Study of a Statistical Series with Two Variables

Linear Regression

Benchikh Tawfik

Faculty of Medicine
1st Year Medicine

October 19, 2025

Plan du cours

1. Statistical Series with Two Variables
2. Linear Regression
3. Exercise

Plan du cours

1. Statistical Series with Two Variables

2. Linear Regression

3. Exercise

Plan du cours

1. Statistical Series with Two Variables

2. Linear Regression

3. Exercise

Plan de cours

1. Statistical Series with Two Variables
   - Introduction
   - Presentation of a Bivariate Series
   - Graphical Representation
   - Mean, Covariance, and Correlation Coefficient
   - Different Types of Scatter Plots
2. Linear Regression
   - Linear Regression
   - Computation of the Regression Line Parameters
   - Nonlinear Regression
3. Exercise

Introduction

- Univariate statistics focus on a single characteristic within a given population, such as the number of children per family or employee salaries within a company.
- When the analysis concerns the **joint study of two characteristics** observed on the same population, we deal with **bivariate statistics**, involving the analysis of **two-variable statistical series**.

General Objective

- The population is studied through **data** collected from a **random sample** of individuals drawn from it.
- **The data** arise from two variables measured simultaneously on the same individuals, providing partial information about the underlying population.
- The objective is to answer specific questions about this population. To achieve this:
  - ⋆ The first step is **descriptive analysis**, which consists in examining the observed data using appropriate descriptive statistical methods.

Descriptive Methods for Bivariate Statistical Series

- The objective is to **visualize** and **quantify through summary measures** the possible **relationships** between two variables observed on sampled data:
  - ⋆ **Graphical representations** and **statistical indices** provide **partial information** about the relationship in the population and serve as a preliminary step toward **inferential analysis**.
- **Modeling relationships** between a quantitative variable to be **explained** and another quantitative variable used as an **explanatory variable**:
  - ⋆ This leads to the use of **linear regression methods**.

PLAN DE COURS

1. STATISTICAL SERIES WITH TWO VARIABLES
   - Introduction
   - Presentation of a Bivariate Series
   - Graphical Representation
   - Mean, Covariance, and Correlation Coefficient
   - Different Types of Scatter Plots
2. LINEAR REGRESSION
   - Linear Regression
   - Computation of the Regression Line Parameters
   - Nonlinear Regression
3. EXERCISE

Definition

- We consider:
  - a population of size n,
  - two statistical variables X and Y (two characteristics).
- Each individual in this population is labeled by an index between 1 and n.
- For each individual i ($1 \leq i \leq n$), we observe a pair (x_i, y_i), where x_i is the observed value (modality) of variable X, and y_i is the corresponding value of variable Y for the same individual.
- The collection of all pairs (x_i, y_i) defines a **bivariate statistical series**.

Example 1

- A doctor measures the systolic blood pressure (Y) of 12 women of different ages (X). The results are as follows:

x (years)	56	42	72	36	63	47
y (mm Hg)	147	125	160	118	149	128
x (years)	55	49	38	42	68	60
y (mm Hg)	150	145	115	140	152	155

⋆ **Population:** women (from a specific city or country).

⋆ **Variable 1:** age (X).

⋆ **Variable 2:** systolic blood pressure (Y).

Example 2

- The following table shows the evolution of life expectancy at birth (in years) for men in Algeria from 2000 to 2009:

X (Year)	2000	2001	2002	2003	2004
Y (Life expectancy)	76.9	77.1	77.3	77.5	77.6
X (Year)	2005	2006	2007	2008	2009
Y (Life expectancy)	77.8	78.0	78.1	78.2	78.2

- **Population:** men in Algeria.
- **Variable 1:** year (X).
- **Variable 2:** life expectancy (Y).

DEFINITION

- When one of the two variables represents time (for example, a year or a date), the bivariate statistical series is called a **time series** or **chronological series** (as in Example 2).

Remark

- From a bivariate statistical series, one can derive the corresponding **univariate series** describing each variable X and Y separately.

x (years)	56	42	72	36	63	47	55	49	38	42	68	60

y (mm Hg)	147	125	160	118	149	128	150	145	115	140	152	155

PLAN DE COURS

1. STATISTICAL SERIES WITH TWO VARIABLES
   - Introduction
   - Presentation of a Bivariate Series
   - Graphical Representation
   - Mean, Covariance, and Correlation Coefficient
   - Different Types of Scatter Plots
2. LINEAR REGRESSION
   - Linear Regression
   - Computation of the Regression Line Parameters
   - Nonlinear Regression
3. EXERCISE

