MAKE SCHOOL

Correlation and Regression

QL 1.1

Learning Outcomes

MAKE SCHOOL

By the end of this session, you should be able to...

1. Draw a scatter plot for a set of ordered pairs.
2. Compute the correlation coefficient.
3. Compute the equation of the regression line.
4. Apply the newly acquired knowledge to calculate correlation and regression in a dataset using numpy and python library.

Relationship among variables

MAKE SCHOOL

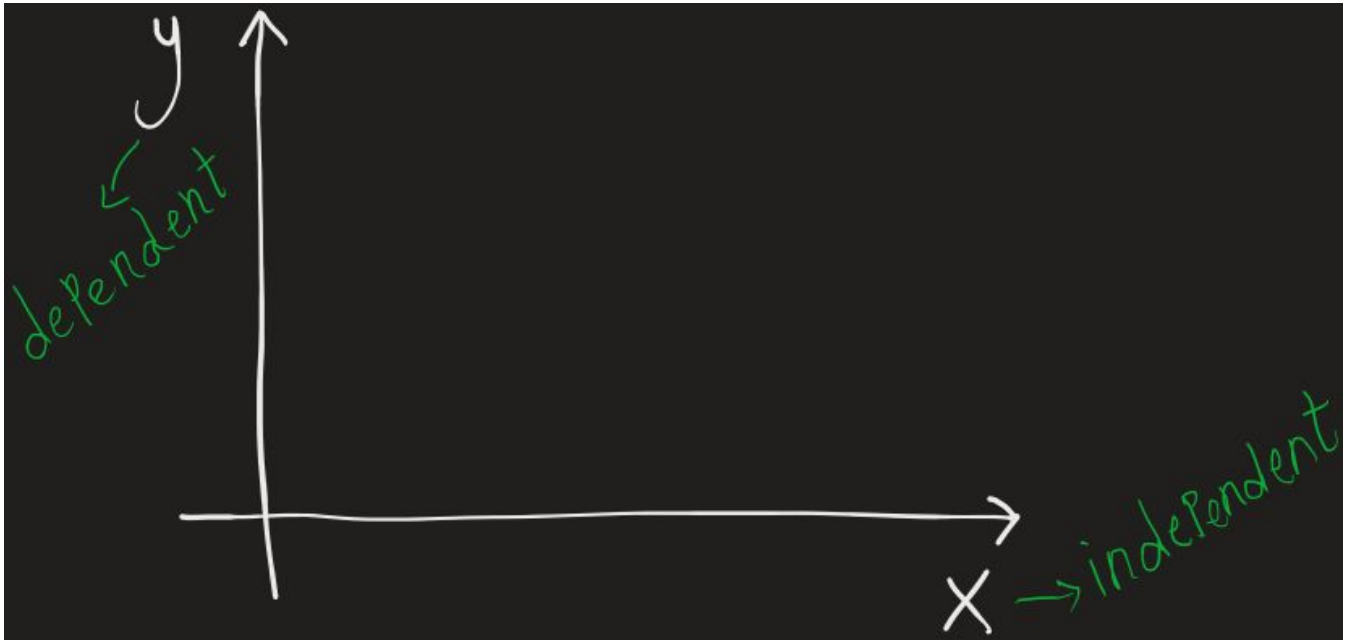
- Effect of dust on respiratory health?
- Study hours ↔ Student's score on an exam
- Is coffee related to heart damage?
- A businessperson wants to know whether the sale for a given month is related to the amount of advertising the firm does that month

Correlation

- **Correlation** is a method to determine whether a relationship between variables exists.

Scatterplot and Correlation

- A scatterplot is a graph of the ordered (x, y) of numbers consisting of the independent variable x and dependent variable y.



Scatterplot example

The following table shows the hours of study for each students and their corresponding grades. Plot the data using scatterplot.

