People's Democratic Republic of Algeria
University Med Khider of Biskra
Faculty of SNVSTU

Linear Regression and Correlation

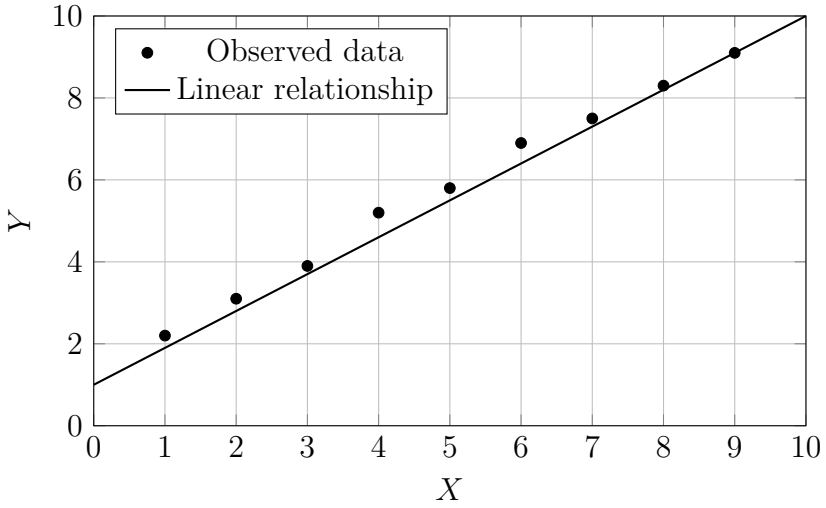
Dr. Ben Gherbal Hanane

Biostatistics Level: M1 Biology
Email: hanane.bengherbal@univ-biskra.dz

1. Simple Linear Regression

Suppose we have a sample of n individuals. For each individual i ($i = 1, \ldots, n$), two variables are measured: y_i, corresponding to the realization of the quantitative variable Y, and x_i, representing the value of the explanatory variable X.

The objective is to analyze the relationship between these two variables, in particular the influence of X on Y.



Assume a linear model (that is, Y is expressed linearly as a function of X) of the form:

$$Y = \alpha X + \beta$$

where α is the slope and β is the intercept.

These two parameters can be estimated by a and b, and we obtain the regression line in the least squares sense from a sample of size n.

Using the least squares method to determine a and b, the estimators of the parameters α and β are those that minimize the function $Q(\alpha, \beta)$ defined by:

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (Y_i - \beta - \alpha x_i)^2$$

This amounts to determining the minimum of the quadratic error function $Q(\alpha, \beta)$, which requires solving the following system:

$$\begin{cases} \dfrac{\partial Q(\alpha, \beta)}{\partial \alpha} = 0 \\ \dfrac{\partial Q(\alpha, \beta)}{\partial \beta} = 0 \end{cases}$$

Solving this system gives:

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{X}\,\overline{Y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$$

$$b = \frac{1}{n}\sum_{i=1}^{n} y_i - a\frac{1}{n}\sum_{i=1}^{n} x_i = \overline{Y} - a\overline{X}$$

2. Correlation and Determination

2.1. The Correlation Coefficient

The correlation coefficient measures the strength and direction of the linear relationship between two quantitative variables. In a linear framework, it is defined as

$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y},$$

where $\mathrm{Cov}(X,Y)$ is the covariance between X and Y, and σ_X, σ_Y are their standard deviations.

The correlation coefficient satisfies

$$-1 \leq r \leq 1.$$

Its sign indicates the direction of the linear association:

- $r > 0$: the variables tend to increase or decrease together (positive linear relationship).
- $r < 0$: one variable tends to increase when the other decreases (negative linear relationship).

The magnitude $|r|$ reflects the strength of the linear relationship. Values of $|r|$ close to 1 indicate a strong linear association, whereas values close to 0 indicate a weak or no linear association.

It is important to note that correlation measures only linear dependence and does not imply causation.

2.2. The Coefficient of Determination

The coefficient of determination, denoted by R, measures the proportion of the total variability in the dependent variable Y that is explained by the linear regression model.

In simple linear regression, it is given by

$$R = r^2,$$

where r is the linear correlation coefficient between X and Y.

