SST (Sum of Squares Total) translates the total variation of Y.
- SSM (Sum of Squares of the Model) expresses the variation explained by the model.
- SSE (Sum of Squares of the Error) expresses the variation unexplained by the model.

Coefficient of determination

The coefficient of determination is the following quantity:

$$\mathrm{R}^2 = \frac{\mathrm{SSM}}{\mathrm{SST}}$$

This coefficient $\mathbf{R}^2$ is in $[0, 1]$, since :

$$0 \leq \mathrm{SSM} \leq \mathrm{SST}$$

- If $\mathbf{R}^2 = 1$, then we have $\mathbf{SSM} = \mathbf{SST}$: all the variation is explained by the model.
- If $\mathbf{R}^2 = \mathbf{0}$, then we have $\mathbf{SSE} = \mathbf{SST}$: no variation is explained by the model.

The risk of over-interpreting

Care must be taken not to over-interpret the coefficient of determination:

- A good linear adjustment is reflected by a $\mathbf{R}^2$ close to 1.
- On the other hand, a $\mathbf{R}^2$ close to 1 does not necessarily translate into a linear link.
- A $\mathbf{R}^2$ close to 0 indicates a poor linear fit, but does not imply that no relationship can be established between the variables.

The risk of over-interpreting

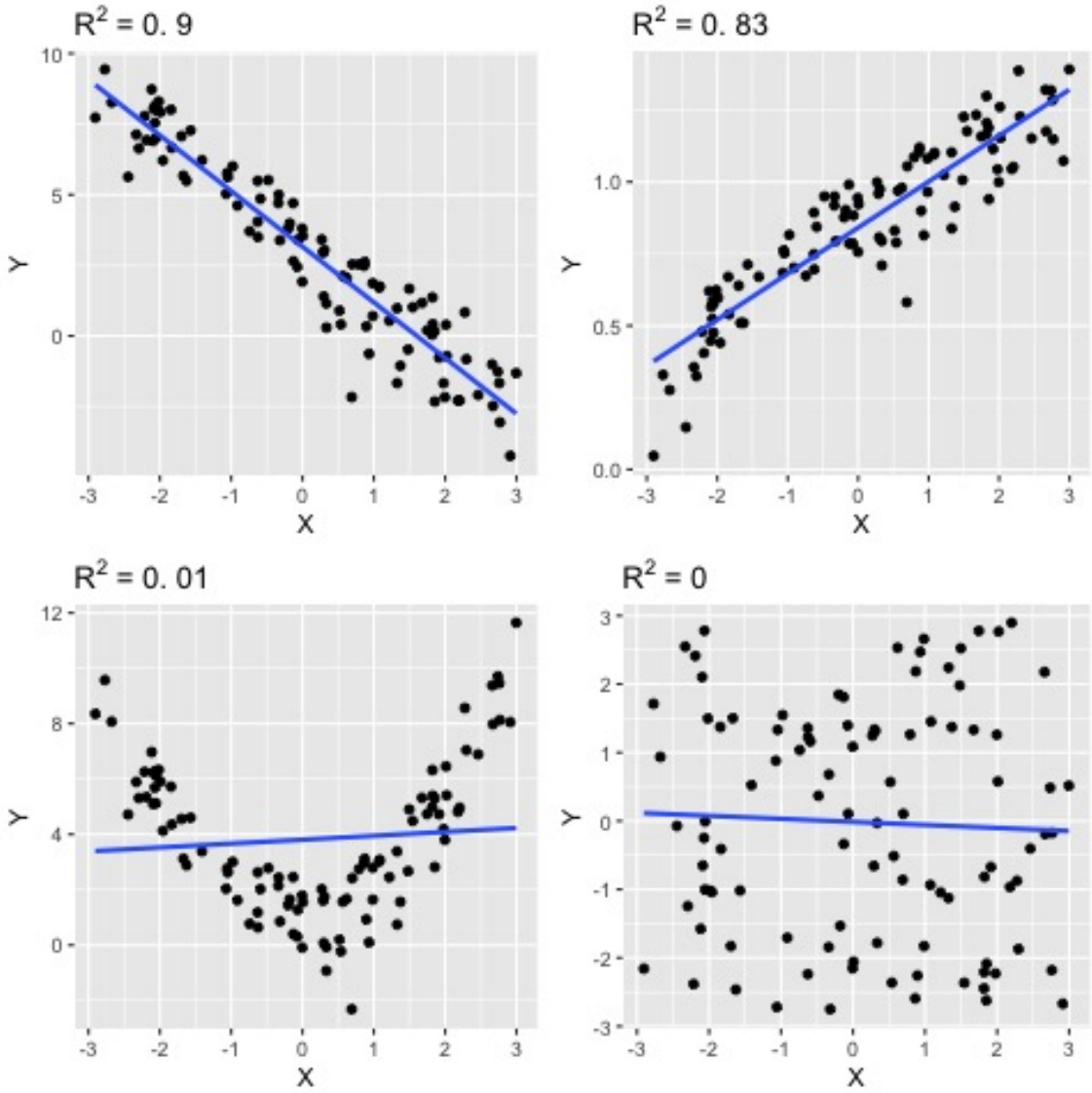


Figure: Examples of coefficients of determination

Evaluate the significance

We have estimated β_1, β_2 and σ^2, and we are also able to calculate the coefficient of determination ... But we cannot test the parameters or establish confidence intervals on these parameters.

To be able to do this, we will therefore add a law hypothesis : we consider that

$$\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

In this framework, $(\varepsilon_i)_{i \in \{1,\ldots,n\}}$ is an i.i.d. sample of law $\mathcal{N}\left(0, \sigma^2\right)$

Test the significance of β_1 and β_2

For $j \in \{1, 2\}$, we test

$$H_0 : \beta_j = 0 \qquad H_1 : \beta_j \neq 0$$

