1

Linear Modeling for Two-Dimensional Data

CONCEPTS COVERED IN THIS CHAPTER.

This brief chapter serves as a reminder of the concepts presented in detail in Volume 1. It primarily provides an overview of basic statistical analysis tools, particularly linear regression and correlation for two-dimensional data.

References: [SAP 11].

1.1. Basic statistics

Consider a population of n elements. Each element i is characterized by the value of a variable $x = x_i$. The n values x_i constitute a one-dimensional statistical series, whose characteristics are:

– The *average* $\bar{x}$ is defined by:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In this definition, it is assumed that all elements have the same statistical weight ($p_i = \frac{1}{n}$). If the weights are not equal, the following expression is used:

Mathematics for Digital Science 3,
by Gérard-Michel COCHARD and Mhand HIFI. © ISTE Ltd 2025.

$$\bar{\bar{x}} = \sum_{i=1}^{n} p_i x_i \text{ with } \sum_{i=1}^{n} p_i = 1$$

where p_i represents the statistical weight of individual i.

– *Variance* $v(x)$ is defined as the average of the squares of the deviations from the average:

$$v(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{or} \quad v(x) = \sum_{i=1}^{n} p_i\,(x_i - \bar{x})^2$$

Huygens' theorem provides another method for calculating variance:

$$v(x) = \overline{x^2} - \bar{x}^2, \text{ where } \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \text{ or } \overline{x^2} = \sum_{i=1}^{n} p_i\, x_i^2$$

This relationship is often summarized as "the average of squares minus the square of the mean".

– *Standard deviation* $\sigma(x)$ is defined as the square root of the variance:

$$\sigma(x) = \sqrt{v(x)}$$

EXAMPLE 1.1.– Consider the statistical series shown in Figure 1.1, which represents the number of rainy days over 10 consecutive years at a given location.

x
50
120
70
90
150
80
100
30
110
40

Figure 1.1. *Statistical series*

The average $\bar{x} = 84$ can be easily calculated by assigning equal statistical weight to each measurement. The average of the squares $\overline{x^2} = 8340$, the variance $v(x) = 1284$ and the standard deviation $\sigma(x) = 35{,}83$ are also determined.

Figure 1.2 shows the graphical representation of the statistical series in the form of a histogram. This histogram illustrates the distribution of data regarding the number of rainy days over the 10 years.

The average is a measure of the position of the statistical series along the number of days axis, while the standard deviation serves as a dispersion parameter, providing an indicator of the spread of the statistical series.

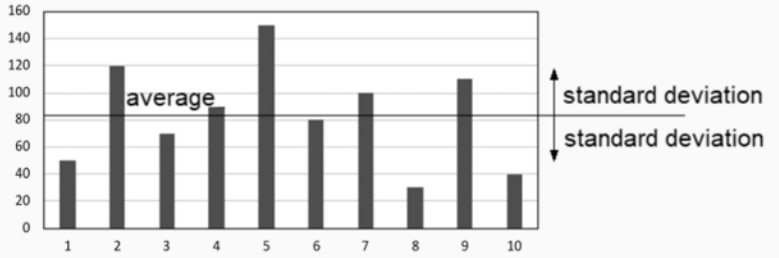

Figure 1.2. *Graphical representation of the statistical series*

1.2. Linear adjustment

Now, consider a two-dimensional statistical series, where each element is characterized by the values of two variables, x and y. For each variable, various statistical measures can be calculated, such as the average, variance and standard deviation.

To graphically represent this two-dimensional series, a two-dimensional Cartesian coordinate system is used. The x-axis represents the variable x, and the y-axis represents the variable y. Each element i of the series is represented as a point (x_i, y_i) in this coordinate system, where the coordinate x_i corresponds to the value of the variable x, and the coordinate y_i corresponds to the value of the variable y. Figure 1.3 shows examples of graphical representations of two two-dimensional series.

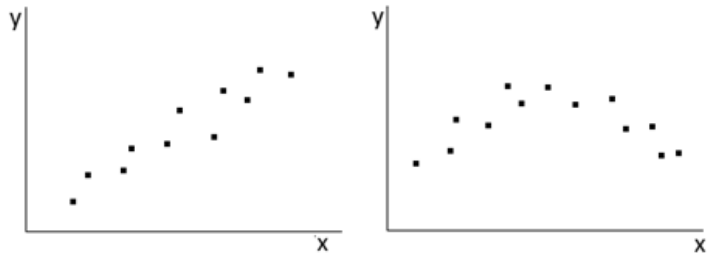

Figure 1.3. *Example of two two-dimensional series*

When observing a two-dimensional series and detecting a certain structure in the set of representative points, we may be inclined to model this structure using a curve. This involves finding a mathematical function that best describes the relationships between the variables x and y. In the examples shown in Figure 1.3, a straight line can be proposed for modeling the first example, and a parabola for the second example, as shown in Figure 1.4. These models are *adjustments* that simplify the representation of trends or relationships observed in the data.