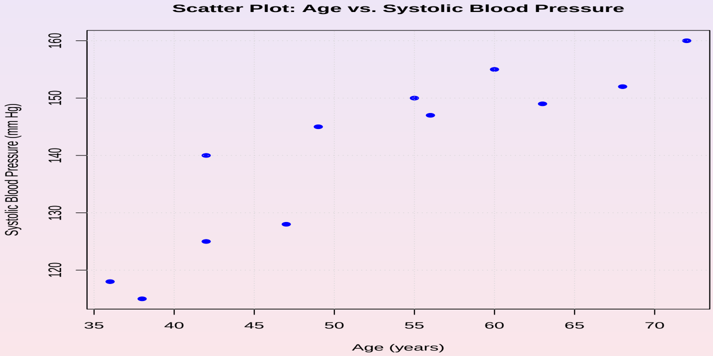
Scatter Plot

- The observations can be represented on a **scatter diagram** (or **scatter plot**), also called a **cloud of points**.
- Each point i corresponds to one individual and is defined by the coordinates:

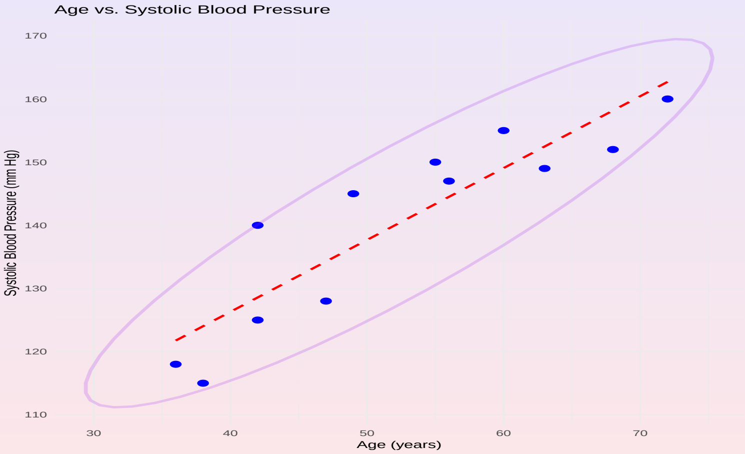
$$\begin{aligned} x_i &= \text{Age}, \\ y_i &= \text{Systolic blood pressure}. \end{aligned}$$

Scatter Plot

- Representation of systolic blood pressure as a function of age:



NUAGE DE POINTS: DROITE DE REGRESSION



Plan de cours

1. Statistical Series with Two Variables
   - Introduction
   - Presentation of a Bivariate Series
   - Graphical Representation
   - Mean, Covariance, and Correlation Coefficient
   - Different Types of Scatter Plots
2. Linear Regression
   - Linear Regression
   - Computation of the Regression Line Parameters
   - Nonlinear Regression
3. Exercise

Mean Point of the Scatter Plot

- The **mean point** $G(x, y)$ of a scatter plot is defined as the point whose coordinates are the means of the two variables X and Y:

$$G = (\overline{X}, \overline{Y})$$

where

$$m_X = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad m_Y = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,.$$

Interpretation

The mean point represents the **center of gravity** of the data cloud.

Mean Point of the Scatter Plot

- The **mean point** $G(x, y)$ of a scatter plot is defined as the point whose coordinates are the means of the two variables X and Y:

$$G = (\overline{X}, \overline{Y})$$

where

$$m_X = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad m_Y = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i \,.$$

Interpretation

The mean point represents the **center of gravity** of the data cloud.

Covariance

- The **covariance** between X and Y is given by:

$$\begin{aligned}\mathrm{Cov}(X,Y) = S_{X,Y} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{X}\,\overline{Y}\,.\end{aligned}$$

- It measures the **joint variability** of the two variables.

Interpretation

- $\mathrm{Cov}(X,Y) > 0$: X and Y tend to increase together (vary in the same direction).
- $\mathrm{Cov}(X,Y) < 0$: when one increases, the other decreases (an opposite trend).
- $\mathrm{Cov}(X,Y) = 0$: no linear relationship between X and Y.

Covariance

- The **covariance** between X and Y is given by:

$$\begin{aligned}\mathrm{Cov}(X,Y) = S_{X,Y} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{X}\,\overline{Y}\,.\end{aligned}$$

- It measures the **joint variability** of the two variables.

Interpretation

- $\mathrm{Cov}(X,Y) > 0$: X and Y tend to increase together (vary in the same direction).
- $\mathrm{Cov}(X,Y) < 0$: when one increases, the other decreases (an opposite trend).
- $\mathrm{Cov}(X,Y) = 0$: no linear relationship between X and Y.

Linear Correlation Coefficient

- For two variables X and Y, the **linear correlation coefficient** $r = \rho(X, Y) = \operatorname{cor}(X, Y)$ is defined as:

$$r = \operatorname{cor}(X, Y) = \frac{\operatorname{cov}(X, Y)}{S_X \, S_Y}$$

- The coefficient r takes values between -1 and 1:
  - $r = 1$: perfect positive linear relationship,
  - $r = -1$: perfect negative linear relationship,
  - $r = 0$: no linear relationship.

Remark

Correlation is a **standardized measure of covariance**, independent of the scale of the variables.

