Correlation analysis

Mohammad Saifuddin
Assistant Professor, Dept. of Mathematics and Statistics, BUBT

2024-11-05

Contents

1 Correlation

Correlation implies **degree** and **direction** of *linear relationship* between two *quantitative variables*. Example:

- As *income* increases *expense* also increases (positive correlation);
- As *resistance* increases *current flow* decreases (negative correlation) etc.

2 Scatter plot: Graphical method to explore correlation

A scatter plot shows the relationship between two quantitative variables measured for the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point on the graph.

- A scatterplot *displays* the strength, direction, and form of the relationship between two quantitative variables (*see* Figure 1).

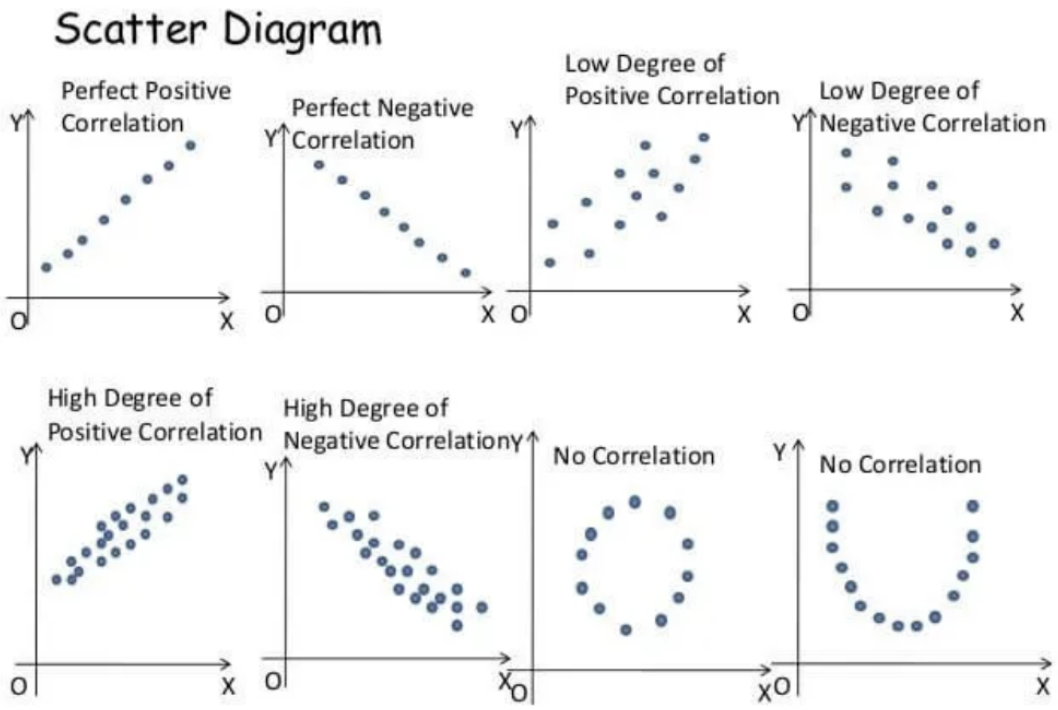

Figure 1: Types of correlations that can be represented using a scatterplot

3 Pearson correlation coefficient

The correlation coefficient *measures* the strength of the relationship between two quantitative variables.

Pearson correlation coefficient or **sample correlation coefficient** is denoted by $\boldsymbol{r}$ and computed as follows:

$$r = \frac{s_{xy}}{s_x s_y} \tag{1}$$

Note that, r is a number between -1 to +1 that is $-1 \leq r \leq +1$.