Student	Hours of study x	Grade y (%)
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

Absences and Final Grades

5 min

Click here for scatterplot activity.

Correlation Coefficient

The correlation coefficient computed from the data measures the **strength** and **direction** of a **linear** relationship between two variables.

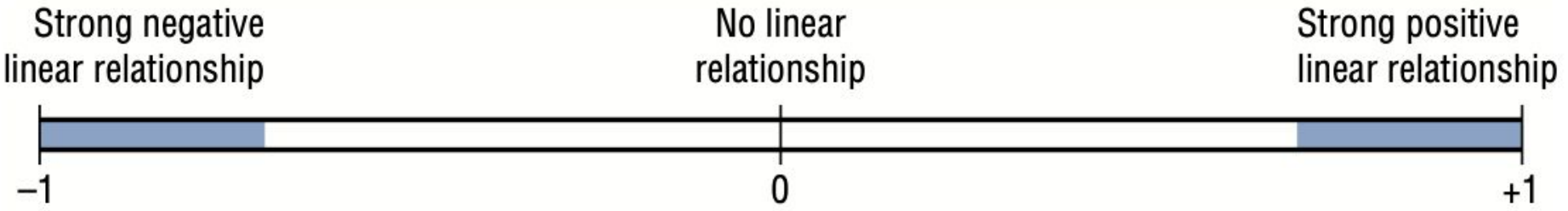
r: **Sample** correlation coefficient

ρ: **Population** correlation coefficient

Correlation Coefficient

MAKE SCHOOL

$$-1 < \text{Correlation Coefficient} < +1$$



Range of Values for the Correlation Coefficient

Correlation Coefficient Formula

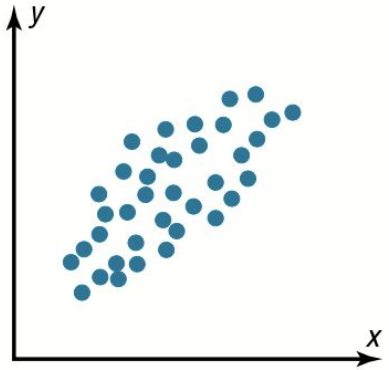
Do not memorize this formula!

Formula for the Correlation Coefficient *r*

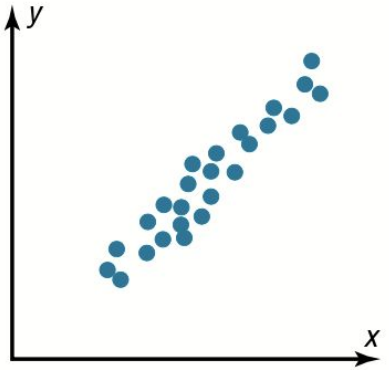
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

where n is the number of data pairs.