The value of R satisfies

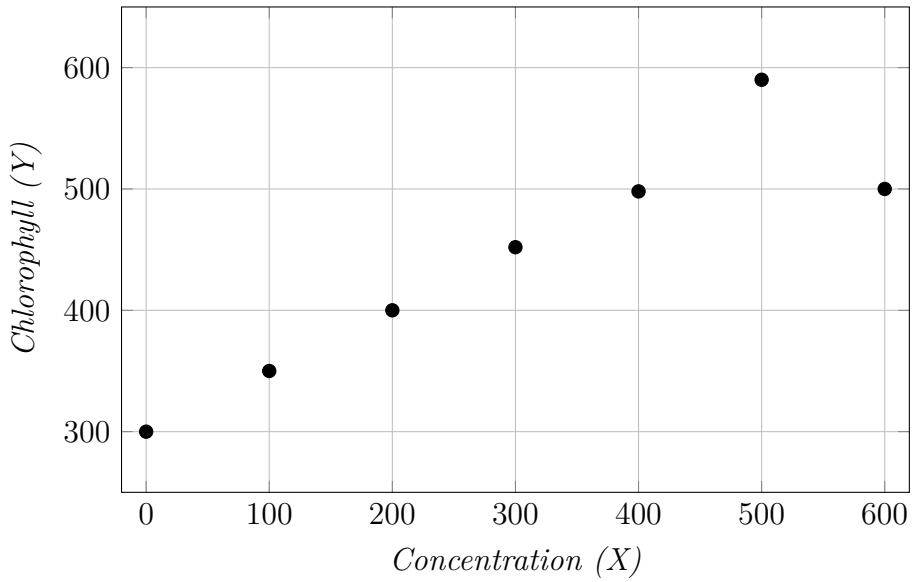
$$0 \leq R \leq 1.$$

An R close to 1 indicates that a large proportion of the variability of Y is explained by the model, whereas an R close to 0 indicates that the model explains little of the variability.

For example, $R = 0.93$ means that 93% of the variability of Y is explained by the linear relationship with X, while the remaining 7% is due to other factors or random error.

Exercise 1. *A study was conducted to analyze the effect of phosphate concentration on chlorophyll production of a certain type of algae. The results obtained are presented in the following table:*

X_i	0	100	200	300	400	500	600
Y_i	300	350	400	452	498	590	500



Solution

We have $n = 7$ observations:

$$(0, 300), (100, 350), (200, 400), (300, 452), (400, 498), (500, 590), (600, 500).$$

$$\sum x_i = 2100, \quad \sum y_i = 3090, \quad \sum x_i^2 = 910000,$$
$$\sum y_i^2 = 1422908, \quad \sum x_i y_i = 1044800.$$

$$\overline{X} = \frac{2100}{7} = 300, \qquad \overline{Y} = \frac{3090}{7} = 441.4286.$$

$$\widehat{\alpha} = \frac{\frac{1}{n}\sum x_i y_i - \overline{X}\,\overline{Y}}{\frac{1}{n}\sum x_i^2 - \overline{X}^2} = \frac{\frac{1044800}{7} - 300 \cdot \frac{3090}{7}}{\frac{910000}{7} - 300^2} = \frac{\frac{117800}{7}}{40000} = \frac{589}{1400} \approx 0.420714.$$

$$\widehat{\beta} = \overline{Y} - \widehat{\alpha}\,\overline{X} = \frac{3090}{7} - \frac{589}{1400} \cdot 300 = \frac{4413}{14} \approx 315.2143.$$

$$\widehat{Y} = 0.420714\,X + 315.2143.$$

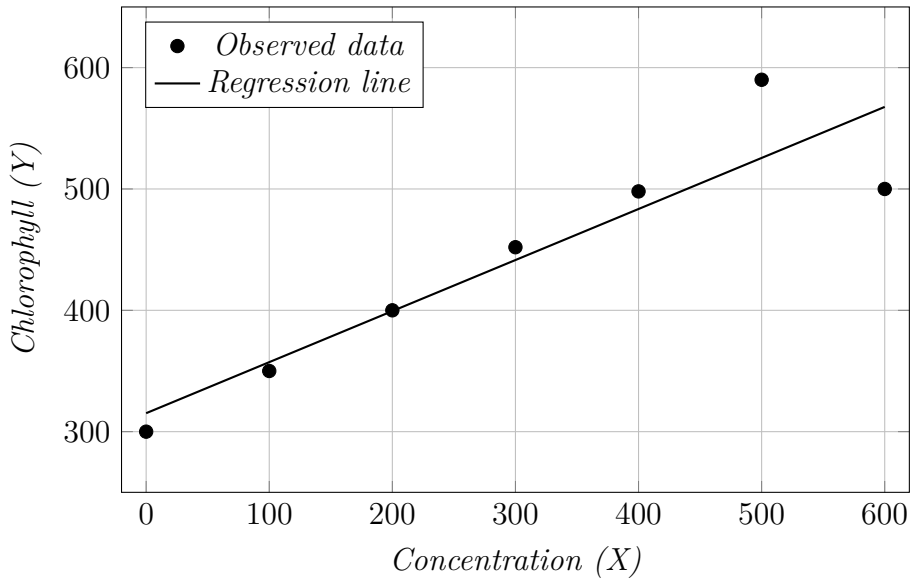
For the correlation coefficient:

$$\mathrm{Cov}(X, Y) = \frac{1}{n}\sum x_i y_i - \overline{X}\,\overline{Y} = \frac{117800}{7}, \quad \mathrm{Var}(X) = \frac{1}{n}\sum x_i^2 - \overline{X}^2 = 40000.$$

$$\mathrm{Var}(Y) = \frac{1}{n}\sum y_i^2 - \overline{Y}^2 = \frac{1422908}{7} - \left(\frac{3090}{7}\right)^2 = \frac{412256}{49}.$$

$$r = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \approx 0.91734, \qquad R = r^2 \approx 0.84152.$$

Conclusion: *The correlation is positive and strong, but it is reduced due to the presence of a relatively distant point (outlier) and another point that deviates moderately from the line.*



3. Multiple Linear Regression

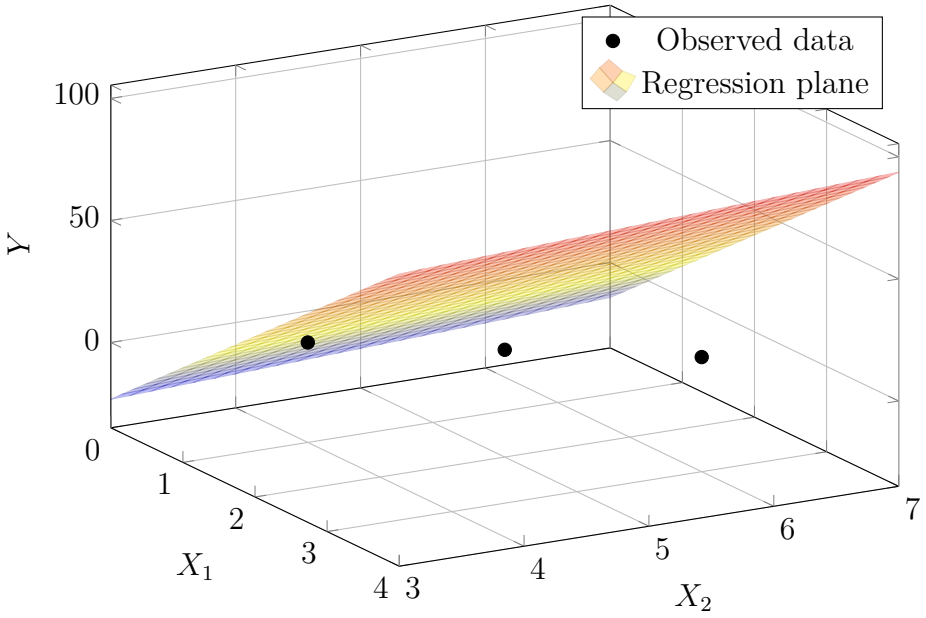
Multiple linear regression is a statistical method that allows modeling the relationship between a dependent variable Y (e.g., plant growth rate) and several explanatory variables $X_1, X_2, ..., X_p$ (e.g., light, water, temperature...). It is used to:

1. predict a value of Y.
2. understand which factors influence Y.
3. quantify the effect of each explanatory variable.

The multiple linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon,$$

- Y dependent variable,
- X_i explanatory variables,
- β_i unknown coefficients to estimate,
- ε random error.



In matrix form:

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

where

- Y is an $n \times 1$ vector,
- X is an $n \times (p+1)$ matrix, the first column is filled with 1 for β_0,
- β is a $(p+1) \times 1$ vector,
- ε is the error vector.

Estimation of the coefficients
We use the least squares method: it minimizes the sum of squared errors ε.
Estimator formula:

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

3.1. Model with Two Explanatory Variables

Special case of the model with two explanatory variables X_1 and X_2:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

We construct the matrix X as:

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ 1 & \cdot & \cdot \\ & \cdot & \cdot \\ 1 & X_{n1} & X_{n2} \end{bmatrix}$$

Example 1. *Suppose you want to study how plant growth (in cm) depends on temperature (in °C) and the amount of water received (in mL/day).*

We therefore have:
Dependent variable: growth Y
Two explanatory variables: temperature X_1 and water X_2.