Here,

$$s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1} = \frac{\sum xy - n\bar{x}\bar{y}}{n-1} \quad (Sample \quad covariance \quad of \quad X,Y)$$

$$s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n-1}} \quad (Sample \quad standard \quad deviation \quad of \quad X)$$

$$s_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n-1}} = \sqrt{\frac{\sum y^2 - n\bar{y}^2}{n-1}} \quad (Sample \quad standard \quad deviation \quad of \quad Y)$$

3.1 An example: Computation and interpretation of r

Six observations taken for two variables follow.

x	30	50	40	55	30	25
y	28	25	25	23	30	32

a. Develop a scatter diagram for these data.
b. What does the scatter diagram indicate about a relationship between x and y?
c. Compute and interpret the sample covariance.
d. Compute and interpret the sample correlation coefficient

Solution

a. Scatter diagram between X and Y

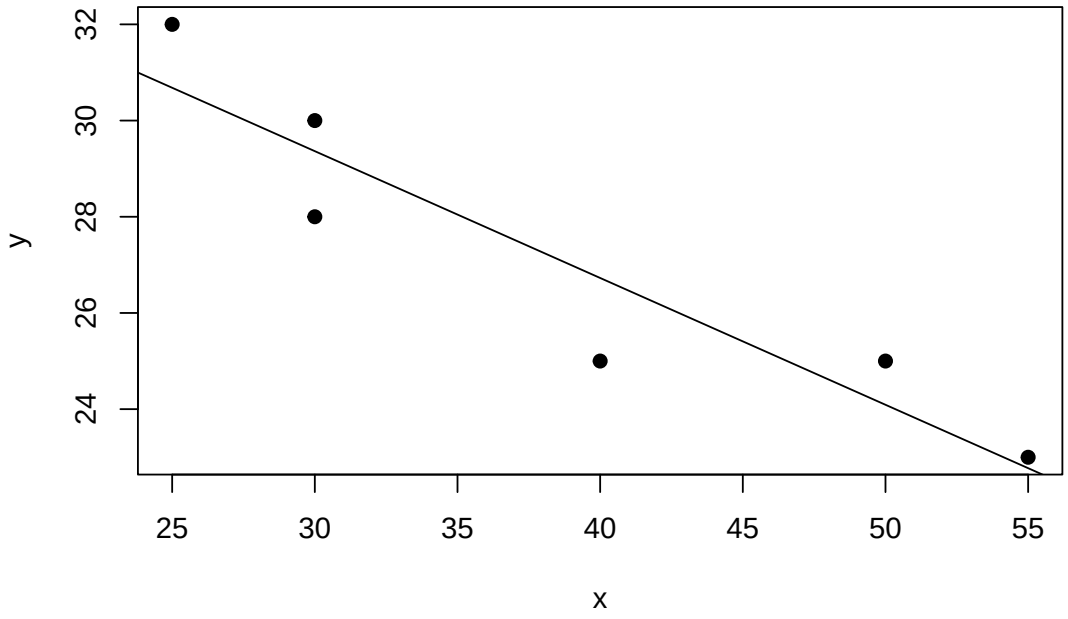

Figure 2: Scatter diagram between X and Y

b. The Figure 2 indicates *negative correlation* between X and Y.

c. We know the sample covariance is:

$$s_{xy} = \frac{\sum xy - n\bar{x}\bar{y}}{n-1}$$

Here,

- Total number of observations is: $n = 6$
- $\sum xy = x_1y_1 + x_2y_2 + ... + x_ny_n = 30 \times 28 + 50 \times 25 + ... + 25 \times 32 = 6055$
- $\sum x = 230$; $\sum y = 163$
- $\sum x^2 = 9550$
- $\sum y^2 = 4487$
- $\bar{x} = \frac{\sum x}{n} = \frac{230}{6} = 38.33$
- $\bar{y} = \frac{\sum y}{n} = \frac{163}{6} = 27.17$

So,

$$s_{xy} = \frac{\sum xy - n\bar{x}\bar{y}}{n-1} = \frac{6055 - 6 * 38.33 * 27.17}{6-1} = -38.67$$

The negative *covariance* indicates X and Y are negatively related.

d. We know, sample correlation coefficient:

$$r = \frac{s_{xy}}{s_x s_y}$$

Now

$$s_x = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{9550 - 6 * 38.33^2}{6-1}} = 12.12$$

and

$$s_y = \sqrt{\frac{\sum y^2 - n\bar{y}^2}{n-1}} = \sqrt{\frac{4487 - 6 * 27.17^2}{6-1}} = 3.40$$

Hence,

$$r = \frac{s_{xy}}{s_x s_y} = \frac{-38.67}{12.12 \times 3.40} = -0.9384$$

Interpretation of r: Since, $r = -0.9384$ so X and Y are strongly negatively correlated. That is, if X increases then Y also decreases and vice-versa.

