

Correlation and regression



Tudor Călinici
2026

Objectives



- Explain the correlation concept
- Interpret the value of the correlation coefficient
- Check the statistical significance of the correlation
- Estimate the dependent variable using the regression line equation
- Create the correlation chart

CORELLATION



- *Determine if there is a relationship between the variables X and Y (quantities / ordinals) and determine a way to measure the intensity of this relationship.*
- *Pearson's (Spearman's) coefficient of correlation*

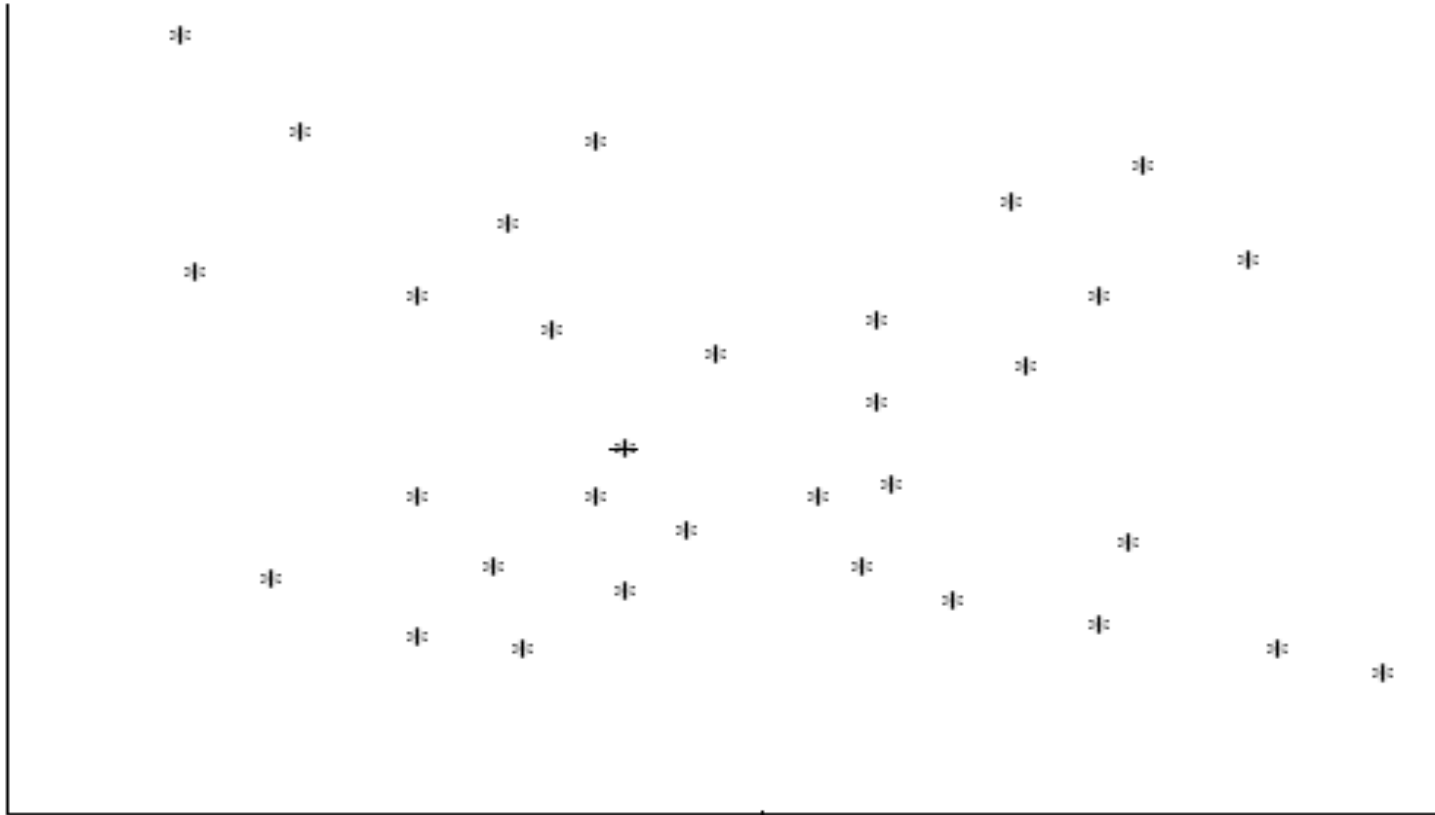
Variables



- Bivariate (numerical) series
- The connection between variables makes sense
- Independent variable/ dependent variable