Linear Correlation Coefficient

- For two variables X and Y, the **linear correlation coefficient** $r = \rho(X,Y) = \mathrm{cor}(X,Y)$ is defined as:

$$r = \mathrm{cor}(X,Y) = \frac{\mathrm{cov}(X,Y)}{S_X\, S_Y}$$

- The coefficient r takes values between -1 and 1:
  - $r = 1$: perfect positive linear relationship,
  - $r = -1$: perfect negative linear relationship,
  - $r = 0$: no linear relationship.

Remark

Correlation is a **standardized measure of covariance**, independent of the scale of the variables.

Plan de cours

1 Statistical Series with Two Variables
- Introduction
- Presentation of a Bivariate Series
- Graphical Representation
- Mean, Covariance, and Correlation Coefficient
- Different Types of Scatter Plots

2 Linear Regression
- Linear Regression
- Computation of the Regression Line Parameters
- Nonlinear Regression

3 Exercise

Types of Scatter Plots

- **Example 1:** Evolution of life expectancy at birth (in years) for men in Algeria from 2000 to 2009.

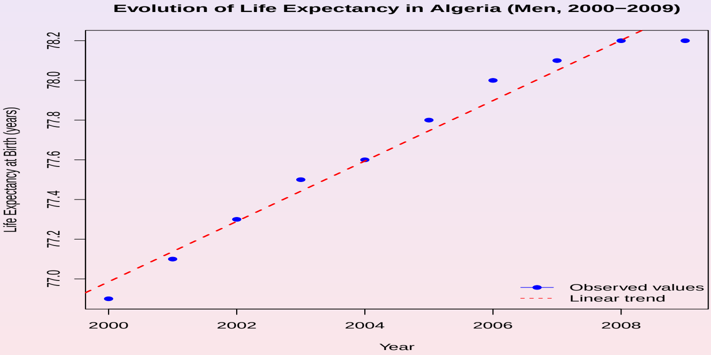


Figure: Life expectancy at birth for men in Algeria (2000-2009).

Example 3

Data

The amount of **metabolized energy** (in calories) over a period of 10 hours was measured for a sparrow exposed to different ambient temperatures (in °C). The results are given below:

x = Temperature (°C)	0	4	10	18	26	32
y = Energy (calories)	25	23	24	19	15	14

Objective

We aim to study the relationship between temperature and the energy metabolized by the sparrow.

EXAMPLE 3

DATA

The amount of **metabolized energy** (in calories) over a period of 10 hours was measured for a sparrow exposed to different ambient temperatures (in °C). The results are given below:

x = Temperature (°C)	0	4	10	18	26	32
y = Energy (calories)	25	23	24	19	15	14

OBJECTIVE

We aim to study the relationship between temperature and the energy metabolized by the sparrow.

- The scatter plot below represents the amount of energy metabolized in 10 hours (in calories) as a function of temperature (°C).

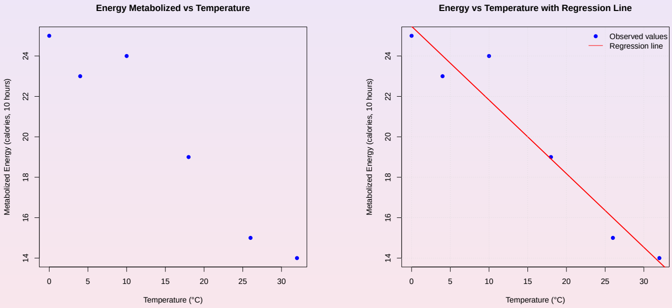


Figure: Relationship between temperature and metabolized energy

INTERPRETATION

As temperature increases, the amount of energy metabolized by the sparrow decreases. The fitted regression line (right plot) confirms a clear negative correlation between the two variables.

Types of Scatter Plots: Example 4

Example 4

We have the following data for a sample of 20 men: their **ages** x_i (in years) and their **heights** y_i (in meters).

X (Age)	22	26	33	35	38	40	42	42	43	44
Y (Height)	1.82	1.71	1.72	1.75	1.77	1.97	1.94	1.76	1.68	1.98
X (Age)	45	48	49	50	51	53	56	61	64	75
Y (Height)	1.79	1.82	1.80	1.87	1.72	1.65	1.90	1.81	1.75	1.68

Objective

We want to study how height varies with age in this sample of adult men.

Types of Scatter Plots: Example 4

Example 4

We have the following data for a sample of 20 men: their **ages** x_i (in years) and their **heights** y_i (in meters).

X (Age)	22	26	33	35	38	40	42	42	43	44
Y (Height)	1.82	1.71	1.72	1.75	1.77	1.97	1.94	1.76	1.68	1.98
X (Age)	45	48	49	50	51	53	56	61	64	75
Y (Height)	1.79	1.82	1.80	1.87	1.72	1.65	1.90	1.81	1.75	1.68

Objective

We want to study how height varies with age in this sample of adult men.

Types of Scatter Plots: Example 4

- **Example 4:** Relationship between age (in years) and height (in meters) for a sample of 20 men.