We use as test statistic:

$$T_j = \frac{\widehat{\beta}_j}{s_{\widehat{\beta}_j}}$$

where the estimation of the error variance is

$$s^2 = \frac{SSE}{n-2}$$

and

$$s(\widehat{\beta}_1)^2 = \frac{s^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$s(\widehat{\beta}_2)^2 = \frac{s^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

The larger this quantity is, the more one is inclined to reject the hypothesis of nullity of the parameter.

Test the significance of β_1 and β_2

It can be shown that, under H_0, T_j has Student's law with $n-2$ degrees of freedom.

$$T_j \sim \mathcal{T}(n-2)$$

We decide to reject H_0 at the α test level if

$$|t_j| > t_{n-2,1-\frac{\alpha}{2}}$$

where $t_{n-2,1-\frac{\alpha}{2}}$ designates the quantile of order $1-\frac{\alpha}{2}$ of the $\mathcal{T}(n-2)$ law. In practice, to say that we reject this hypothesis H_0 for β_2 is equivalent to keeping the variable X as explanatory.

Obtain confidence intervals for β_1 and β_2

For $j \in \{1, 2\}$, the parameter β_j admits as a confidence interval of level $1 - \alpha$:

$$\left[\widehat{\beta}_j - t_{n-2,1-\frac{\alpha}{2}} s(\widehat{\beta}_j); \widehat{\beta}_j + t_{n-2,1-\frac{\alpha}{2}} s(\widehat{\beta}_j)\right]$$

It is also possible to establish a simultaneous confidence region for the two parameters β_1 and β_2 (via a Fisher's law), as well as a confidence interval for σ^2 (via a chi-square law).

Analyze the results

After estimating a linear regression model, the next step is to analyze the results:

- The significance of the parameters: a correct model must have significant parameters.
- The atypicality and possible influence of certain data: atypical and influential data can be removed.
- Possible problems of collinearity.
- Possible problems of heteroscedasticity (when the variance of the residuals cannot be considered as constant).

