Statistical series with two characters

LAIDI.M

National Higher School of Autonomous Systems

2025/2026

Introduction

- In the previous chapter, we studied the distribution of a single statistical variable and used numerical and graphical tools to describe it.
- Often, we are interested in the relationship between two (or more) variables.
- Specifically, we want to know:
  - Whether the value of one variable affects the other (i.e., if there is a correlation).
  - How to fit one variable with respect to another using a mathematical equation to make predictions.

Distributions and Characteristics

- To study the joint distribution and marginal distributions of two statistical variables, we consider two discrete variables (just for simplify our study).
- If one or both variables are continuous, we categorize values into class intervals.
- Let X and Y be two discrete variables for a population of n individuals.
- Denote:
  - Values of X: $x_1 < x_2 < \cdots < x_k$
  - Values of Y: $y_1 < y_2 < \cdots < y_l$

Absolute and Relative Frequencies

- For each pair (x_i, y_j), we define:
  - **Absolute frequency** n_{ij}: Number of individuals where $X = x_i$ and $Y = y_j$.
  - **Relative frequency** f_{ij}: Given by $f_{ij} = \frac{n_{ij}}{n}$.
- The sum of all absolute frequencies is n:

$$\sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} = n$$

- The sum of all relative frequencies is 1:

$$\sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} = 1$$

Contingency Table

- The statistical series for (X, Y) can be represented using a **contingency table**, showing the absolute or relative frequencies of each value pair.

X \ Y	y_1	y_2	$\dots$	y_j	$\dots$	y_l	**Total**
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1l}	$\sum_j n_{1j}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2l}	$\sum_j n_{2j}$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{il}	$\sum_j n_{ij}$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kl}	$\sum_j n_{kj}$
Total	$\sum_i n_{i1}$	$\sum_i n_{i2}$	$\dots$	$\sum_i n_{ij}$	$\dots$	$\sum_i n_{il}$	n

Contingency Table: Example

Consider the following data set:

$$(3,6),(2,6),(2,4),(1,4),(3,8),(3,2),(3,2),(3,4),(1,8),$$
$$(2,8),(3,10),(3,8),(3,4),(2,2),(3,8),(1,10),(1,8),(2,2),$$
$$(3,8),(3,4),(1,6),(3,8),(1,2),(2,2),(3,2),(3,6),(2,2),$$
$$(3,2),(2,6),(2,10),(2,6),(1,2),(3,2),(2,8),(3,2),(1,8),$$
$$(2,4),(3,10)$$

$(1,2),(1,2),(1,4),(1,6),(1,8),(1,8),(1,8),(1,10),$
$(2,2),(2,2),(2,2),(2,2),(2,4),(2,4),(2,6),(2,6),$
$(2,6),(2,8),(2,8),(2,10),$
$(3,2),(3,2),(3,2),(3,2),(3,2),(3,2),$
$(3,4),(3,4),(3,4),(3,6),(3,6),(3,8),(3,8),(3,8),(3,8),(3,8),$
$(3,10),(3,10)$

X \ Y	2	4	6	8	10	**Total**
1	2	1	1	3	1	
2	4	2	3	2	1	
3	6	3	2	5	2	
Total						

Marginal Distributions

Marginal distributions refer to the statistical distributions of each variable (X or Y) individually.

Definition of Marginal Absolute Frequencies of X

Definition The i^{th} marginal absolute frequency of X is the count of individuals for which $X = x_i$, regardless of the value of Y. This is given by:

$$n_{i\bullet} = \sum_{j=1}^{J} n_{ij}$$

Definition of Marginal Relative Frequencies of X

Definition : The i^{th} marginal relative frequency of X is the proportion of individuals for which $X = x_i$, regardless of the value of Y. This is given by:

$$f_{i\bullet} = \sum_{j=1}^{J} f_{ij} = \frac{n_{i\bullet}}{n}$$

Marginal Frequencies of Y

Similarly, we can define the marginal absolute and relative frequencies of Y, given by:

$$n_{\bullet j} = \sum_{i=1}^{k} n_{ij} \quad \text{and} \quad f_{\bullet j} = \sum_{i=1}^{k} f_{ij} = \frac{n_{\bullet j}}{n}$$

Properties of Marginal Frequencies

The sum of the marginal absolute frequencies for X and Y equals the total population n:

$$\sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j} = n$$

Similarly, the sum of the marginal relative frequencies for X and Y is equal to 1:

$$\sum_{i=1}^{k} f_{i\bullet} = \sum_{j=1}^{l} f_{\bullet j} = 1$$

X \ Y	2	4	6	8	10	**Total** $n_{i\bullet}$
1	2	1	1	3	1	8
2	4	2	3	2	1	12
3	6	3	2	5	2	18
Total $n_{\bullet j}$	12	6	6	10	4	**38**

Marginal characteristics

Marginal Mean of X

The marginal mean of X, denoted by $\bar{X}$, is defined as:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} n_{i\bullet} x_i$$

where $n_{i\bullet}$ represents the marginal absolute frequency of X for each x_i.