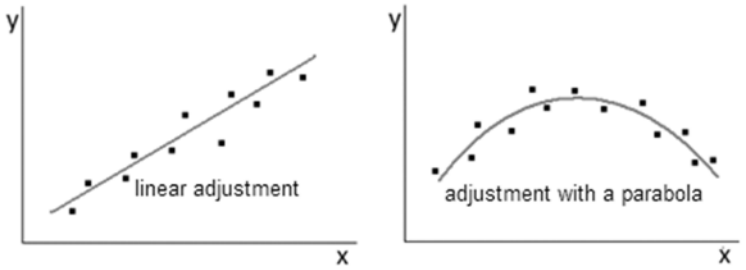

Figure 1.4. *Examples of adjustments*

The linear adjustment is the simplest of all analytical adjustments. It involves obtaining the equation of the straight line that "best fit" the set of representative points of the series.

A classic method for obtaining the equation of the line in linear adjustment is the least squares method. This method involves minimizing the sum of the squares of the deviations between the observed values and the values predicted by the line. For the variables x and y, the respective means, denoted by $\bar{x}$ and $\bar{y}$, are calculated assuming equal statistical weight for each value of i:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Next, the deviations from these averages for each point in the series are calculated (convenient to work with "centered" coordinates):

$$X_i = (x_i - \bar{x}) \text{ and } Y_i = (y_i - \bar{y}).$$

It is easy to verify that:

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i = 0\,.$$

The squares of these deviations are obtained by squaring these values:

$$(x_i - \bar{x})^2 \quad \text{and} \quad (y_i - \bar{y})^2.$$

The least squares method involves finding the coefficients a and b of the equation of the line $y = ax + b$. Alternatively, using the centered coordinates, the equation becomes $Y' = AX + B$, where for each of the representative points, $Y_i' = AX_i + B$. The relationship between (A, B) and (a, b) is:

$$a = A \quad \text{and} \quad b = B - a\bar{x} + \bar{y}$$

The goal is to optimize the sum of squared deviations to a minimum. Mathematically, this involves minimizing the following objective function:

$$F(a,b) = \sum_{i=1}^{n} \left[\left(y_i - (ax_i + b) \right)^2 \right]$$

In other words, the aim is to minimize the following quantity:

$$M = \sum_{i=1}^{n} (Y_i - Y'_i)^2$$
$$= \sum_{i=1}^{n} (Y_i - AX_i - B)^2 = \sum_{i=1}^{n} Y_i^2 - 2A\sum_{i=1}^{n} X_i Y_i + A\sum_{i=1}^{n} X_i^2 + nB^2 = f(A,B)$$

The minimum of M corresponds to the cancellation of the first derivatives with respect to A and B, the only unknowns in M. Taking the partial derivatives:

$$\frac{\partial M}{\partial A} = -2\sum_{i=1}^{n} X_i Y_i + 2A\sum_{i=1}^{n} X_i^2 = 0 \quad \Leftrightarrow \quad \frac{\partial M}{\partial B} = 2nB = 0$$

which leads to:

$$A = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \text{ and } B = 0$$

These conditions lead to the following equations:

$$a = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \text{ and } b = \bar{y} - a\bar{x}$$

EXAMPLE 1.2.– Let us consider the statistical series showing the number of rainy days (x) and umbrella sales in local currency (y) (see Figure 1.5).

x	y
50	65000
120	180000
70	90000
90	140000
150	200000
80	125000
100	130000
30	50000
110	145000
40	65000

Figure 1.5. *Statistical series (x, y)*

x	y	X	Y	X^2	XY
50	65000	-34	-54000	1156	1836000
120	180000	36	61000	1296	2196000
70	90000	-14	-29000	196	406000
90	140000	6	21000	36	126000
150	200000	66	81000	4356	5346000
80	125000	-4	6000	16	-24000
100	130000	16	11000	256	176000
30	50000	-54	-69000	2916	3726000
110	145000	26	26000	676	676000
40	65000	-44	-54000	1936	2376000
averages	84	119000		sums	12840	16840000