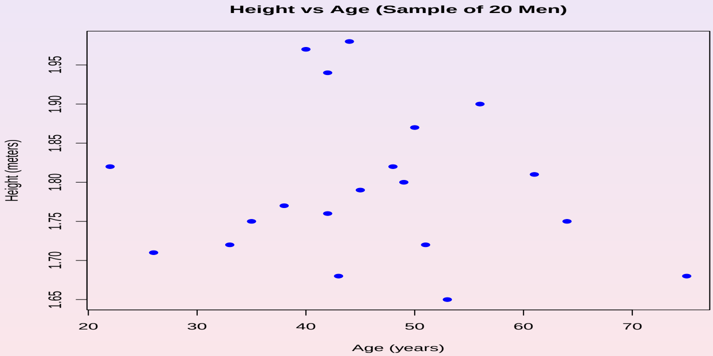


Figure: Scatter plot of height versus age for 20 male subjects.

Nonlinear Regression: Example 5

- In the absence of mortality, we aim to describe how a bacterial population grows over time.
- Daily counts starting from the second day provide the following results:

Data: Bacterial Growth

Days (t)	Bacteria Count (N)
2	55
3	90
4	135
5	245
6	403
7	665
8	1100
9	1810
10	3300
11	4450
12	7350

Scatter Plot: Example 5

- Representation of bacterial population growth as a function of time (in days).

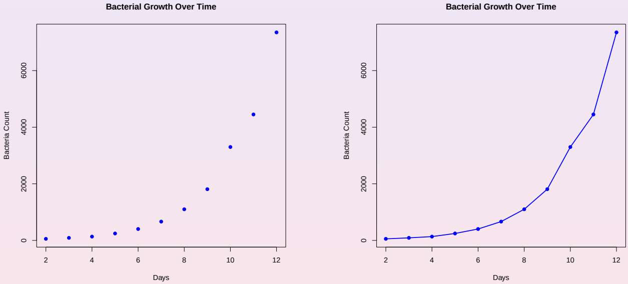


Figure: Scatter plot and nonlinear smoothing of bacterial growth over time.

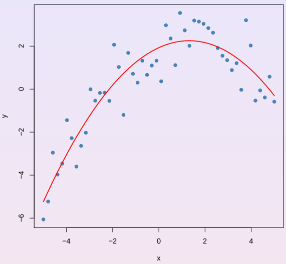
TYPES OF SCATTER PLOTS: REMARKS

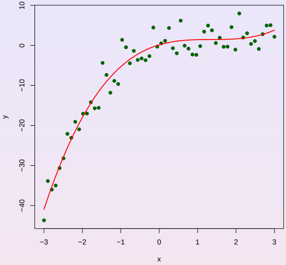
- By observing these graphs, we can see that the type of function to be fitted depends on the data:
  1. For Examples 1 and 2, the points are aligned in an increasing pattern. A suitable model would be an increasing linear function (a straight line).
  2. For Example 3, the points are aligned in a decreasing pattern. A decreasing linear function can be fitted.
  3. For Example 4, the points are scattered randomly with no clear trend.
  4. For Example 5, the number of bacteria increases rapidly-an exponential growth pattern.

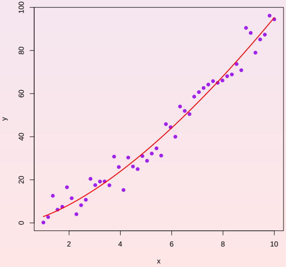
Types of Scatter Plots: Generalization

- When we have a scatter plot (x_i, y_i), several situations may occur:
  - A) The points are randomly distributed: we say that the two variables (e.g., Age and Height) are **independent**.
  - B) The points are distributed around a certain curve: in this case, we can perform a **graphical fitting**, i.e., draw the curve that best represents the data.
    - The simplest curve that can be fitted is a straight line.
    - Sometimes, a parabola, a cubic function, a power function, or an exponential function may provide a better fit.

Other example


Parabolic Fit: y = a + bx + cx²


Cubic Fit: y = a + bx + cx² + dx³


Power Function Fit: y = a * x^b

Plan de cours

1. Statistical Series with Two Variables
   - Introduction
   - Presentation of a Bivariate Series
   - Graphical Representation
   - Mean, Covariance, and Correlation Coefficient
   - Different Types of Scatter Plots
2. Linear Regression
   - Linear Regression
   - Computation of the Regression Line Parameters
   - Nonlinear Regression
3. Exercise

Linear Regression

- **Linear regression:** the simplest regression model:

$$\boxed{Y = f(X) + \varepsilon = \alpha + \beta X + \varepsilon}$$

  - Interpretation
  - Parameter estimation
  - Prediction

LINEAR REGRESSION

- **Linear regression:** the simplest regression model:

$$\boxed{Y = f(X) + \varepsilon = \alpha + \beta X + \varepsilon}$$

  - Interpretation
  - Parameter estimation
  - Prediction

Linear Regression

- **Linear regression:** the simplest regression model:

$$Y = f(X) + \varepsilon = \alpha + \beta X + \varepsilon$$

  - Interpretation
  - Parameter estimation
  - Prediction

LINEAR REGRESSION

- **Linear regression:** the simplest regression model:

$$Y = f(X) + \varepsilon = \alpha + \beta X + \varepsilon$$

  - Interpretation
  - Parameter estimation
  - Prediction

Linear Regression

- α represents the intercept, and β represents the slope of the regression line.
- Greek letters are used to denote the intercept and slope, emphasizing that they are **unknown parameters**.
- Their true values would be known only if the entire population were observed-which is never the case in practice. Therefore, we need to **estimate** them from the available sample.