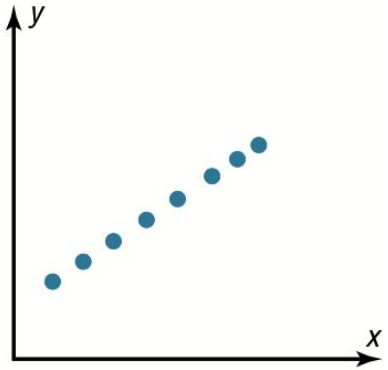
Correlation Coefficient and Scatterplot



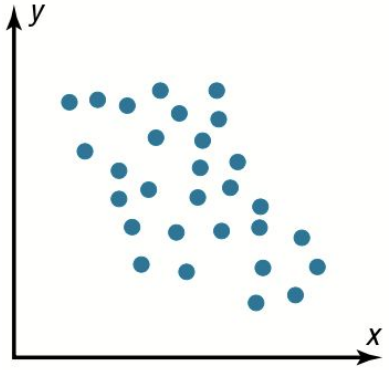
(a) $r = 0.50$



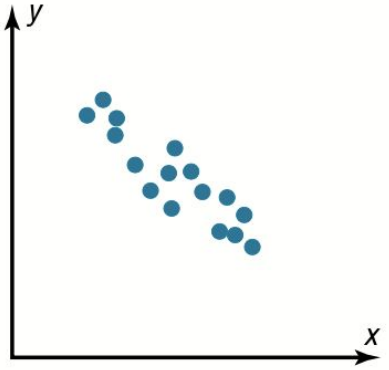
(b) $r = 0.90$



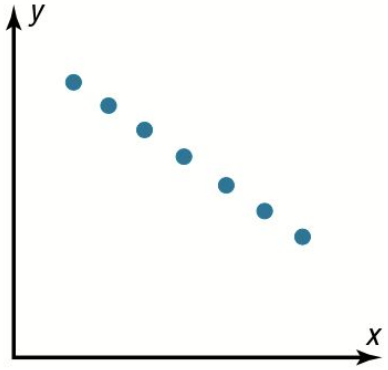
(c) $r = 1.00$



(d) $r = -0.50$



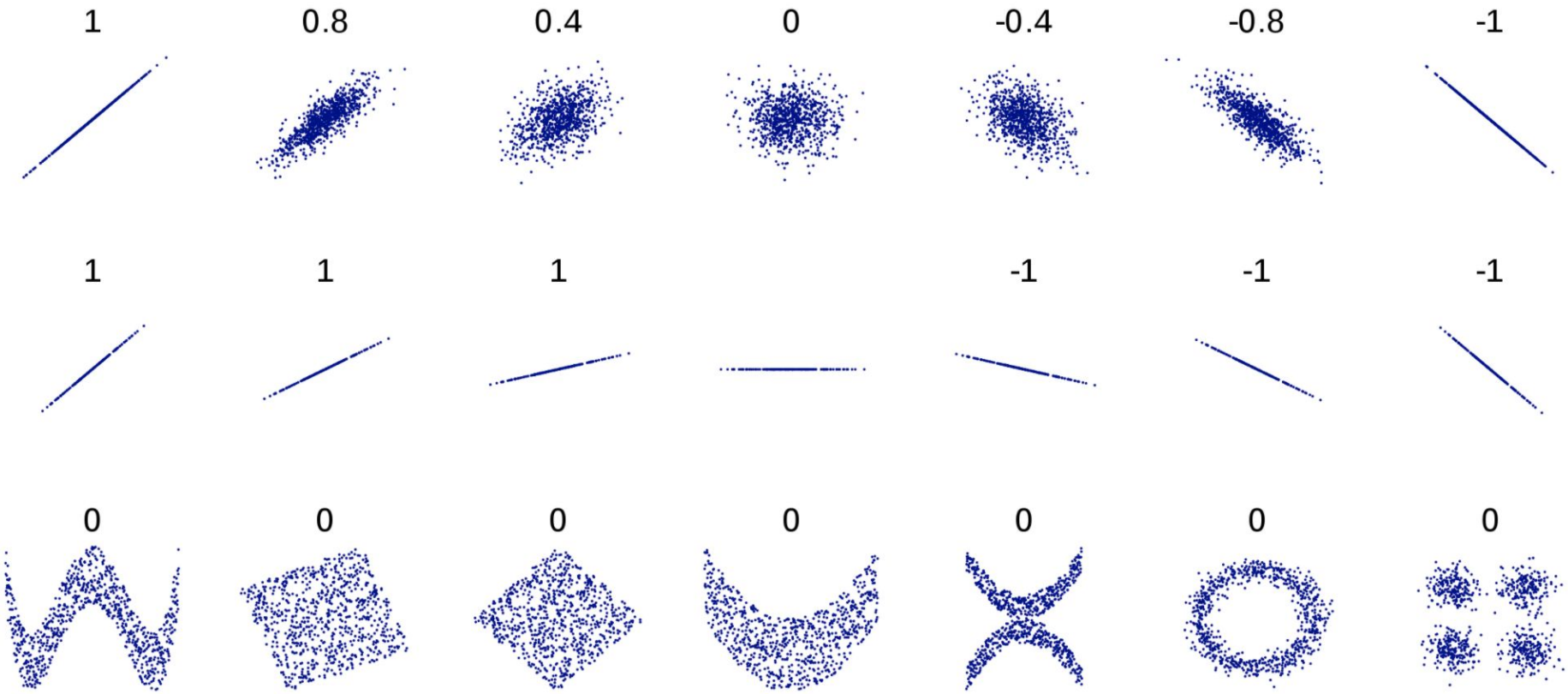
(e) $r = -0.90$



(f) $r = -1.00$

Break 10 mins

Correlation Coefficient and Scatterplot



Look at the scatterplot before calculating correlation!

MAKE SCHOOL

- If the scatterplot doesn't indicate there's at least somewhat of a linear relationship, the correlation doesn't mean much.
- you can take the idea of no linear relationship two ways:
  a. If no relationship at all exists, calculating the correlation doesn't make sense because correlation only applies to linear relationships
  b. If a strong relationship exists but it's not linear, the correlation may be misleading, because in some cases a strong curved relationship exists.

When to get excited?

TIP

How close is close enough to –1 or +1 to indicate a strong enough linear relationship? Most statisticians like to see correlations beyond at least +0.5 or –0.5 before getting too excited about them. Don't expect a correlation to always be 0.99 however; remember, these are real data, and real data aren't perfect.

Computing Correlation Coefficient