3.2 Another example:

The following data shows the *Miles/gallon (mpg)* and *Horse power (hp)* of 10 automobiles.

hp	110	110	93.0	110.0	175.0	105.0	245.0	62.0	95.0	123.0
mpg	21	21	22.8	21.4	18.7	18.1	14.3	24.4	22.8	19.2

(a) **Construct** a *scatter plot.*

(b) **Compute** *correlation coefficient* between *mpg* and *wt*. Explain your finding.

Solution (a)

Figure 3 shows positive correlation between *horse power* and *mpg*.

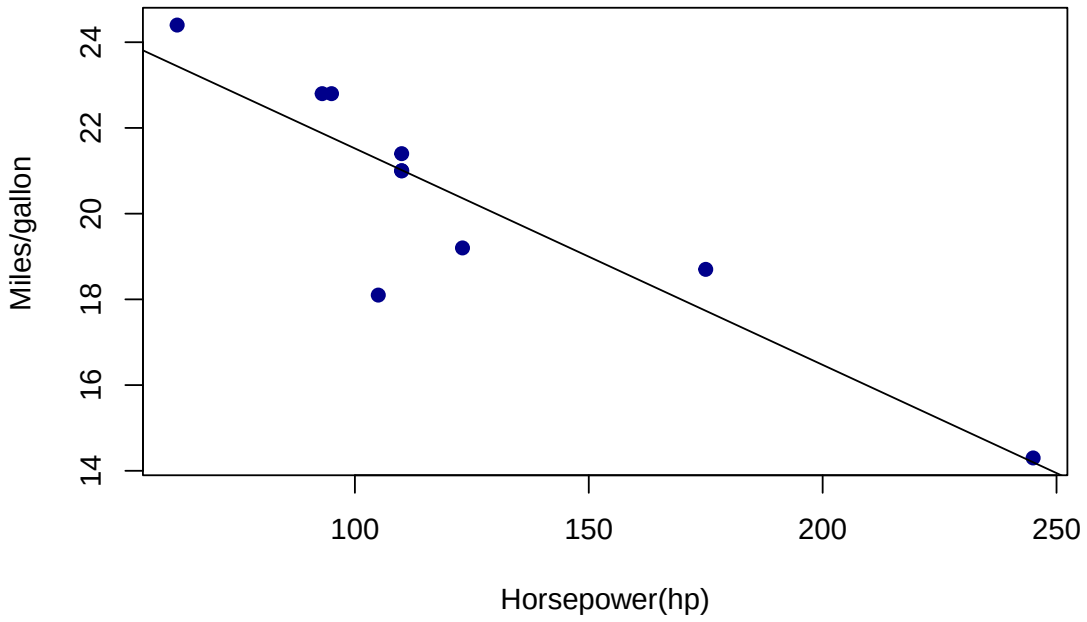

Figure 3: Scatter plot between HP and MPG

Solution (b)

- The correlation coefficient between *horse power* and *mpg* is -0.9308.
- It indicates that *mpg* is strongly and negatively correlated with *horse power*.

3.3 Correlation vs. causation

Correlation does not always imply *causation*. For example,

- A study(Messerli, 2012) found that there was a significant ($r = 0.791$) positive correlation between *chocolate consumption per capita* and *number of Nobel laureates per 10 million persons*. This does not necessarily implies that more a country consumes chocolate, more the chance of getting a Nobel prize. Rather differences in socioeconomic status from country to country and geographic and climatic factors may play some role to win a Nobel proze.
- We might find that there is a positive correlation between the time spent driving on road and the number of accidents but this does not mean that spending more time on road causes accident.Because in that case, in order to avoid accidents one may drive fast so that time spent on road is less (Selvamuthu & Das, 2024).

References

Messerli, F. H. (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates. *New England Journal of Medicine*, *367*(16), 1562–1564. https://doi.org/10.1056/nejmon1211064

Selvamuthu, D., & Das, D. (2024). *Introduction to probability, statistical methods, design of experiments and statistical quality control*. Springer Nature Singapore. https://doi.org/10.1007/978-981-99-9363-5