PLAN DE COURS

1. STATISTICAL SERIES WITH TWO VARIABLES
   - Introduction
   - Presentation of a Bivariate Series
   - Graphical Representation
   - Mean, Covariance, and Correlation Coefficient
   - Different Types of Scatter Plots
2. LINEAR REGRESSION
   - Linear Regression
   - Computation of the Regression Line Parameters
   - Nonlinear Regression
3. EXERCISE

Regression Line

- **Regression line:**
  - Provides the best summary of the scatter plot.
    - $\Rightarrow$ It lies as close as possible to all the data points.
    - $\Rightarrow$ The errors ε_i are as small as possible.

Computation Principle

- Estimate α and β so that the residuals ε_i are as small as possible.

1. Each residual ε_i represents the deviation between the observed point and the regression line:

$$y_i = \alpha + \beta x_i + \varepsilon_i \Rightarrow \varepsilon_i = y_i - (\alpha + \beta x_i) .$$

2. Minimize the **Sum of Squared Errors (SSE)**:

$$SSE = \sum_{i=1}^{n} (\varepsilon_i)^2$$

3. Estimate α and β such that SSE is minimal.

Computation of the Slope β

- The slope β is given by:

$$b = \frac{\mathsf{Cov}(X, Y)}{\mathsf{Var}(X)}$$

- It represents the **average change in Y** for a **one-unit increase in X**.

Estimation of the Intercept α

- The regression line passes through the point of means $(\overline{X}, \overline{Y})$:

$$\overline{Y} = a + b\,\overline{X}$$

- Hence, the intercept is given by:

$$\boxed{a = \overline{Y} - b\,\overline{X}}$$

where

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Estimation of the Intercept α

- The regression line passes through the point of means $(\overline{X}, \overline{Y})$:

$$\overline{Y} = a + b\,\overline{X}$$

- Hence, the intercept is given by:

$$\boxed{a = \overline{Y} - b\,\overline{X}}$$

where

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

EXAMPLE

- Covariance between systolic blood pressure and age:

$$\text{Cov}(\text{Systolic BP}, \text{Age}) = \text{Cov}(X, Y) = S_{XY} = 160.4242$$

- Variance of age: $S^2(\text{Age}) = S^2(X) = 140.9697$
- Estimation of the slope β:

$$b = \frac{\text{Cov}(\text{Systolic BP}, \text{Age})}{S^2(\text{Age})} = \frac{S_{XY}}{S^2(X)} = 1.1380$$

- Estimation of the intercept α:

$$a = \overline{Y} - b\,\overline{X} = M_{\text{Systolic BP}} - b\,M_{\text{Age}} = 80.778$$

Example

- Covariance between systolic blood pressure and age:

$$\text{Cov}(\text{Systolic BP}, \text{Age}) = \text{Cov}(X, Y) = S_{XY} = 160.4242$$

- Variance of age: $S^2(\text{Age}) = S^2(X) = 140.9697$

- Estimation of the slope β:

$$b = \frac{\text{Cov}(\text{Systolic BP}, \text{Age})}{S^2(\text{Age})} = \frac{S_{XY}}{S^2(X)} = 1.1380$$

- Estimation of the intercept α:

$$a = \overline{Y} - b\,\overline{X} = M_{\text{Systolic BP}} - b\,M_{\text{Age}} = 80.778$$

EXAMPLE

- Covariance between systolic blood pressure and age:

$$\text{Cov}(\text{Systolic BP}, \text{Age}) = \text{Cov}(X, Y) = S_{XY} = 160.4242$$

- Variance of age: $S^2(\text{Age}) = S^2(X) = 140.9697$
- Estimation of the slope β:

$$b = \frac{\text{Cov}(\text{Systolic BP}, \text{Age})}{S^2(\text{Age})} = \frac{S_{XY}}{S^2(X)} = 1.1380$$

- Estimation of the intercept α:

$$a = \overline{Y} - b\overline{X} = M_{\text{Systolic BP}} - b\,M_{\text{Age}} = 80.778$$

Example

- Covariance between systolic blood pressure and age:

$$\text{Cov}(\text{Systolic BP}, \text{Age}) = \text{Cov}(X, Y) = S_{XY} = 160.4242$$

- Variance of age: $S^2(\text{Age}) = S^2(X) = 140.9697$
- Estimation of the slope β:

$$b = \frac{\text{Cov}(\text{Systolic BP}, \text{Age})}{S^2(\text{Age})} = \frac{S_{XY}}{S^2(X)} = 1.1380$$