MAKE SCHOOL

The following table shows the hours of study for each students and their corresponding grades. Compute the correlation coefficient.

Student	Hours of study x	Grade y (%)
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

Absences and Final Grades

5 mins

Click here for the activity.

(go to the activity 2)

Correlation and Causation

5 mins

Assume there is a high correlation between two variables. Then what we can say about the two variables?

Discuss it in groups.

Correlation & Causation

MAKE SCHOOL

1. There is a **direct** cause-and-effect relationship between the two variables. That is, x causes y.
2. There is a **reverse** cause-and-effect relationship between the variables. That is, y causes x.
3. The relationship between the variables may be caused by a third variable (lurking variable)
4. There may be a complexity of interrelationships among many variables.
5. The relationship may be coincidental.

Regression

MAKE SCHOOL

You are studying the relationship between two variables:

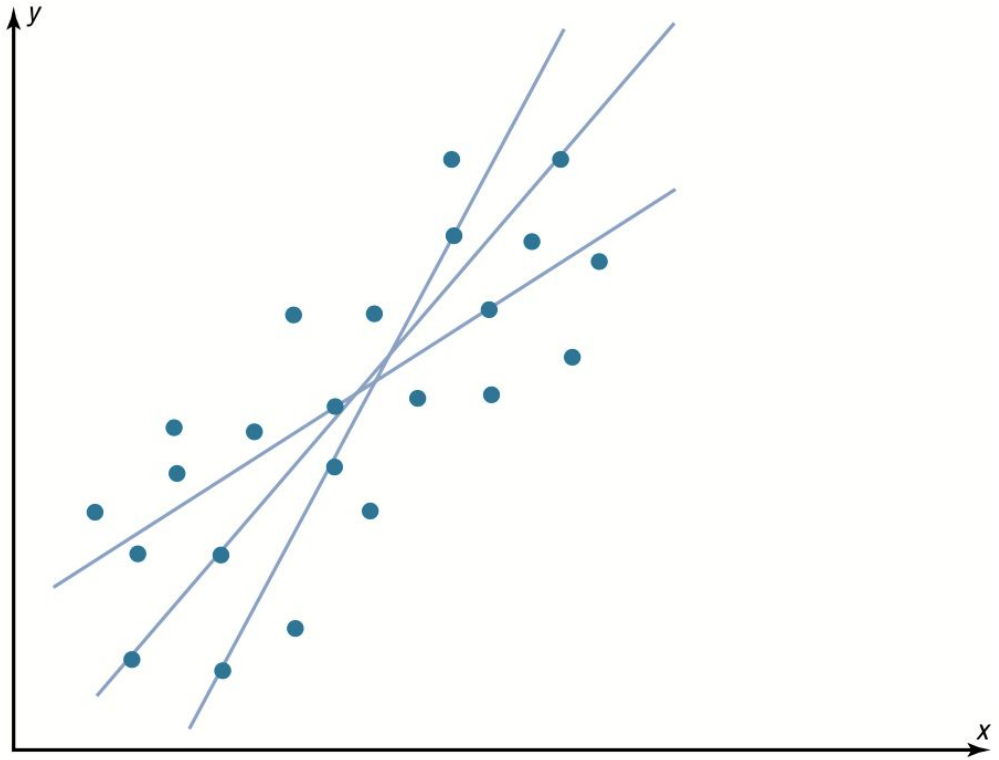
1. You collect data
2. Construct a scatterplot
3. Compute the correlation coefficient
4. Test the significance of the relationship (* p-value, out of ql1.1 scope)
5. If the value of corr. Coef. is significant, then the next step is to determine the equation of the **regression line.**