Exercice: determine the height of a tree using regression

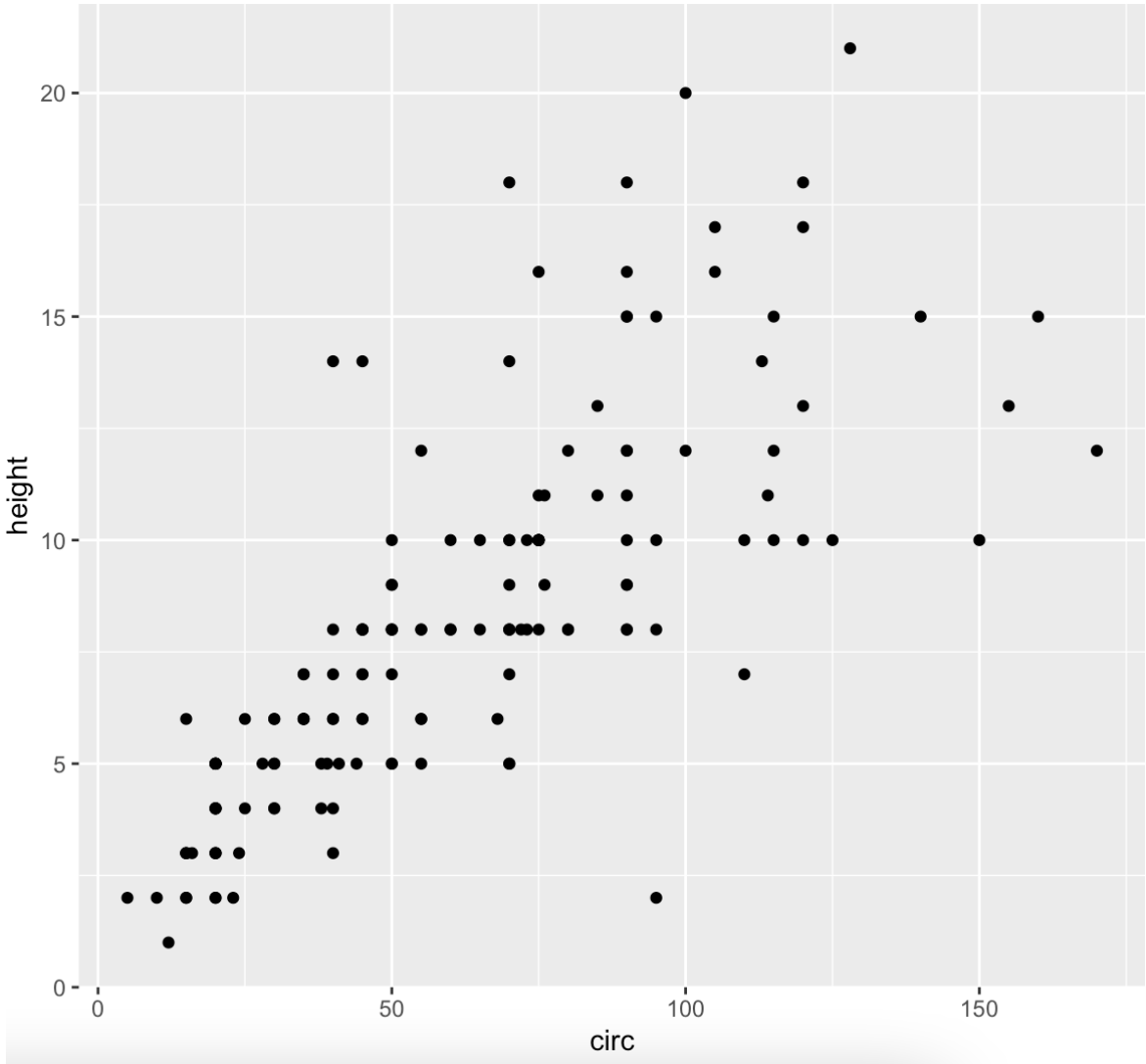
A sawmill wants to determine the height of a tree in order to predict its raw material requirements. They only have data on the circumference of the trees. You are going to create a model to predict the height of a tree based on its circumference.
The sawmill has a data set which contains the following characteristics for 201 spruce trees (https://opendata.paris.fr/explore/dataset/les-arbres/table/). The variables of interest are:

- Circumference : circumference of the tree (in cm)
- heigh : height of the tree (in m)

1. Visualise the scatter plot (graph of height versus circumference).
2. Perform the linear regression of height versus circumference.
3. Give and interpret the coefficient of determination as well as the anova
4. Analyse the significance of the parameters (parameters are tested for nullity at the 5% test level).