Marginal Mean of Y

Similarly, the marginal mean of Y, denoted by $\bar{Y}$, is defined as:

$$\bar{Y} = \frac{1}{n} \sum_{j=1}^{l} n_{\bullet j} y_j$$

where $n_{\bullet j}$ represents the marginal absolute frequency of Y for each y_j.

Marginal Variance of X

The marginal variance of X, $\mathrm{Var}(X)$, is defined by:

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\left(x_i - \bar{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i^2 - (\bar{X})^2$$

This measures the spread of X around its marginal mean $\bar{X}$.

Marginal Variance of Y

The marginal variance of Y, $\mathrm{Var}(Y)$, is defined by:

$$\mathrm{Var}(Y) = \frac{1}{n}\sum_{j=1}^{J} n_{\bullet j}\left(y_j - \bar{Y}\right)^2 = \frac{1}{n}\sum_{j=1}^{J} n_{\bullet j} y_j^2 - (\bar{Y})^2$$

This measures the spread of Y around its marginal mean $\bar{Y}$.

Conditional distribution

Conditional Distribution of X

Definition : The conditional distribution of X is the distribution of X that corresponds to a fixed value y_j of Y.

X \ Y	y_1	y_2	$\dots$	y_j	$\dots$	y_l	**Total**
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1l}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2l}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{il}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kl}	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet l}$	n

Conditional Relative Frequency of X Given $Y = y_j$

The i-th conditional relative frequency of X given $Y = y_j$ is the proportion of individuals for which $X = x_i$ in the sub-population where $Y = y_j$. It is given by:

$$f_{i/Y=y_j} = \frac{n_{ij}}{n_{\bullet j}}$$

with the property that:

$$\sum_{i=1}^{k} f_{i/Y=y_j} = 1$$

Conditional Distribution of Y Given $X = x_i$

Similarly, we define the conditional distribution of Y given $X = x_i$. The conditional relative frequency of Y given $X = x_i$ is:

$$f_{j/X=x_i} = \frac{n_{ij}}{n_{i\bullet}}$$

with the property that:

$$\sum_{j=1}^{I} f_{j/X=x_i} = 1$$

Conditional Mean of X Given $Y = y_j$

The conditional mean of X given $Y = y_j$ is defined by:

$$\bar{X}_{/Y=y_j} = \frac{1}{n_{\bullet j}} \sum_{i=1}^{k} n_{ij} x_i$$

Conditional Mean of Y Given $X = x_i$

The conditional mean of Y given $X = x_i$ is defined similarly:

$$\bar{Y}_{/X=x_i} = \frac{1}{n_{i\bullet}} \sum_{j=1}^{l} n_{ij} y_j$$

Conditional Variance of X Given $Y = y_j$

The conditional variance of X given $Y = y_j$ is defined by:

$$\mathrm{Var}\left(X/Y = y_j\right) = \frac{1}{n_{\bullet j}} \sum_{i=1}^{k} n_{ij} \left(x_i - \bar{X}_{/Y=y_j}\right)^2 = \frac{1}{n_{\bullet j}} \sum_{i=1}^{k} n_{ij} x_i^2 - \left(\bar{X}_{/Y=y_j}\right)^2$$

Conditional Standard Deviation of X Given $Y = y_j$

The conditional standard deviation of X given $Y = y_j$ is defined by:

$$\sigma\left(X/Y = y_j\right) = \sqrt{\text{Var}\left(X/Y = y_j\right)}$$

Covariance of two characters

Definition : The covariance of the two statistical variables X and Y is defined by

$$\text{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}\left(x_i - \bar{X}\right)\left(y_j - \bar{Y}\right) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_i y_j - \bar{X}\,\bar{Y}$$

Properties of the covariance

Proposition : Let X and Y be two discrete statistical variables taking respectively the values $x_1 < x_2 < \cdots < x_k$ and $y_1 < y_2 < \cdots < y_l$ and let $a, b, c, d \in \mathbb{R}$ be some constants. We have

1. $\text{cov}(X, Y) = \text{cov}(Y, X)$.
2. $\text{cov}(X, X) = \text{Var}(X)$.
3. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{cov}(X, Y)$.
4. $\text{cov}(aX + b, cY + d) = ac\,\text{cov}(X, Y)$.
5. $|\text{cov}(X, Y)| \leq \sigma_X \sigma_Y$.