X_1	X_2	Y
1	4	7
2	5	10
3	7	14

The model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

In matrix form:

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

Step 1: Build X and Y

$$X = \begin{bmatrix} 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 7 \end{bmatrix}, \qquad Y = \begin{bmatrix} 7 \\ 10 \\ 14 \end{bmatrix}.$$

Step 2: Compute X^TX and X^TY

$$X^T = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 4 & 5 & 7 \end{bmatrix}.$$

$$X^TX = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 4 & 5 & 7 \end{bmatrix} \begin{bmatrix} 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 7 \end{bmatrix} = \begin{bmatrix} 3 & 6 & 16 \\ 6 & 14 & 35 \\ 16 & 35 & 90 \end{bmatrix}.$$

$$X^TY = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 4 & 5 & 7 \end{bmatrix} \begin{bmatrix} 7 \\ 10 \\ 14 \end{bmatrix} = \begin{bmatrix} 7+10+14 \\ 1\cdot 7+2\cdot 10+3\cdot 14 \\ 4\cdot 7+5\cdot 10+7\cdot 14 \end{bmatrix} = \begin{bmatrix} 31 \\ 69 \\ 176 \end{bmatrix}.$$

Step 3: Invert X^TX

$$\left(X^TX\right)^{-1} = \begin{bmatrix} 35 & 20 & -14 \\ 20 & 14 & -9 \\ -14 & -9 & 6 \end{bmatrix}.$$

Step 4: Compute $\widehat{\beta} = (X^TX)^{-1}X^TY$

$$\begin{aligned} \widehat{\beta} &= \begin{bmatrix} 35 & 20 & -14 \\ 20 & 14 & -9 \\ -14 & -9 & 6 \end{bmatrix} \begin{bmatrix} 31 \\ 69 \\ 176 \end{bmatrix} \\ &= \begin{bmatrix} 35\cdot 31+20\cdot 69-14\cdot 176 \\ 20\cdot 31+14\cdot 69-9\cdot 176 \\ -14\cdot 31-9\cdot 69+6\cdot 176 \end{bmatrix} \\ &= \begin{bmatrix} 1085+1380-2464 \\ 620+966-1584 \\ -434-621+1056 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}. \end{aligned}$$

Thus,

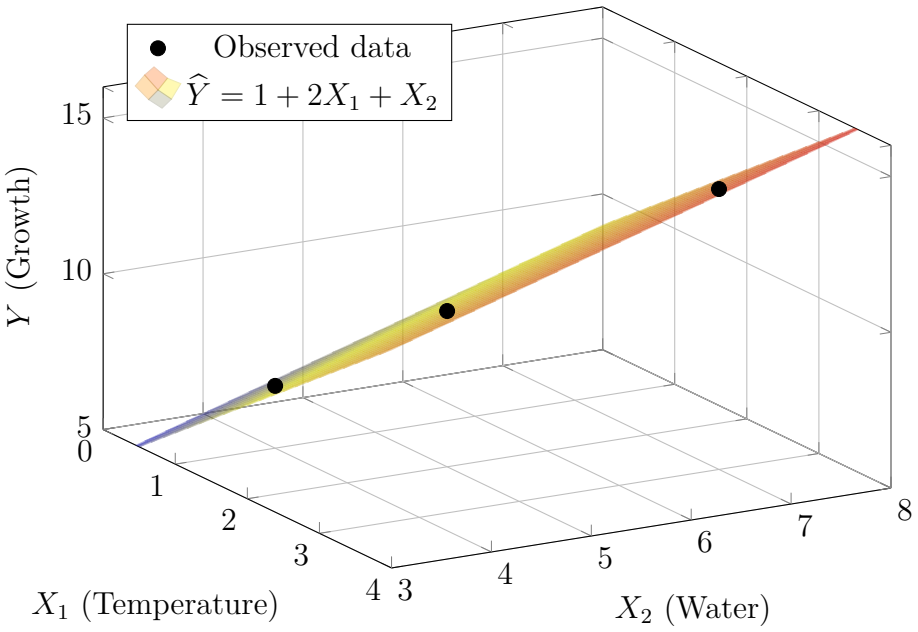
$$\widehat{\beta}_0 = 1, \qquad \widehat{\beta}_1 = 2, \qquad \widehat{\beta}_2 = 1.$$

Estimated model

$$\widehat{Y} = 1 + 2X_1 + X_2.$$

Check (predicted values and residuals)

(X_1, X_2, Y)	$(1,4,7)$	$(2,5,10)$	$(3,7,14)$
$\widehat{Y} = 1 + 2X_1 + X_2$	7	10	14
$\varepsilon = Y - \widehat{Y}$	0	0	0


Observed data
$\widehat{Y} = 1 + 2X_1 + X_2$
Y (Growth)
15
10
5
0
1
2
3
4
X_1 (Temperature)
3
4
5
6
7
8
X_2 (Water)