- Estimation of the intercept α:

$$a = \overline{Y} - b\,\overline{X} = M_{\text{Systolic BP}} - b\,M_{\text{Age}} = 80.778$$

Estimated Regression Equation

The estimated regression line is:

$$\text{Systolic Blood Pressure} = 80.778 + 1.138 \times \text{Age} + \varepsilon$$

Remark

Once the parameters a and b have been estimated, we can compute:

$$\hat{y}_i = a + bx_i$$

the **fitted values**, and then obtain the **residuals**:

$$\varepsilon_i = y_i - \hat{y}_i.$$

ESTIMATED REGRESSION EQUATION

The estimated regression line is:

$$\text{Systolic Blood Pressure} = 80.778 + 1.138 \times \text{Age} + \varepsilon$$

REMARK

Once the parameters a and b have been estimated, we can compute:

$$\hat{y}_i = a + bx_i$$

the **fitted values**, and then obtain the **residuals**:

$$\varepsilon_i = y_i - \hat{y}_i.$$

Coefficient of Determination

- **Percentage of variance explained:**

$$R^2 = \frac{\text{Variance explained by the regression}}{\text{Total variance}}$$

  - R^2 represents the proportion of the total variability in Y that is **explained by** its linear relationship with X (i.e., by the model).
  - When $R^2 \approx 1$, the model fits the data very well-knowing the values of X allows for an accurate prediction of Y.
  - When $R^2 \approx 0$, X provides little or no useful information about Y-knowledge of X does not help predict Y.

Decomposition of Variance and Formula for R^2

- In linear regression, the total variation of Y can be decomposed as:

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{Y})^2}_{\text{Total Sum of Squares (SST)}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2}_{\text{Explained Sum of Squares (SSR)}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{Residual Sum of Squares (SSE)}}$$

- The coefficient of determination is then defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = r^2.$$

- **Interpretation:**
  - R^2 measures the proportion of the total variability in Y that is explained by the regression model.
  - A higher R^2 indicates a better fit.

Decomposition of Variance and Formula for R^2

- In linear regression, the total variation of Y can be decomposed as:

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{Y})^2}_{\text{Total Sum of Squares (SST)}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\overline{Y})^2}_{\text{Explained Sum of Squares (SSR)}} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{Residual Sum of Squares (SSE)}}$$

- The coefficient of determination is then defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = r^2.$$

- **Interpretation:**
  - R^2 measures the proportion of the total variability in Y that is explained by the regression model.
  - A higher R^2 indicates a better fit.

Decomposition of Variance and Formula for R^2

- In linear regression, the total variation of Y can be decomposed as:

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{Y})^2}_{\text{Total Sum of Squares (SST)}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2}_{\text{Explained Sum of Squares (SSR)}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{Residual Sum of Squares (SSE)}}$$

- The coefficient of determination is then defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = r^2.$$

- **Interpretation:**
    - R^2 measures the proportion of the total variability in Y that is explained by the regression model.
    - A higher R^2 indicates a better fit.

Example 1: Correlation and Determination Coefficients

- Estimated correlation coefficient between X and Y:

$$\begin{aligned} r &= \text{cor}(\text{Systolic Blood Pressure, Age}) \\ &= \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\,\sqrt{\text{Var}(Y)}} = 0.8961 \end{aligned}$$

- Estimation of R^2:

$$R^2 = r^2 = 0.8031$$

- This indicates a **very strong relationship** between age and systolic blood pressure: about **80% of the variance** in systolic blood pressure is explained by age.

Example 1: Correlation and Determination Coefficients

- Estimated correlation coefficient between X and Y:

$$\begin{aligned} r &= \text{cor}(\text{Systolic Blood Pressure, Age}) \\ &= \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\,\sqrt{\text{Var}(Y)}} = 0.8961 \end{aligned}$$

- Estimation of R^2:

$$R^2 = r^2 = 0.8031$$

- This indicates a **very strong relationship** between age and systolic blood pressure: about **80% of the variance** in systolic blood pressure is explained by age.

REMARKS: SPURIOUS CORRELATIONS

- What is a correlation? It is a positive or negative association between two phenomena-but it is not absolute.
  - **Example:** There is a positive correlation between height and weight among men: those who are 1.80 m tall generally weigh more than those who are 1.60 m. However, there are also short, heavy individuals and tall, thin ones.

Remarks: Spurious Correlations

- In many cases, a correlation reflects a cause-and-effect relationship. Usually, we know which variable is the cause and which is the effect:
  - For instance, smoking causes lung cancer-not the other way around. But in other cases, the direction of causality is much less clear, and sometimes both variables may influence each other.

REMARKS: SPURIOUS CORRELATIONS

- However, many statistical correlations do **not** result from any causal relationship and can therefore be misleading.
  - This often occurs when two time series evolve in parallel due to general economic or scientific progress. For example, if life expectancy increases while movie theater attendance decreases (a negative correlation), no one would seriously claim that people live longer because they go to the cinema less often.
  - Yet in many cases-especially when trying to support a point-people may interpret a mere parallel evolution between two variables as a cause-and-effect relationship.