Correlation Coefficient : Definition and Interpretation

- The correlation coefficient, denoted by $\rho_{X,Y}$ (population) or r (sample), measures the **strength and direction** of the linear relationship between two variables X and Y.
- Defined by:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

  where $\text{cov}(X, Y)$ is the covariance, and σ_X, σ_Y are standard deviations.
- Interpretation of $\rho_{X,Y}$:
  - $\rho_{X,Y} > 0$: Positive correlation, as X increases, Y tends to increase.
  - $\rho_{X,Y} < 0$: Negative correlation, as X increases, Y tends to decrease.
  - $\rho_{X,Y} \approx 0$: No linear relationship between X and Y.

Strength of Correlation

- The absolute value $|\rho_{X,Y}|$ indicates the **strength** of the linear relationship:
  - Perfect correlation: $\rho_{X,Y} = 1$ (perfect positive) or $\rho_{X,Y} = -1$ (perfect negative).
  - Strong correlation: $|\rho_{X,Y}| \approx 1$.
  - Weak correlation: $|\rho_{X,Y}| \approx 0$.
- **Typical thresholds** for interpretation:
  - $0.7 \leq |\rho_{X,Y}| \leq 1$: Strong correlation
  - $0.3 \leq |\rho_{X,Y}| < 0.7$: Moderate correlation
  - $0 \leq |\rho_{X,Y}| < 0.3$: Weak correlation