Figure 1.6. *Detailed adjustment calculations*

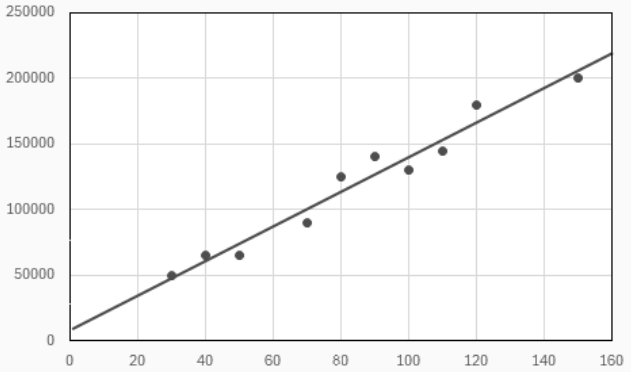

Figure 1.7. *Adjustment line. For a color version of this figure, see www.iste.co.uk/cochard/mathematics3.zip*

Figure 1.6 summarizes the calculations required to determine the best-fit adjustment line, with the values of $a = 1311.53$ and $b = 8831.78$. Figure 1.7 displays the best-fit adjustment line.

1.3. Linear correlation

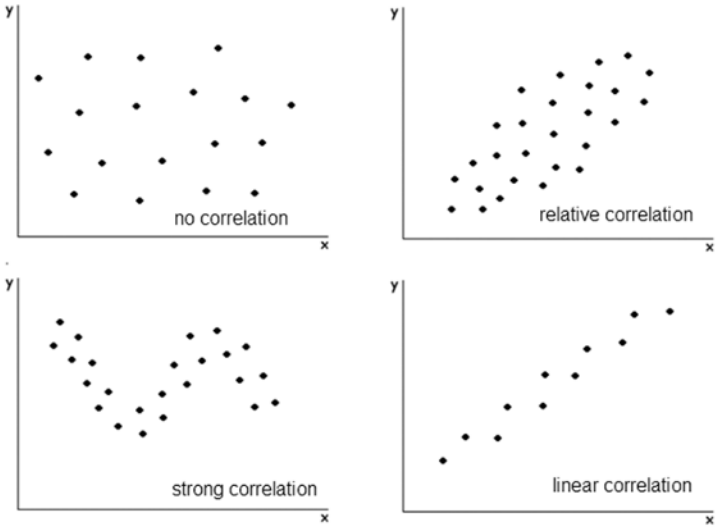

Figure 1.8. *Different correlation situations*

In the case of adjustment, the goal is to express y as a function of x. This choice is arbitrary, as x could be expressed as a function of y. In this case, two adjustment lines would be obtained, both intersecting at the point $(\bar{x}, \bar{y})$:

$$y = ax + b \quad \text{and} \quad x = a'y + b'$$

By treating the variables x and y symmetrically, the concept of *correlation* between these variables can be introduced. Correlation measures the relationship between two variables and quantifies the possible influence of one on the other. Figure 1.8 presents various examples of scatter plots to illustrate different correlation situations.

In particular, in the case of linear correlation, it is interesting to note that when the two best-fit adjustment lines, $y = f(x)$ and $x = f'(y)$, coincide, this indicates maximum linear correlation between the variables x and y.

EXAMPLE 1.3.– For the series in Example 1.2, the following two best-fit adjustment lines are obtained:

$$y = 1311.53x + 8831.78 \text{ and } x = 0.0007\ y - 3.17.$$

Figure 1.9 shows that the two straight lines are very close to each other, indicating a strong correlation between the variables.

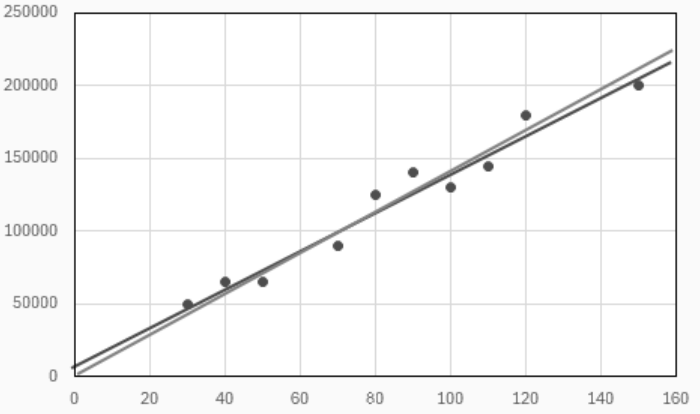

Figure 1.9. *Adjustment lines. For a color version of this figure, see www.iste.co.uk/cochard/mathematics3.zip*

The two best-fit adjustment lines have direction coefficients a and a'. If the lines coincide, then $a = \frac{1}{a'}$ or equivalently, $a \times a' = 1$. Now,

$$a = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}, a' = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} Y_i^2}, \text{so } a \times a' = \frac{(\sum_{i=1}^{n} X_i Y_i)^2}{(\sum_{i=1}^{n} X_i^2)(\sum_{i=1}^{n} Y_i^2)}.$$

The maximum correlation corresponds to the following equality (known as the Cauchy–Schwarz equality):

$$\left(\sum_{i=1}^{n} X_i Y_i\right)^2 = \left(\sum_{i=1}^{n} X_i^2\right)\left(\sum_{i=1}^{n} Y_i^2\right).$$

The analytical definition of the linear correlation is:

$$r = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}}$$

which is simply $\sqrt{a \times a'}$.