- 
- For a group of n people, we observe two continuous characteristics, between which we have reasons to believe that there is a connection
 - For each person we will have the pair (x_i, y_i)
 - This pair can be graphically represented as a point having the coordinates (x_i, y_i)

Graphical representation



Correlation coefficients

Pearson's coefficient of correlation

$$r = \frac{COV(X, Y)}{S_X \cdot S_Y}$$
$$COV(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

Spearman's rank coefficient of correlation

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

ρ = Spearman's rank correlation coefficient

d_i = difference between the two ranks of each observation

n = number of observations

When to use



- If the data are normal distributed

- Pearson's R

- If the data are not normal distributed

- Spearman's ρ

Interpretation

- The sign of the coefficient - direction
 - If the sign of r (ρ) is + the variables are directly proportional - the increase of the independent variable implies the increase of the dependent variable
 - If the sign of r (ρ) is - the variables are inversely proportional - the increase of the independent variable implies the decrease of the dependent variable

Interpretation



- The value of the coefficient will provide the strength of the connection

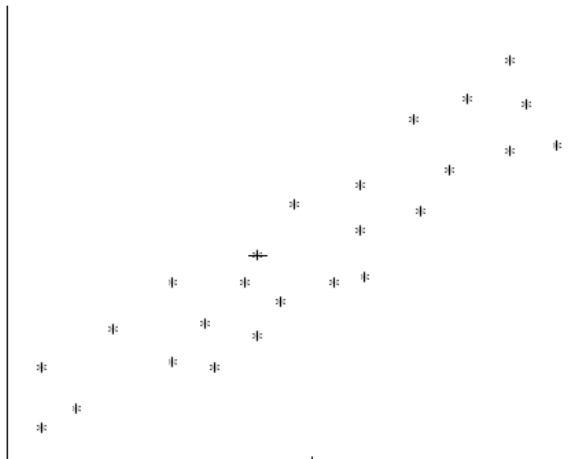
Correlation

□ Colton's Empirical Rule

- $r(p)$ between -0.25 and 0.25 = there is no correlation
- $r(p)$ between 0.25 and 0.50 (or -0.50 and -0.25) = a weak, to acceptable level of association
- $r(p)$ between 0.5 and 0.75 (or -0.5 and -0.75) = a moderate (acceptable) level of association to good
- $r(p)$ between 0.75 and 1 (or -0.75 and -1) = a good to very good association level