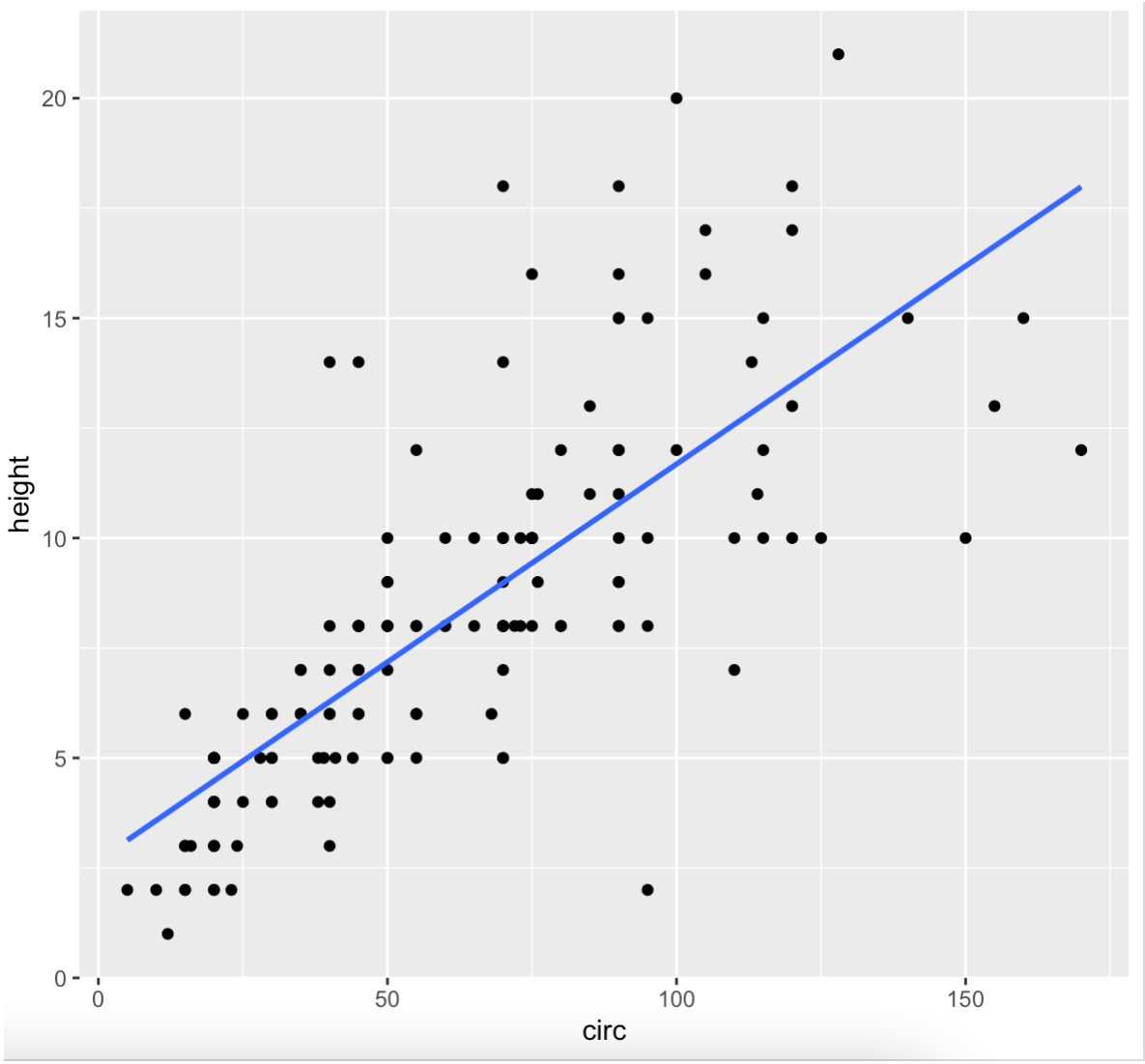
Exercice: determine the height of a tree using regression

Visualise the scatter plot (graph of height versus circumference).



Exercice: determine the height of a tree using regression

Perform the linear regression of height versus circumference.