Regression Line

The purpose of the regression line is to enable the researcher **to see the trend and make predictions. (inferential statistics)**

Note: Determining the regression line when r is not significant and then making prediction using the regression line are meaningless.

Line of best fit

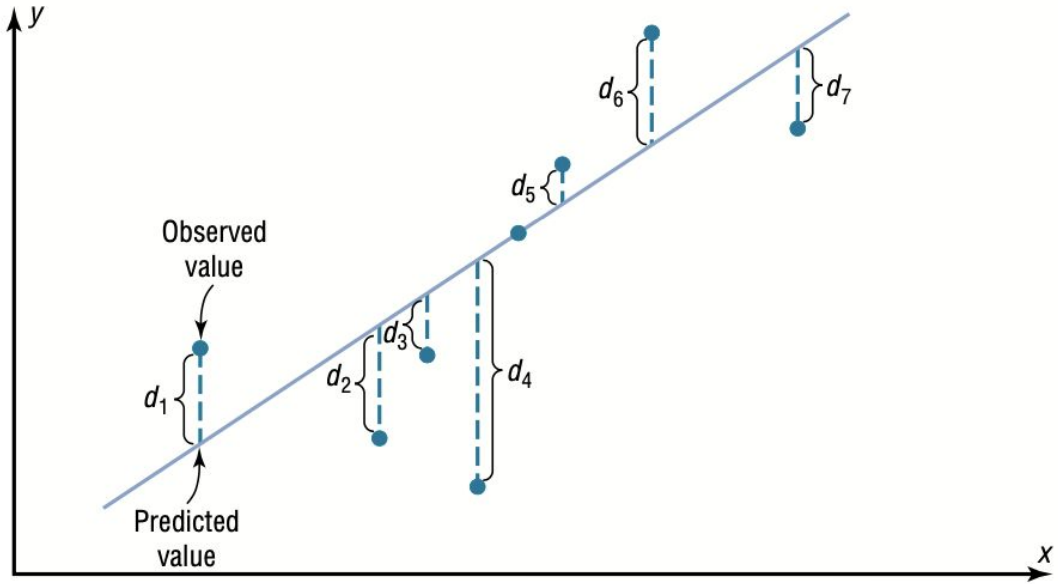


Scatter Plot with Three Lines Fit to the Data

How to find the line with best fit?