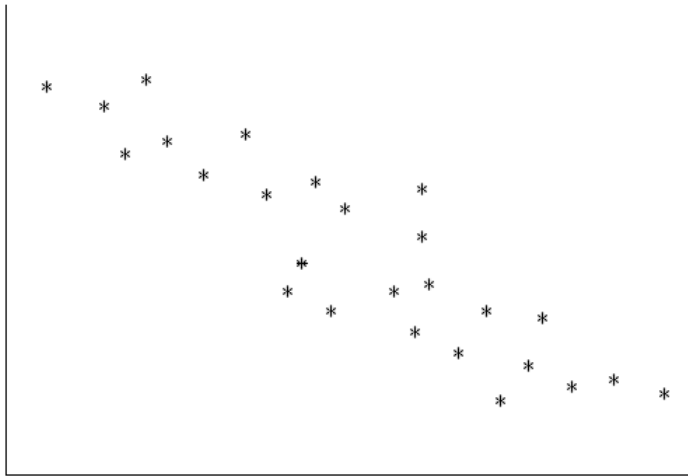
Example 1

□ Cholesterol – Blood Pressure; $R = 0,7$



Example 2

□ Age – calcium density; $R = -0,45$



Coefficient of determination

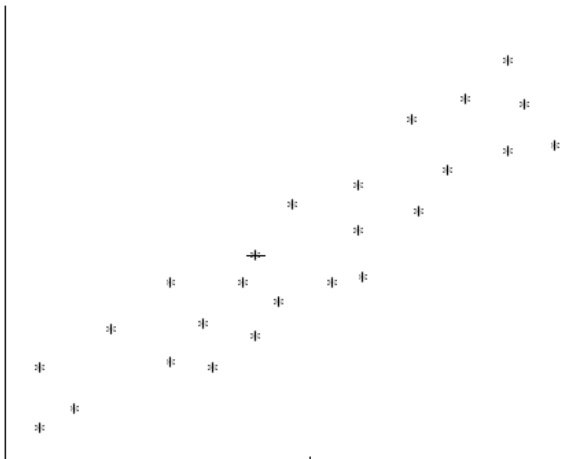
- R^2 expresses as a percentage how much of the variation of the dependent variable can be explained by the variation of the independent variable
- Example – Cholesterol and Blood pressure; $R=0,7$, $R^2=0,49$

Statistical significance

- If the correlation exists, we can find if it is statistically significant
- We'll use a statistical test for correlation
 - The null hypothesis: The correlation is not statistically significant
 - The alternative hypothesis: The correlation is statistically significant
 - The decision – according to the comparison $p < 0,05$

Example 1

□ Cholesterol – Blood Pressure; $R = 0,7$,
 $p=0,03$



Regressions



- *Numerical variables*
- *Find a way to estimate Y as a function of X*
 - *Regression function*
- Regression line $Y(X)$:
$$y = a + b x$$

Situation



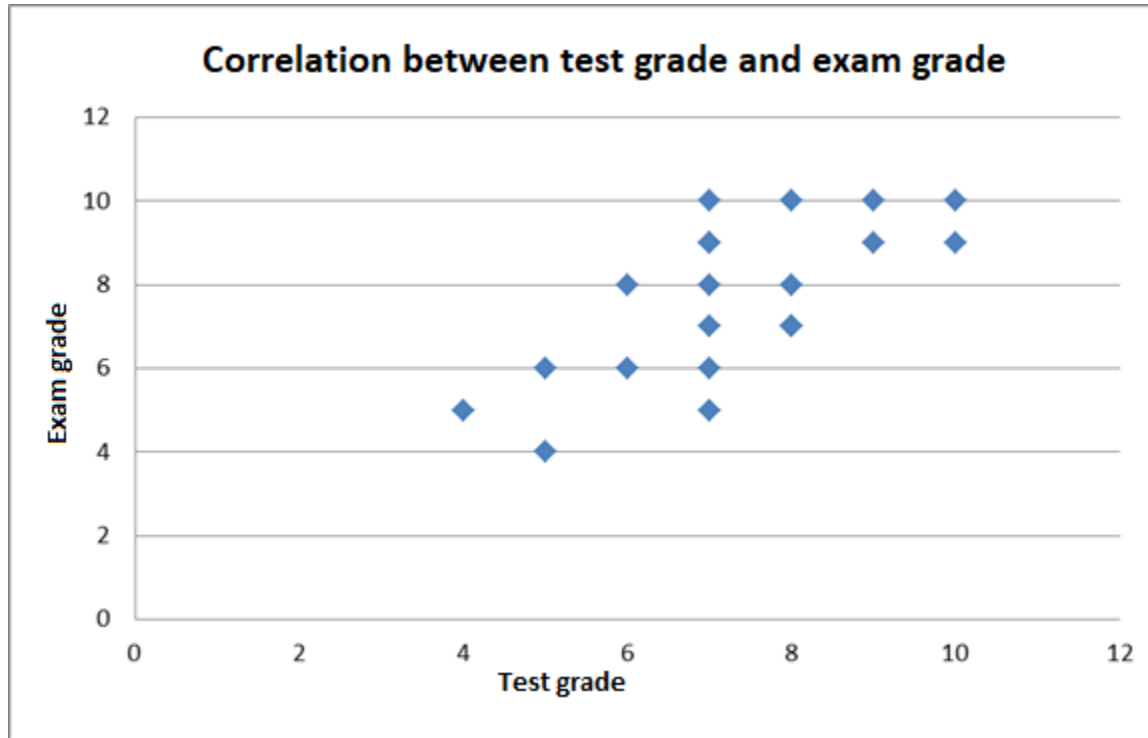
- 100 students took part in an exam. In order to have an idea regarding the obtained results, 20 tests were corrected, the obtained grades being displayed (anonymously).
- 8, 4, 6, 6, 9, 7, 9, 10, 8, 10, 10, 7, 9, 6, 7, 9, 8, 10, 5, 5
- Mean= 7.7; Standard Deviation = 1.9

