Week 8 - Introduction to Regression and Correlation Models

Introduction:

Welcome to this seminar on linear regression in the context of mathematical modeling. Linear regression and Correlation are related statistical techniques used to model the relationship between one or more independent variables and a dependent variable.

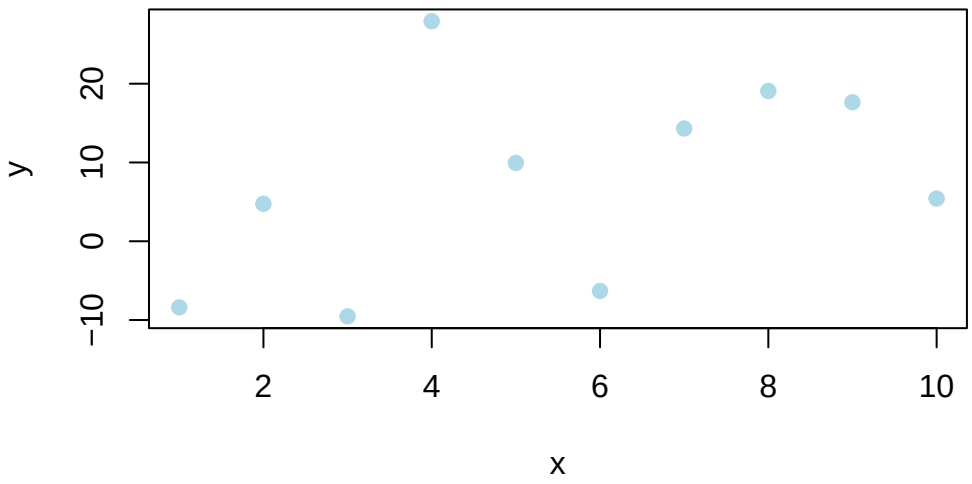
In this seminar, we will explore the principles of linear regression, correlation, and confidence intervals. We will also look at its applications, and how it fits into mathematical modeling.

Understanding Correlations

We may be interested in seeing if there is a linear relationship between 2 (scale) variables.

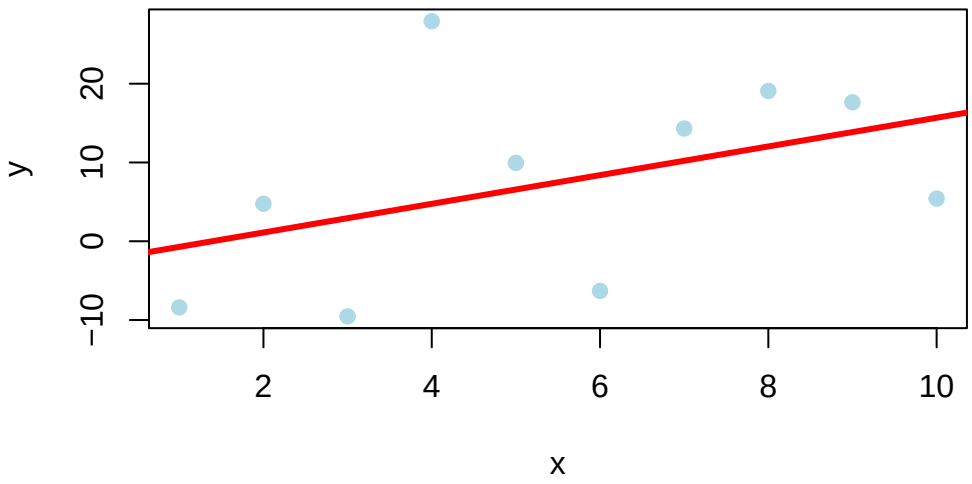
We can plot the values on a scatter diagram and see how close the points are to a straight line

Consider the following plot, where each point is one data point. Each point represents the data for one person. The scatterplot displays two numerical variables, x (independent or predictor) and y (dependent)



The relationship between the two variables is not exact – there is an unpredictable or random element: every person is different.

We can model the underlying relationship by a straight line of best fit (or trendline) We can measure how close those points are to that line (below, the trendline is in red)



Pearsons Product Moment Correlation, r

The correlation coefficient, r, measures how close the points are to the line ie. the strength of the linear relationship.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

where:

$$S_{xy} = \sum(x - \bar{x})(y - \bar{y}) = \sum xy - \frac{\sum x \sum y}{n})$$

$$s_{xx} = \sum(x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$s_{yy} = \sum(y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