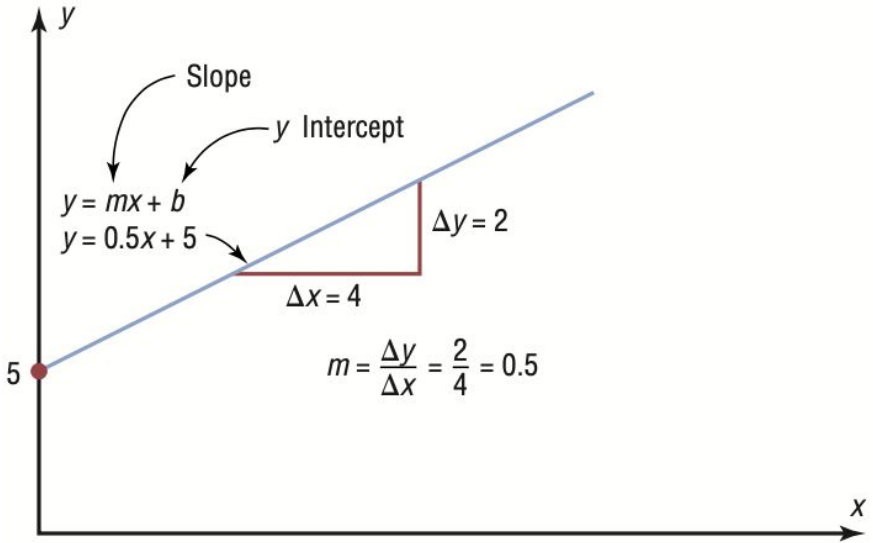
The best fit in the least-squares sense minimizes the sum of squared residuals. That is the line that gives you the minimal of

$$d_1^2 + d_2^2 + d_3^2 + \ldots + d_n^2$$

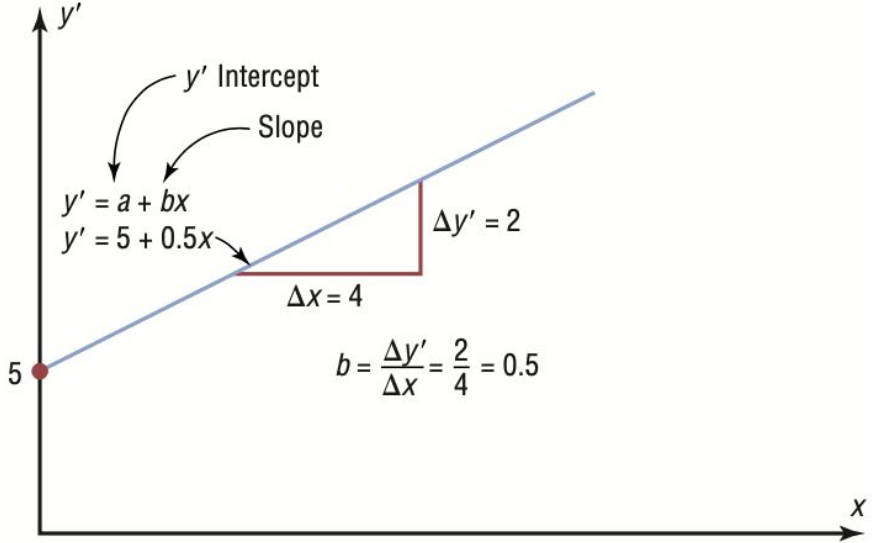


Regression Line Notation

A Line as Represented in Algebra and in Statistics



(a) Algebra of a line



(b) Statistical notation for a regression line

$$y = mx + b$$

$$y' = a + bx$$

Formulas for the Regression Line y' = a + bx

MAKE SCHOOL

Formulas for the Regression Line $y' = a + bx$

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

where a is the y' intercept and b is the slope of the line.

Calculating Regression Line - Example

Find the equation of the regression line for the data below, and graph the line on the scatter plot of the data.

Company	Cars (in ten thousands)	Revenue (in billions)
A	63.0	$7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

Source: *Auto Rental News.*

Calculating Regression Line

7 mins

Calculating Regression Activity link