EXAMPLE 1.3 (CONTINUED).– Let us return to Example 1.3. The equations of the adjustment lines are:

$$y = 1311.53x + 8831.78 \quad \text{and} \quad x = 0.0007\, y - 3.17$$

The linear correlation coefficient is close to 1, i.e. $r = 0.98 \approx 1$. This indicates an almost maximal linear correlation between the variables x and y. In this case, a strong relationship exists between x and y.

The linear correlation coefficient r is often written in another form, using the standard deviations $\sigma(x)$ and $\sigma(y)$:

$$\sigma(x) = \sqrt{v(x)} = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n}} \quad \text{and} \quad \sigma(y) = \sqrt{v(y)} = \sqrt{\frac{\sum_{i=1}^{n} Y_i^2}{n}}$$

Furthermore, the covariance $\text{cov}(x, y)$ is defined by:

$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n} X_i Y_i}{n}$$

It follows that:

$$r = \frac{\mathrm{cov}(x, y)}{\sigma(x)\sigma(y)}$$

In the case of linear fitting, the expression for M is:

$$M = \sum_{i=1}^{n} Y_i^2 - 2A \sum_{i=1}^{n} X_i Y_i + A \sum_{i=1}^{n} X_i^2 + nB^2$$

The minimum is found by replacing A and B with the values obtained:

$$M_{min} = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} X_i Y_i\right)^2}{\sum_{i=1}^{n} X_i^2} = \frac{\left(\sum_{i=1}^{n} X_i^2\right)\left(\sum_{i=1}^{n} Y_i^2\right) - \left(\sum_{i=1}^{n} X_i Y_i\right)^2}{\sum_{i=1}^{n} X_i^2}$$

By definition, M and therefore M_{min} are positive or zero quantities. This leads to the Cauchy–Schwarz inequality:

$$\left(\sum_{i=1}^{n} X_i^2\right)\left(\sum_{i=1}^{n} Y_i^2\right) \geq \sum_{i=1}^{n} (X_i Y_i)^2$$

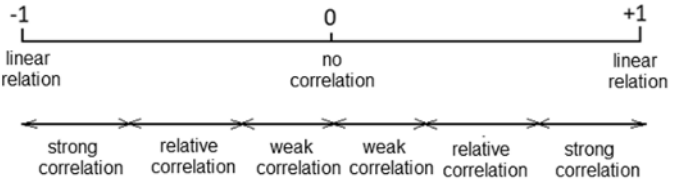

Figure 1.10. *Variations in the linear correlation coefficient*

This inequality implies that the linear correlation coefficient lies in the range $-1 \leq r \leq 1$. This means that the linear correlation coefficient can take values between -1 and 1, inclusive. Figure 1.10 shows such a correlation scale, where different ranges of r values are associated with specific degrees of correlation.

Note that there can be “accidental” correlations, as the following example shows. It is important to distinguish between correlation and causation.

EXAMPLE 1.4.– A statistical study of the French town of Perpette-les-Oisettes (Marne-et-Garonne) has determined the values of the following two variables over 10 successive years (see Figure 1.11): the annual number of personal computers purchased (x) and the annual number of recorded mental illnesses (y).

year	x	y
2013	124	25
2014	132	30
2015	144	35
2016	155	40
2017	166	45
2018	170	50
2019	180	55
2020	185	60
2021	195	65
2022	200	70

Figure 1.11. *The statistical series*

The calculations described in Figure 1.12 allow for the determination of the correlation coefficient r. The result is $r = 0.98$, indicating a strong correlation. The conclusion is left to the reader!

year	x	y	X	Y	X^2	Y^2	XY
2013	124	25	-41,10	-22,50	1689,21	506,25	924,75
2014	132	30	-33,10	-17,50	1095,61	306,25	579,25
2015	144	35	-21,10	-12,50	445,21	156,25	263,75
2016	155	40	-10,10	-7,50	102,01	56,25	75,75
2017	166	45	0,90	-2,50	0,81	6,25	-2,25
2018	170	50	4,90	2,50	24,01	6,25	12,25
2019	180	55	14,90	7,50	222,01	56,25	111,75
2020	185	60	19,90	12,50	396,01	156,25	248,75
2021	195	65	29,90	17,50	894,01	306,25	523,25
2022	200	70	34,90	22,50	1218,01	506,25	785,25
averages	165,10	47,50		sums	6086,90	2062,50	3522,50

Figure 1.12. *Calculation details*