Strength of Correlation

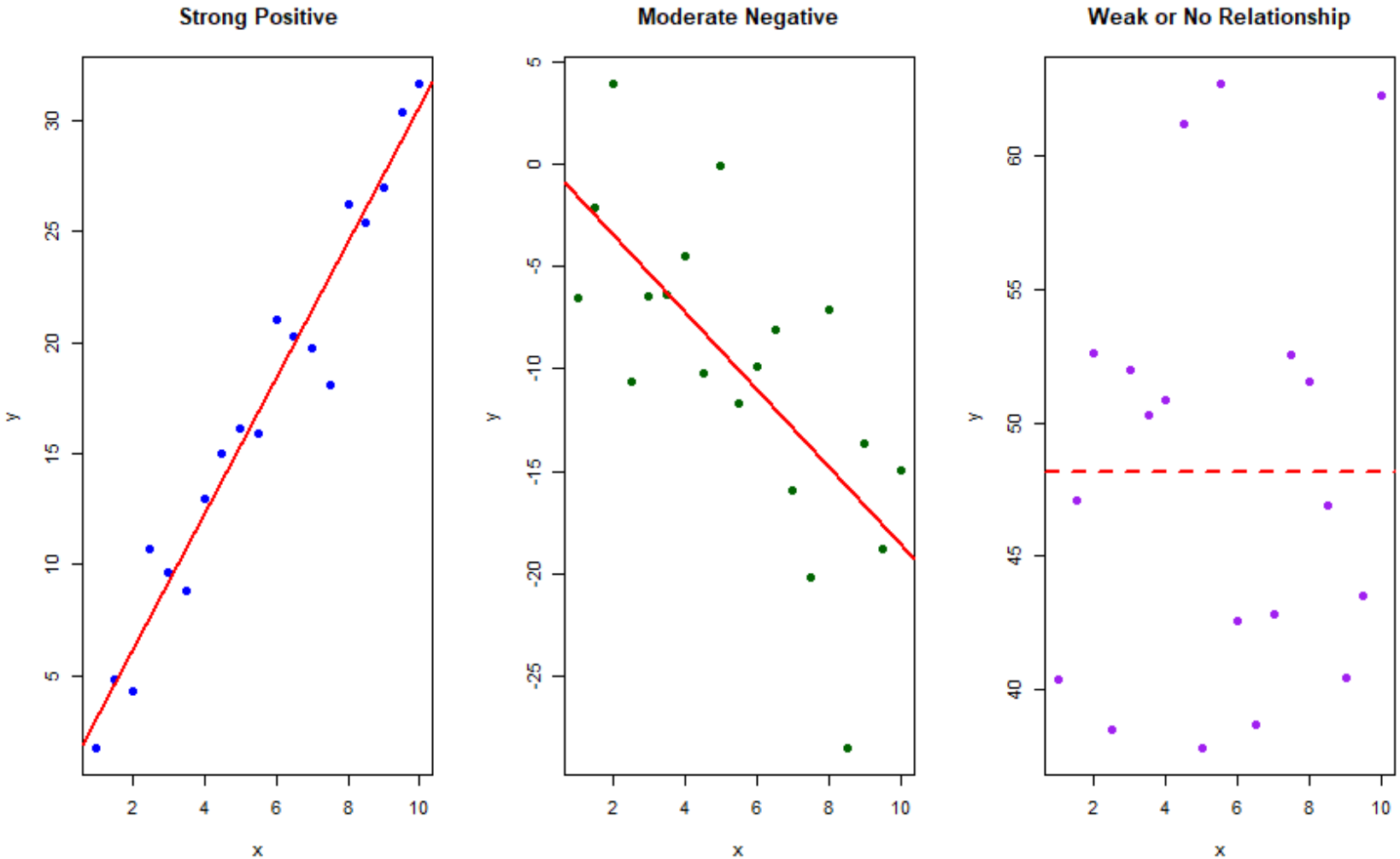

Figure:

Fittings

1) Linear fitting

The line of best fit (or the regression line) of Y on X using the **least square method** is given by $Y = aX + b$, where

$$\begin{cases} a = \frac{\text{cov}(X,Y)}{\text{Var}(X)} \\ b = \bar{Y} - a\bar{X} \end{cases}$$

X is called explanatory or independant variable and Y is called response or dependant variable.