Situation

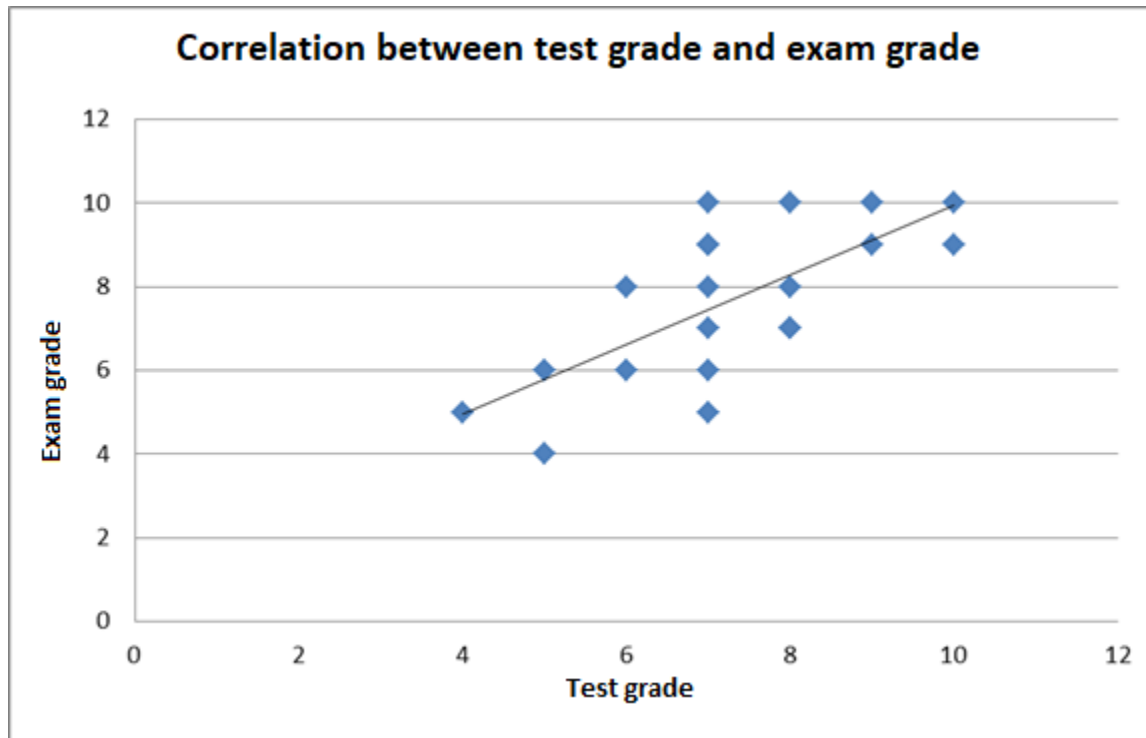
- In the middle of the semester, the students took a test. For those students, whose grades were displayed, the teacher also displayed the grade obtained in the test.