Exercice: determine the height of a tree using regression

Exploiting what R is given us

```
> anova(simple_reg)
Analysis of Variance Table

Response: height
           Df Sum Sq Mean Sq F value    Pr(>F)
circ        1 1508.1 1508.13  197.57 < 2.2e-16 ***
Residuals 148 1129.8    7.63
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SSM explained variance)

n-2

SSE (residual variance)

Exercice: determine the height of a tree using regression

```
Call:
lm(formula = height ~ circ, data = myData)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2321 -1.6180 -0.2804  1.1280  9.0187

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.679057   0.455838   5.877 2.66e-08 ***
circ        0.090032   0.006405  14.056  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.763 on 148 degrees of freedom
Multiple R-squared:  0.5717,    Adjusted R-squared:  0.5688
F-statistic: 197.6 on 1 and 148 DF,  p-value: < 2.2e-16
```

s(beta_1)

T(beta_1)

beta_1

beta_2

R^2

s

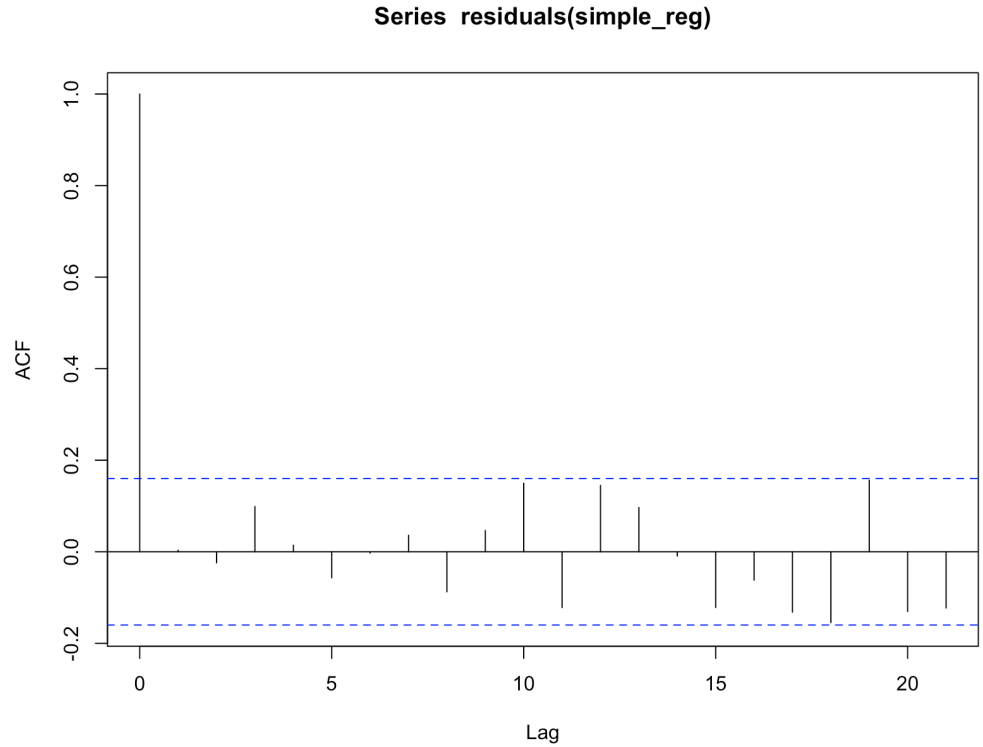
Exercice: determine the height of a tree using regression

Checking the hypothesis

- Identification of outliers and points that strongly contribute to the determination of the model:
- Check the hypothesis on the errors: iid and normality (preliminary to the interpretation of the tests)

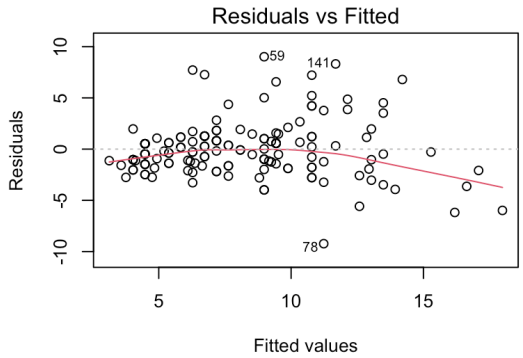
Exercice: determine the height of a tree using regression

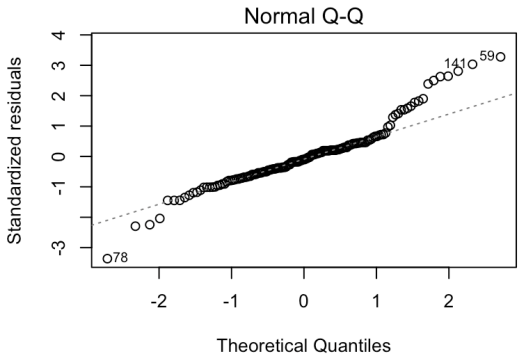
Evaluation of the hypothesis of independence of residuals

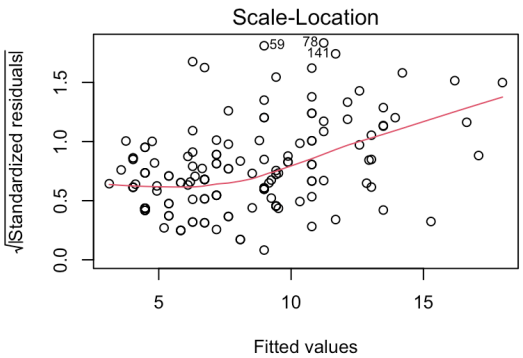


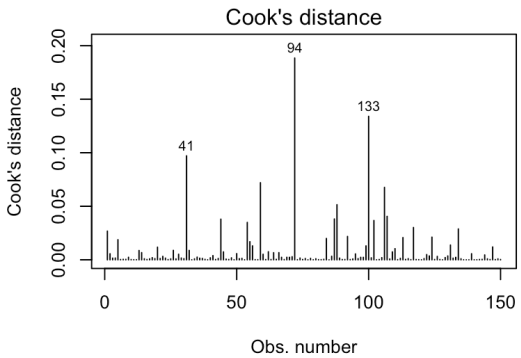
The horizontal dotted lines are the confidence intervals of the correlation coefficient equal to 0. The vertical lines represent the correlation coefficients between the residuals of each point and those of the points of the following line (lag=1), or those separated by two lines (lag=2) etc...

Exercice: determine the height of a tree using regression









- Plot 1 : must be without structure distributed on both sides of the x-axis
- Plot 2 : must follow the bisector
- Plot 3 : must be without structure
- Plot 4 : Cook's distances : to detect possible outliers

Table of contents

Multivariate linear regression - the model

A natural extension of the simple linear regression model, the multivariate linear regression model assumes that :

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

where:

- Y is a v.a.r., observable ;
- $(X_1, \ldots, X_p)$ are deterministic (non-random), observable;
- $(\beta_1, \ldots, \beta_p\ )$ are unknown parameters (not observable)
- ε, the error of the model, is a centred v.a.r. of unknown variance σ^2 (it is also a parameter of the model).

Multivariate linear regression - the data

We consider here that we have n observations $(x_{i1}, \ldots, x_{ip}, y_i)_{i \in \{1,\ldots,n\}}$ of an i.i.d. sample of $(X_1, \ldots, X_p, Y)$;

$$\forall i \in \{1, \ldots, n\} : y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

In the same way as the simple linear regression, the errors $(\varepsilon_i)_{i \in \{1,\ldots,n\}}$ verify for $(i,j) i \in \{1, \ldots, n\}^2$

- $\mathbb{E}(\varepsilon_i) = 0$ (they are centred around 0)
- $\mathrm{Var}(\varepsilon_i) = \sigma^2$ (their variance, unknown, is constant and equal to σ^2);
- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ if $i \neq j$ (they have no linear dependence).

Matrix writting

Mathematically, we can rewrite the problem in the following form:

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \mathbb{X} = \begin{bmatrix} x_{11} \dots x_{1p} \\ \vdots\vdots \\ \vdots \\ x_{n1} \dots x_{np}. \end{bmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Multivariate linear regression with or without constant

In the presence of a constant term in the model, we will consider that the first variable X_1 is equal to 1 :

$$\forall i \in \{1, \ldots, n\} : x_{i1} = 1$$

We are then in the presence of $\boldsymbol{p} - \mathbf{1}$ true explanatory variables and $\boldsymbol{p}$ parameters to be estimated (with in addition σ^2 which remains to be estimated whatever the case).

Apply the Ordinary Least Squares method

The ordinary least squares (OLS) estimators of $= (\beta_1, \ldots, \beta_p)^\top$ the vector $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_1, \ldots, \widehat{\beta}_p\right)^\top$ minimizing the criteria:

$$S(\beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

The solution is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \mathbf{Y}$$

The matrix X is made up of all the variables observed on all the individuals. The matrix-column Y is given by the set of y values observed on all the individuals.

Apply the Ordinary Least Squares method

As for the rest, everything works like the simple linear regression:

- The fitted values (or estimated values) are obtained from the following formula:

$$\widehat{\mathbf{Y}} = \mathbb{X}\widehat{\boldsymbol{\beta}}$$

  These are always the values that would have been obtained for all observations from the regression model.

- The residuals always measure the differences between the observed values (for Y) and the estimated values:

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$$

Statistical properties of the parameters

It can be shown that $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$:

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

This means that, on average, the OLS estimator will lead us to the correct solution.
The residual variance is :

$$s^2 = \frac{1}{n-p} \sum_{i=1}^{n} e_i^2$$

This is an unbiased estimator of σ^2.

Coefficient of determination

Again, the coefficient of determination is the following quantity:

$$\mathrm{R}^2 = \frac{\mathrm{SSM}}{\mathrm{SST}}$$

In practice, this coefficient has an important disadvantage: we could artificially introduce pseudo-explicative variables and increase the coefficient of determination.

- The greater the number of variables, the lower the adjustment error and therefore the coefficient of determination close to 1.
- However, the predictive quality of the model decreases, making the model less robust

In order to take into account the number of explanatory variables, we often consider the adjusted coefficient of determination:

$$\mathrm{R}^2_{\text{adjust}} = 1 - \frac{n}{n-p}\left(1 - \mathrm{R}^2\right)$$

Testing the model

We add the following hypothesis

$$\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

We therefore deduce that $(\varepsilon_i)_{i \in \{1,\ldots,n\}}$ is an i.i.d. sample of law $\mathcal{N}\left(0, \sigma^2\right)$.
To test the significance of the model, we have 2 levels:

- A global test, obtained thanks to a Fisher statistic. In practice, the H_0 hypothesis of this test is often rejected, so the model is often significant globally.
- A test of significance on each of the explanatory variables taken one by one. In this case, it is a Student's t test, just like in simple linear regression. Here, testing one of the parameters has a real meaning: if a variable is not significant, it must be removed from the model. If it is not removed, it is possible that the prediction error of the model is higher.