Plan de cours

1. Statistical Series with Two Variables
   - Introduction
   - Presentation of a Bivariate Series
   - Graphical Representation
   - Mean, Covariance, and Correlation Coefficient
   - Different Types of Scatter Plots
2. Linear Regression
   - Linear Regression
   - Computation of the Regression Line Parameters
   - Nonlinear Regression
3. Exercise

Nonlinear Model

- Let us return to Example 5, which examines the evolution of the number of bacteria over time (in days).

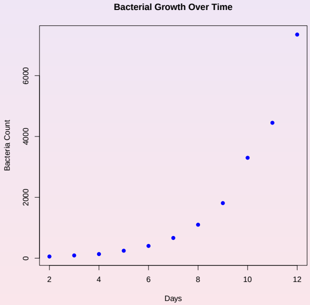


Figure: Relationship between temperature and metabolized energy

NONLINEAR MODEL

- From the graph, we observe that the number of bacteria (Example 5) increases very rapidly-following an **exponential growth pattern.**
- Linear correlation coefficient $r(t, N) = 0.8647$.
- However, note that this relationship is **nonlinear**. A high correlation does not necessarily imply that a linear model is appropriate.

Nonlinear Model

- From the graph, we observe that the number of bacteria (Example 5) increases very rapidly-following an **exponential growth pattern.**
- Linear correlation coefficient $r(t, N) = 0.8647$.
- However, note that this relationship is **nonlinear**. A high correlation does not necessarily imply that a linear model is appropriate.

Nonlinear Model

- From the graph, we observe that the number of bacteria (Example 5) increases very rapidly-following an **exponential growth pattern.**
- Linear correlation coefficient $r(t, N) = 0.8647$.
- However, note that this relationship is **nonlinear.** A high correlation does not necessarily imply that a linear model is appropriate.

Logarithmic Transformation

- To explain N as a function of t, we apply a **logarithmic transformation** only to the variable N (since it takes very large values).
- We set $y = \ln(N)$ and $x = t$.

Logarithmic Transformation

- To explain N as a function of t, we apply a **logarithmic transformation** only to the variable N (since it takes very large values).
- We set $y = \ln(N)$ and $x = t$.

Logarithmic Transformation

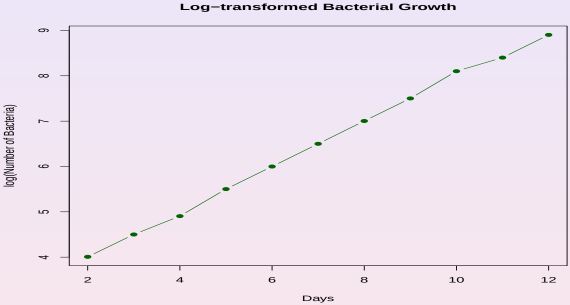


Figure: Relationship between temperature and log-metabolized energy

Logarithmic Transformation

- The resulting graph shows a clear **linear relationship** between y and x.
- This transformation allows us to use a **linear regression model** to estimate the parameters of the exponential growth.

Logarithmic Transformation

- The resulting graph shows a clear **linear relationship** between y and x.
- This transformation allows us to use a **linear regression model** to estimate the parameters of the exponential growth.

ESTIMATION OF THE MODEL PARAMETERS

- The linear correlation coefficient is:

  $r = cor(t, \log(N)) = 0.9996615.$

- The linear fit of $Y = \ln(N)$ as a function of X is well justified.

Estimation of the Model Parameters

- The linear correlation coefficient is:

  $r = cor(t, \log(N)) = 0.9996615.$

- The linear fit of $Y = \ln(N)$ as a function of X is well justified.

Estimation of the Model Parameters

- We find $a = 3.014$ and $b = 0.494$.
- The least-squares regression line is given by

$$Y = 0.494t + 3.014$$

- The sum of squared residuals $SSR = 0.04499$ is very small.
- The coefficient $R^2 = 0.9993$ is very close to 1, therefore we can conclude that the fit is of very high quality.
- In summary, we deduce that the evolution of the number of bacteria as a function of time (days) follows the equation:

$$N(t) = e^{0.494t+3.014} = 20.36871e^{0.494t}.$$

Estimation of the Model Parameters

- We find $a = 3.014$ and $b = 0.494$.
- The least-squares regression line is given by

$$Y = 0.494X + 3.014$$

- The sum of squared residuals $SSR = 0.04499$ is very small.
- The coefficient $R^2 = 0.9993$ is very close to 1, therefore we can conclude that the fit is of very high quality.
- In summary, we deduce that the evolution of the number of bacteria as a function of time (days) follows the equation:

$$N(t) = e^{0.494t+3.014} = 20.36871e^{0.494t}.$$

Estimation of the Model Parameters

- We find $a = 3.014$ and $b = 0.494$.
- The least-squares regression line is given by