TG	EG
7	8
5	4
7	6
6	6
10	9
8	7
7	9
9	10
6	8
10	10
8	10
8	7
7	9
5	6
7	7
9	9
8	8
7	10
7	5
4	5

Graphical representation



Regression line



Coefficients of the regression line

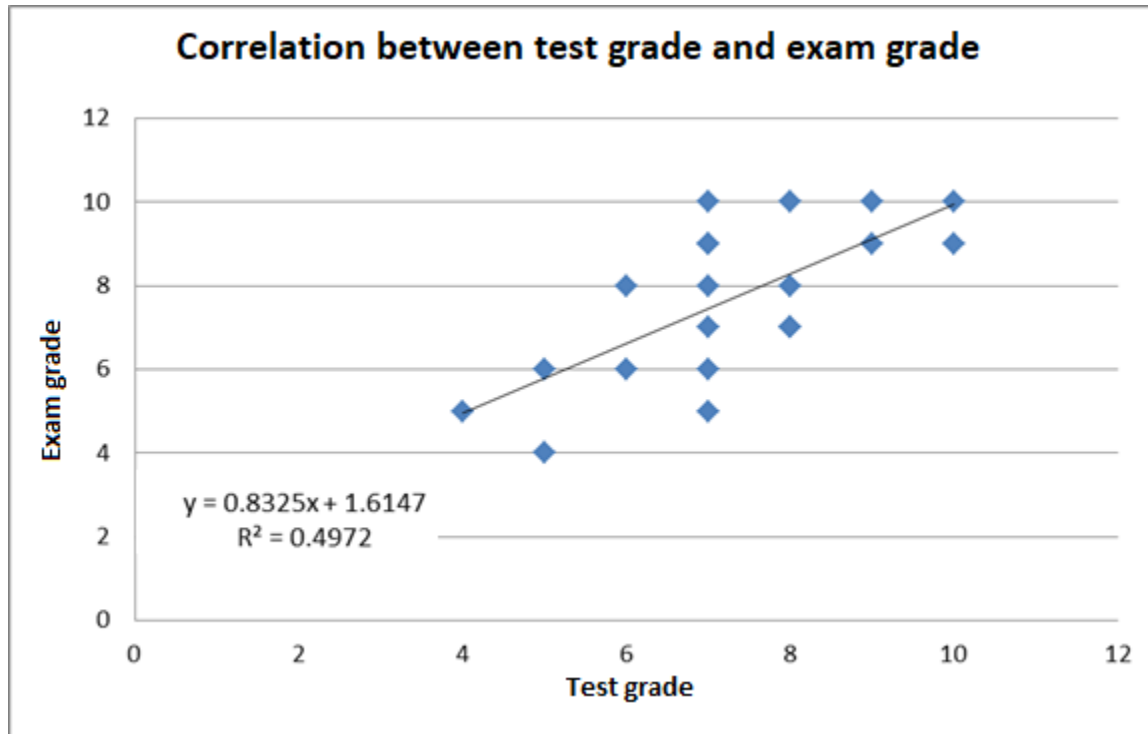
Regression line $Y(X)$

$$\min_{a, b \in \mathbb{R}} \sum_{i=1}^n (a + bX_i - Y_i)^2$$

$$b = \frac{\hat{\text{COV}}(X, Y)}{s_x}.$$

$$a = \bar{Y} - b \cdot \bar{X}$$

Correlation chart



A student having 7 on the test

$$(\text{Exam grade}) = 7 * 0,8325 + 1,6147 = 7,47$$

A student having 5 on the test

$$(\text{Exam grade}) = 5 * 0,8325 + 1,6147 = 5,78$$

Correlation chart



- It is scatter type
- At all times, the first variable (the one on the left) will be represented on the OX axis
- It NEVER has a legend!
- It is not complete until the regression line is drawn
- It contains the equation of the regression line as well as the coefficient of determination

Regression line



- The equation of the regression line allows the estimation of the dependent variable as a function of the independent variable
- The estimation is better when the correlation is stronger (the value of the correlation coefficient is close to 1 and -1 respectively)

CORRELATIONS AND REGRESSIONS – REMEMBER!

Correlations



- You must check the type of the variables
- If the variables are numerical, the connection must logically exist

Variables



- Be careful when you are choosing the variables – which variable is the independent, which variable is dependent!

Coefficient of correlation



- When the data are quantitative normal distributed, use Pearson's coefficient of correlation – R
- When the data are not quantitative / normal distributed use the Spearman's coefficient of correlation – ρ

Direction of the relationship



- The sign of the correlation coefficient (r or ρ)

or

- The sign of the x coefficient from the equation of the regression line



**Correlation is NOT
causality!!!!**