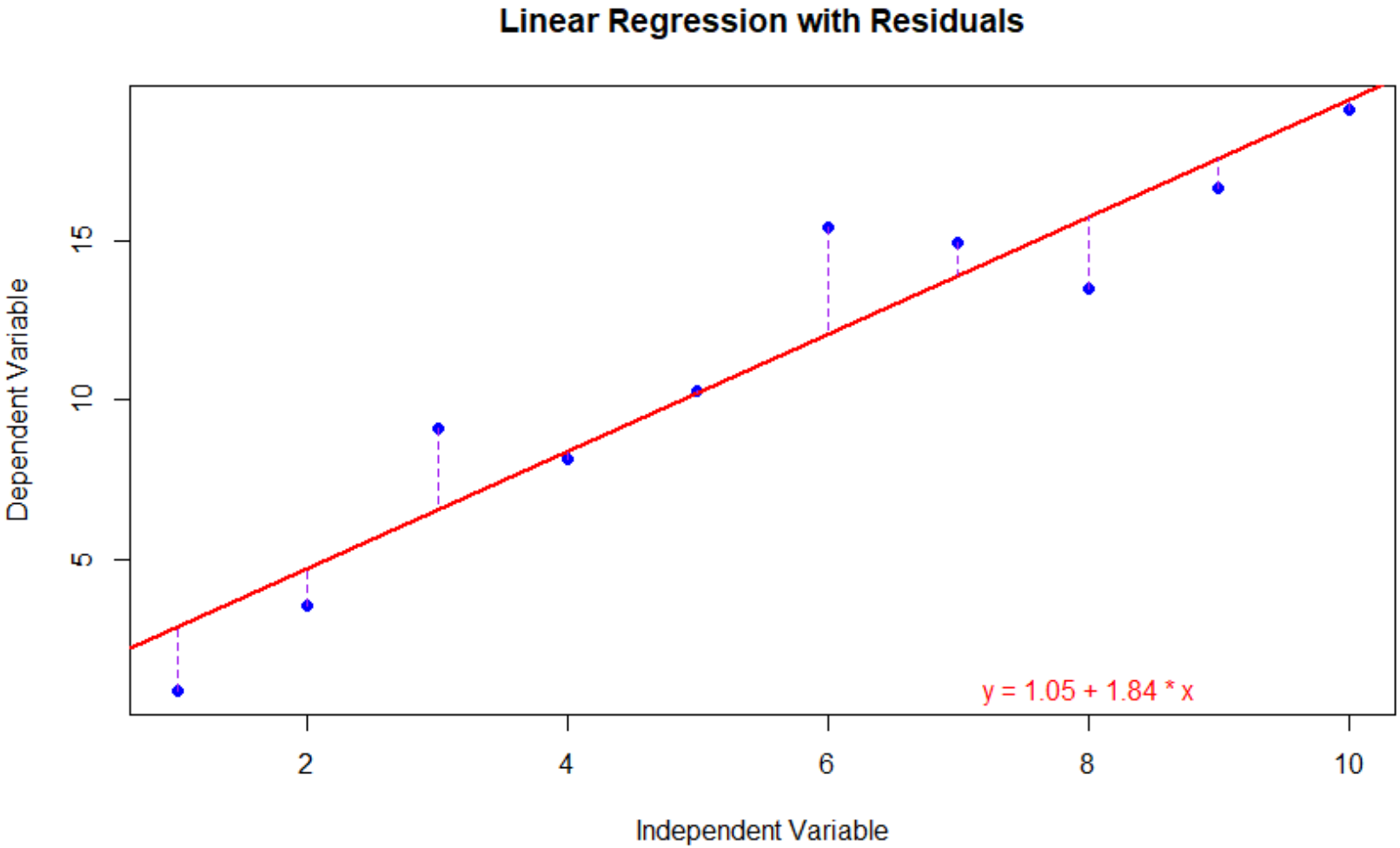

Figure: Linear Regression with Residuals.

Remarks :

- The regression line of X on Y is given by $X = cY + d$, where

$$\begin{cases} c = \dfrac{\text{cov}(X, Y)}{\text{Var}(Y)} \\ d = \bar{X} - c\bar{Y} \end{cases}$$

- The two lines pass through the point $(\bar{X}, \bar{Y})$.

Example : Consider the example below where the mass, y (grams), of a chemical is related to the time, x (seconds), for which the chemical reaction has been taking place according to the table:

Time, x (seconds)	5	7	12	16	20
Mass, y (grams)	40	120	180	210	240

Find the equation of the regression line.

Solution: Calculating the Regression Line

To find the equation of the regression line, $y = ax + b$, we need to calculate:

- The slope $a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- The intercept $b = \bar{y} - a\bar{x}$

where:

$$\bar{x} = \frac{\sum_{i=1}^{n} x}{n}, \quad \bar{y} = \frac{\sum_{i=1}^{n} y}{n}$$

Solution: Step-by-Step Calculation

1. Calculate the means $\bar{x}$ and $\bar{y}$:

$$\bar{x} = \frac{5 + 7 + 12 + 16 + 20}{5} = 12,$$

$$\bar{y} = \frac{40 + 120 + 180 + 210 + 240}{5} = 158$$

2. Use these to calculate b and a:

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \mathbf{12.21}, \quad b = \bar{y} - a\bar{x} = \mathbf{11.51}$$

Solution: Final Regression Line Equation

Substitute the values for a and b to find the final equation:

$$y = 12.21 \times x + 11.51$$

This equation represents the best-fit line for predicting y (mass in grams) based on x (time in seconds).