Test the overall significance of the model

In the case of the regression with constant, we test:

$$\begin{cases} H_0 : \beta_2 = \ldots = \beta_p = 0 \\ H_1 : \exists j \in \{2, \ldots, p\} / \beta_j \neq 0 \end{cases}$$

The test statistic used is :

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

It can be shown that under H_0 :

$$F \sim \mathcal{F}(p-1, n-p)$$

We decide to reject H_0 at the α test level if $f > f_{(p-1,n-p),1-\alpha}$.
The overall significance test is usually presented in the form of the analysis of variance table, with the notations $\mathrm{MSM} = \frac{\mathrm{SSM}}{p-1}$ and $\mathrm{MSE} = \frac{\mathrm{SSE}}{n-p}$.
Here is the analysis of variance table:

Source	*df*	*SC*	*MS*	*F*	*p*-valeur
Model	$p-1$	*SSM*	MSM	$\frac{MSM}{MSE}$	$\mathbb{P}(\mathcal{F}(p-1, n-p) > f)$
Residuals	$n-p$	*SSE*	MSE		
Total	$n-1$	*SST*			

Test the significance of one of the parameters

We test:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

for $j \in \{1, \ldots, p\}$ The test statistic used is:

$$T_j = \frac{\widehat{\beta}_j}{s\sqrt{(\mathrm{X}^\top \mathrm{X})^{-1}_{jj}}}$$

where $\left(\mathbb{X}^\top \mathbb{X}\right)^{-1}_{jj}$ denotes the j-th diagonal term of the matrix $\left(\mathbb{X}^\top \mathbb{X}\right)^{-1}$.
It can be shown that under H_0 :

$$T \sim \mathcal{T}(n-p)$$

We decide to reject H_0 at the test level α if $t > t_{n-p,1-\alpha}$.
In practice, to say that we reject this hypothesis $\boldsymbol{H_0}$ is to keep the variable X_j as explanatory.
The test is passed, congratulations! We can now analyse our results.

Analyze the results

After estimating a linear regression model, the next step is to analyze the results:

- The significance of the parameters: a correct model must have significant parameters.
- The atypicality and possible influence of certain data: atypical and influential data can be removed.
- Possible problems of collinearity.
- Possible problems of heteroscedasticity (when the variance of the residues cannot be considered as constant).

In the case where several regressors are available, it is necessary to choose the best model by using a choice criterion and a search algorithm.

Automatically select a model

- To use our model for forecasting purposes, it is best to make it as parsimonious as possible, i.e. to include only those variables which are really useful.
- If you want to test all possible models, depending on the size of the dataset (i.e. number of individuals and variables), this can be very expensive in terms of time and computer memory.
- The idea here is to use a model search algorithm that will seek to optimise a certain statistical criterion.

Automatically select a model

As far as statistical criteria are concerned, we generally consider criteria such as R^2. However, we know that this is not the most sensible indicator. One can also use information such as AIC or BIC and Mallows' C_p. The latter criteria seek a compromise between :

- the fit of the model ;
- the quality of the forecast.

Indeed, these two quantities do not vary in the same direction: when the number of parameters increases, the better the adjustment, but the worse the forecast.

Determine the adapted regressors

Suppose there are K potential regressors and you want to determine the k best-fit regressors. The following selection criteria can be considered for each of the models tested:

- the coefficient of determination (adjusted or not) ;
- the residual variance;
- criteria based on Kullback information (AIC or BIC, for example);
- the Mallows statistic.

Study selection criteria in detail

These criteria are a compromise between the adjustment of the model (a low residual variance is preferred) and the parsimony of this model (a low number of explanatory variables is preferred). They include (for a $\mathcal{M}_k$ model with k explanatory variables):

- The Akaike Criterion (AIC: Akaike Information Criterium):

$$\mathrm{AIC}\left(\mathcal{M}_k\right) = n \ln\left(s^2_{\mathcal{M}_k}\right) + 2k$$

- The Schwarz Criterion (BIC: Bayesian Information Criterium, also noted SBC for Schwarz Bayesian Criterium) :
- The Mallows statistic: $\mathrm{C_p}$ is defined by:

$$\mathrm{C_p}\left(\mathcal{M}_k\right) = \frac{\mathrm{SSR}\left(\mathcal{M}_k\right)}{s^2} - n + 2k$$

where s^2 is the variance of the complete model (with all regressors) and $\mathrm{SRS}\left(\mathcal{M}_k\right)$ is the sum of the squares of the residuals of the model $\mathcal{M}_k$.

Discover the iterative choice procedures

The most complete procedure, but also the most tedious, consists in selecting the model which minimizes one of the previous criteria for all the potential regression models with k regressors, for $k \in \{1, \ldots, K\}$.
Be aware that there are also alternative procedures:

- Forward procedure: The procedure is initialized by integrating only the constant, then the regressors are introduced one by one, the principle being to retain at each step the variable which contributes most to increasing the sum of the explained squares.
- Top-down or backward procedure The procedure is initialized by integrating all the regressors, then the regressor associated with the smallest decrease in the sum of squares explained is eliminated at each step (the constant is always retained).
- Stepwise procedure. This method is a forward selection procedure, with the possibility of eliminating variables that have become insignificant (in a backward step).

Exercise: Improve tree height predictions

In the previous section, you determined the height of a tree as a function of its circumference, using simple linear regression. In this activity, you will go one step further and improve your prediction, this time using a multivariate linear regression!

1. Add a column to the sample. Name it circ_sqrt and fill it with the square root of the circumference of each tree.
2. Perform the multivariate linear regression of height on the basis of:
   - of circumference ;
   - of circ_sqrt.
3. Analyse the significance of the parameters, and remove any non-significant parameters.
4. Give and interpret the coefficient of determination of the model finally retained.

Conclusion

1. You need a model to perform your regression
2. You need to check whether the underlying hypothesis of this model are reasonable or not

This model will allow you to:

1. Assess and quantify the effect of parameters on the response
   - Parameters are estimated as a whole, using all the measurements
2. Extrapolate within the range of parameters you tried
3. Detect outstanding points (those with a high residual and/or with a high lever)

This model will guide on how to design your experiments:

- e.g., the linear model assumes some uniformity of interest over the parameter space range
- if your system is heteroscedastic, you should perform more measurements for parameters that lead to higher variance