$$Y = 0.494X + 3.014$$

- The sum of squared residuals $SSR = 0.04499$ is very small.
- The coefficient $R^2 = 0.9993$ is very close to 1, therefore we can conclude that the fit is of very high quality.
- In summary, we deduce that the evolution of the number of bacteria as a function of time (days) follows the equation:

$$N(t) = e^{0.494t+3.014} = 20.36871e^{0.494t}.$$

Estimation of the Model Parameters

- We find $a = 3.014$ and $b = 0.494$.
- The least-squares regression line is given by

$$Y = 0.494X + 3.014$$

- The sum of squared residuals $SSR = 0.04499$ is very small.
- The coefficient $R^2 = 0.9993$ is very close to 1, therefore we can conclude that the fit is of very high quality.
- In summary, we deduce that the evolution of the number of bacteria as a function of time (days) follows the equation:

$$N(t) = e^{0.494t+3.014} = 20.36871e^{0.494t}.$$

Graph

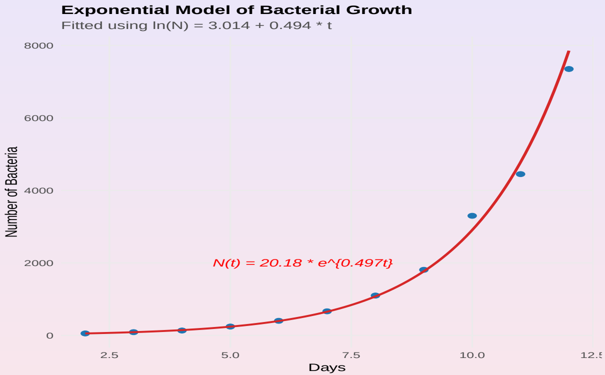


Figure: Relationship between temperature and metabolized energy

EXERCISE 1

One of the key measurements taken when investigating respiratory conditions is the **Forced Expiratory Volume in one second (FEV).** In a random sample of 8 healthy individuals aged between 30 and 35 years, both their height T (in meters) and their FEV (V in liters per second) were measured, yielding the following results:

Subject	1	2	3	4	5	6	7	8
*Height*T	1.85	1.72	1.51	1.62	1.60	1.80	1.75	1.68
V	4.5	3.6	2.7	3.1	3.6	4.4	4.3	3.8

Exercise 1

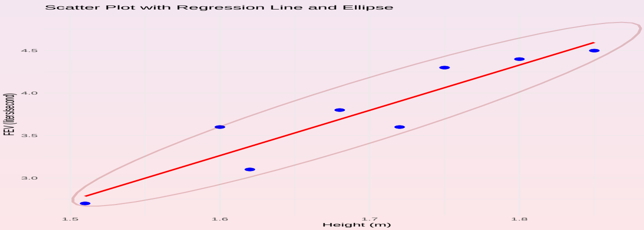
1. Plot and comment on the scatter plot of these observations (with T on the x-axis and V on the y-axis).
2. Compute the linear correlation coefficient between T and V.
3. On the same plot, draw the regression line of V on T.
4. A ninth subject has a height of 1.70 m. What FEV$_1$ value would you predict for this individual? If their actual FEV$_1$ is 4 liters, what is the prediction error?

Note: $\sum t_i = 13.53$, $\sum t_i^2 = 22.9703$, $\sum v_i = 30$, $\sum v_i^2 = 115.36$, $\sum t_i v_i = 51.205$.

Solution

1. **Scatter Plot:** The data points are almost perfectly aligned, suggesting a **linear relationship** between height T and V:

$$V = b \times T + a.$$



Solution

2. With $n = 8$, the correlation coefficient is given by:

$$\text{cor}(T, V) = \frac{S_{T,V}}{S_T \times S_V}.$$

- $\overline{T} = \frac{1}{n}\sum t_i = 1.6913$, $\quad S^2(T) = \frac{1}{n}\sum t_i^2 - (\overline{T})^2 = 0.0125$, $S_T = 0.112$.
- $\overline{V} = \frac{1}{n}\sum v_i = 3.75$, $\quad S^2(V) = \frac{1}{n}\sum v_i^2 - (\overline{V})^2 = 0.409$, $S_V = 0.639$.
- $\text{Cov}(T, V) = S_{T,V} = \frac{1}{n}\sum t_i v_i - \overline{T}\,\overline{V} = 0.0668$.

Thus:

$$\text{cor}(T, V) = \frac{S_{T,V}}{S_T \times S_V} = 0.9335 \approx 0.93.$$

SOLUTION

3. The regression line of V on T is given by:

$$V = a + bT,$$

where:

$$b = \frac{S_{T,V}}{S_T^2}, \quad a = \overline{V} - b\,\overline{T}.$$

We find: $b = 5.33$, $\quad a = -5.267$.

Hence, the regression equation is:

$$V = 5.33T - 5.267.$$

4. Predicted V for T = 1.70: $V = 5.33 \times 1.70 - 5.267 = 3.794.$
5. Prediction error: error $= V_{obs} - V_{pred} = 4 - 3.794 = 0.206$.