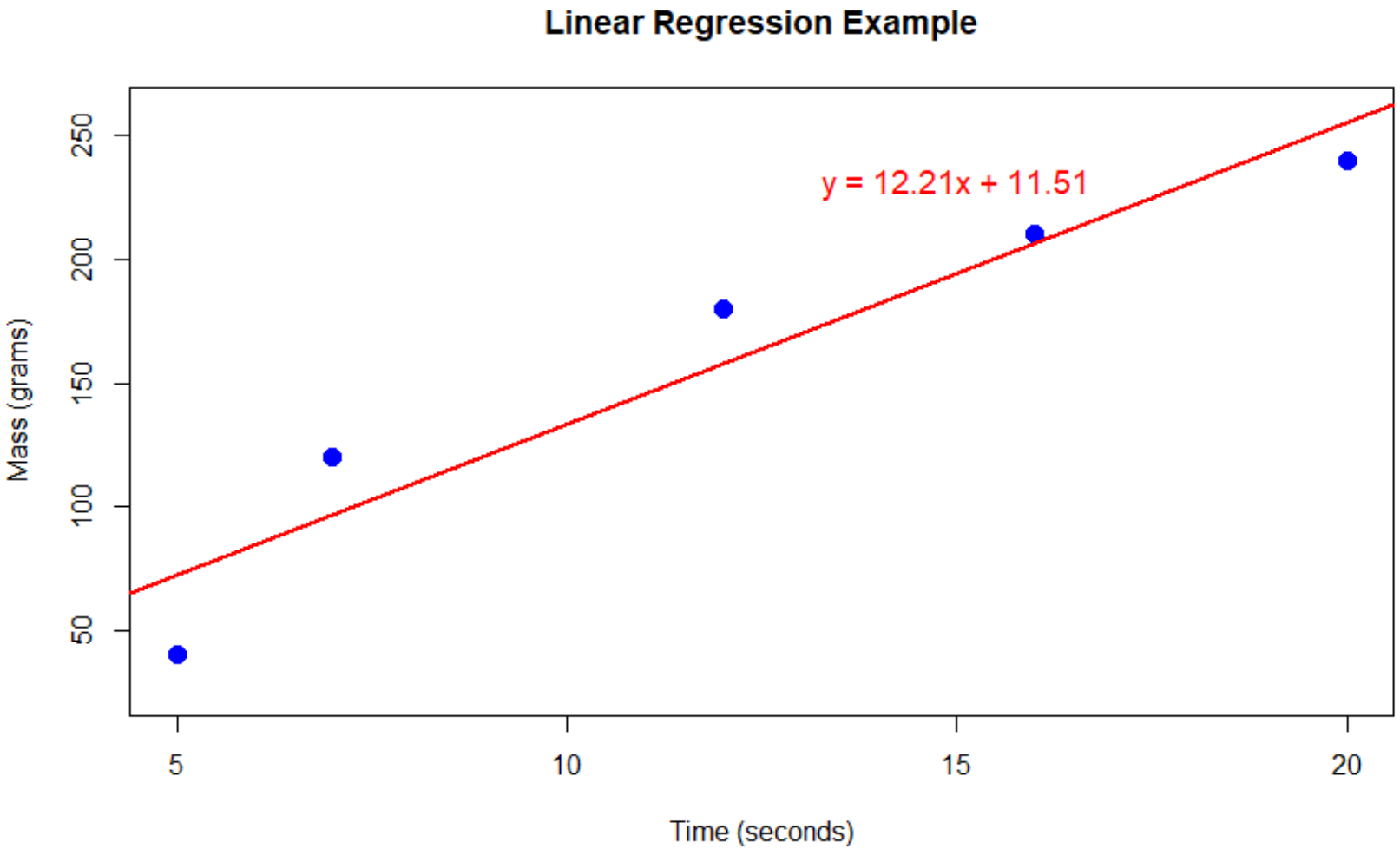

Figure: Regression Line Equation of Mass based on time

Fittings

1) Non-Linear fitting

- The regression line is not always appropriate to describe the relation between two variables X and Y.
- In some cases, the scatter plot suggests other forms, such as an exponential function of the form:

$$Y = B \times A^X (\text{or} \quad B \times X^A)$$

Example

An epidemic has broken out in a city and the number of cases reported in each day is given in the following table.

The day (X)	1	2	3	4	5	6
Le number of cases (Y)	4	13	38	106	330	965

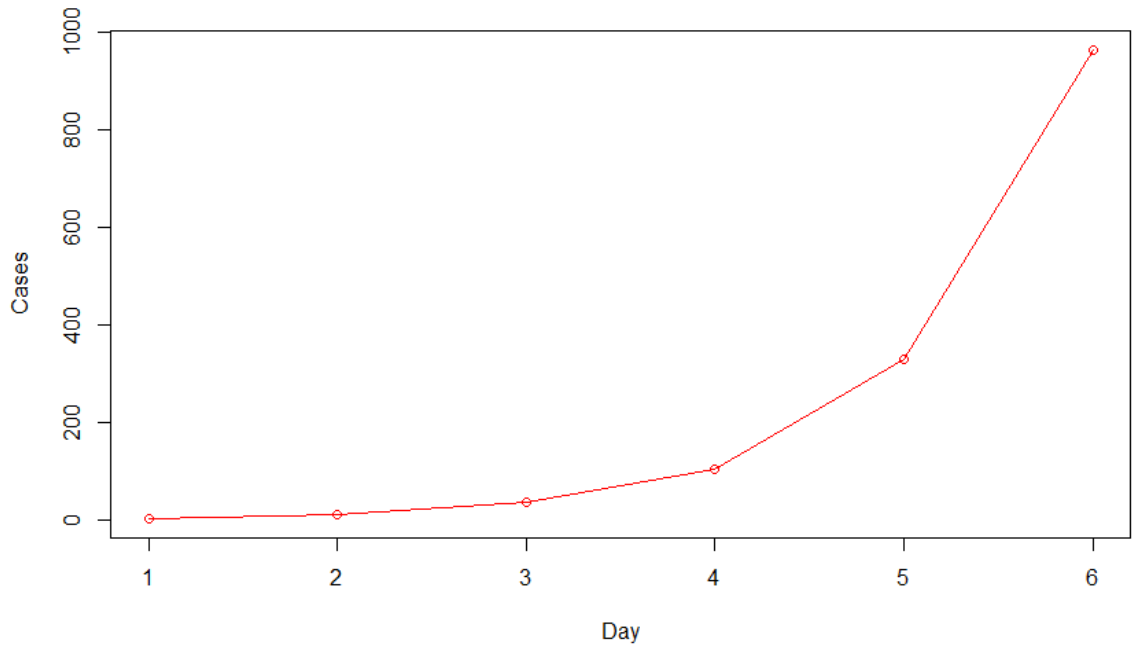

Figure: Number of Cases Over Days

Linearization of the Equation

To determine A and $B(A > 0, B > 0)$, we linearize the equation:

$$Y = B \times A^X \quad \Longrightarrow \quad \ln(Y) = \ln(B) + X \ln(A)$$

Substituting:

- Let $Z = \ln(Y)$, $a = \ln(A)$, $b = \ln(B)$,
- The equation becomes:

$$Z = b + aX$$

Coefficients Calculation

The coefficients a and b can be calculated using:

$$a = \frac{\text{cov}(X, Z)}{\text{Var}(X)}$$
$$b = \bar{Z} - a\bar{X}$$

Once a and b are known:

$$A = e^a$$
$$B = e^b$$

The day (X)	1	2	3	4	5	6
Le number of cases (Y)	4	13	38	106	330	965
$Z = \ln(Y)$	1.39	2.56	3.64	4.66	5.80	6.87

Coefficients Calculation for the example

The coefficients a and b can be calculated using:

$$a = \frac{\text{cov}(X, Z)}{\text{Var}(X)} = \frac{3.179788}{2.916667} = 1.090213$$

$$b = \bar{Z} - a\bar{X} = 0.3381698$$

So

$$A = e^a = 2.974907$$

$$B = e^b = 1.402379$$

Finally :

$$\boxed{Y = 1.402379 \times 2.974907^X}$$

Exercise

Determine A and B when

$$Y = B \times X^A$$

Definition of Statistical Independence

Two variables X and Y are said to be independent if and only if:

$$f_{ij} = f_{i\bullet} \times f_{\bullet j}$$

or equivalently:

$$n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}, \quad \forall i = 1, \ldots, k \quad \text{and} \quad \forall j = 1, \ldots, l.$$

Interpretation:

- If X and Y are independent, joint frequencies (n_{ij}) can be calculated from marginal distributions.
- Independence means that observing X or Y separately provides the same information as observing them together.

Example: Testing Independence of X and Y

Survey Data: An investigation was conducted on 100 households to observe:

- X: Monthly expenditures (in 1000s of DA),
- Y: Monthly income (in 1000s of DA).

The results are summarized in the following contingency table:

Y	$[4, 10[$	$[10, 20[$	$[20, 40[$	$n_{i\bullet}$
$[3, 5[$	20	10	0	$n_{1\bullet} = 30$
$[5, 15[$	10	20	10	$n_{2\bullet} = 40$
$[15, 35[$	0	10	20	$n_{3\bullet} = 30$
$n_{\bullet j}$	30	40	30	$n = 100$

Example: Testing Independence (Continued)

Test for Independence: To verify independence, calculate n_{ij} using the formula:

$$n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}.$$

For $i = 2$ and $j = 1$:

$$n_{21} = \frac{n_{2\bullet} \times n_{\bullet 1}}{n} = \frac{40 \times 30}{100} = 12.$$

However, the observed value is:

$$n_{21} = 10.$$

Conclusion: Since $n_{21} \neq \frac{n_{2\bullet} \times n_{\bullet 1}}{n}$, X and Y are not independent.

Exercise

Consider two variables X and Y measured in a dataset.
Show that if X and Y are statistically independent, then their covariance is zero:

$$Cov(X, Y) = 0.$$