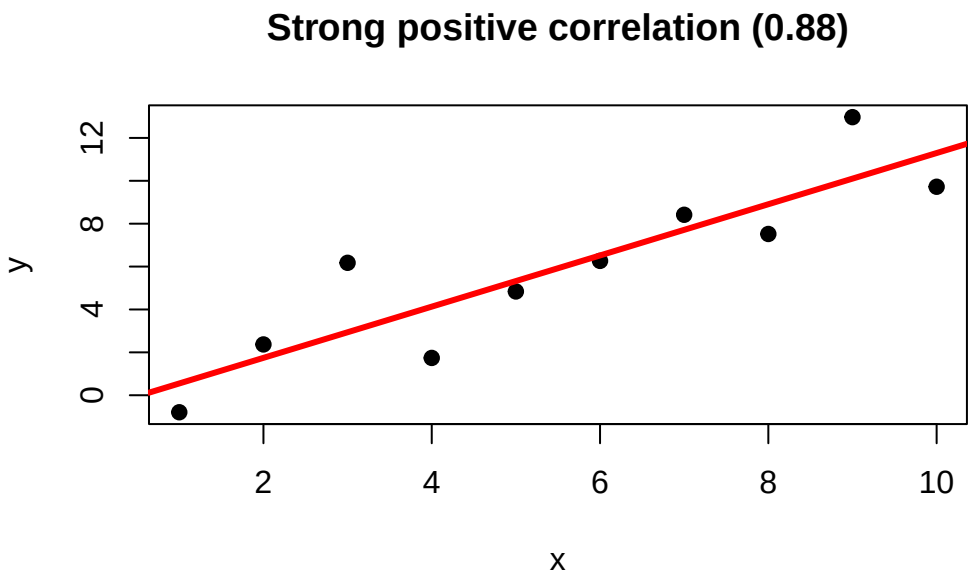
The attributes of the Pearson Correlation Coefficient (r) are:

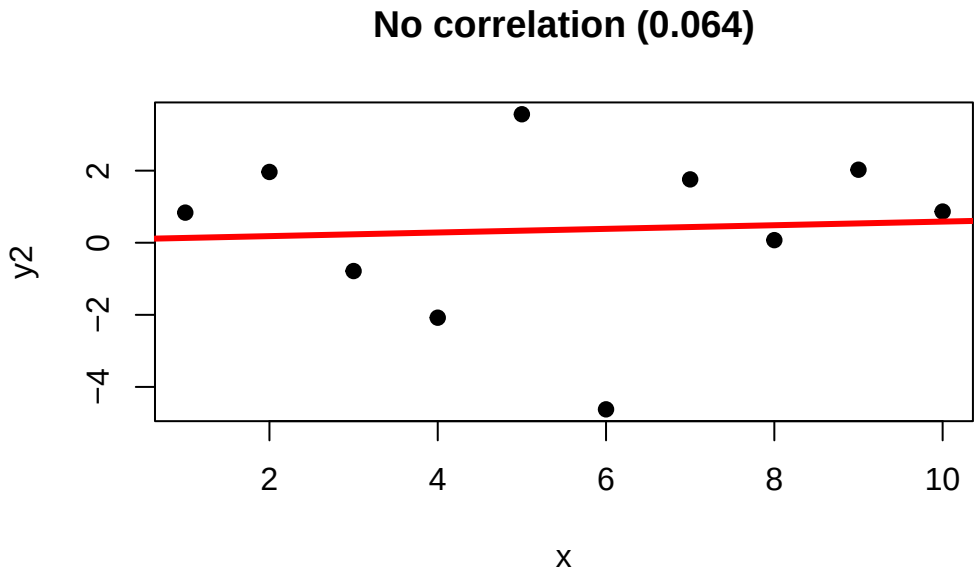
- Its value is between -1 and 1

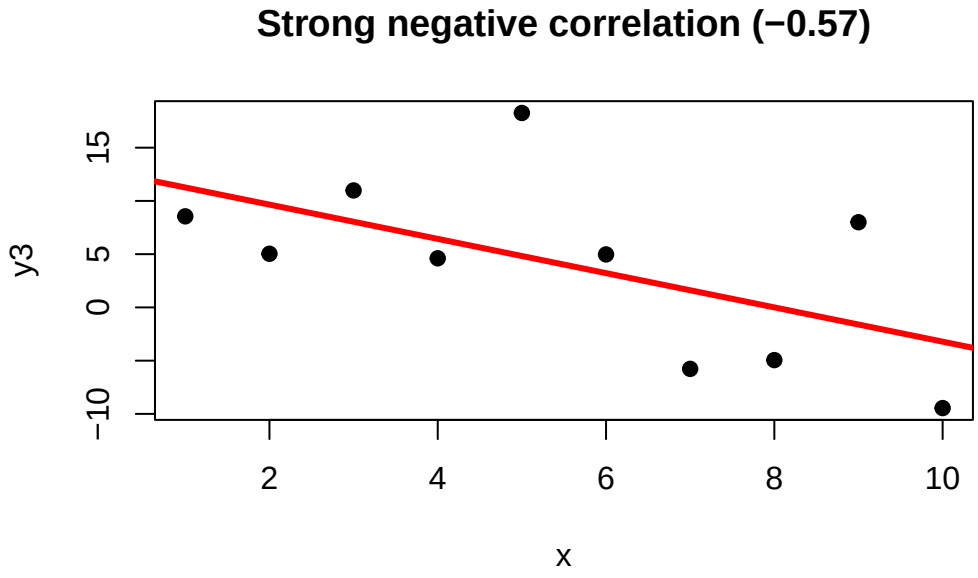
- A value close to 1 or -1 represents a strong linear relationship.
- A positive value suggests a positive relationship ie. as one variable increases the other increases
- A negative value suggests a negative relationship ie. as one variable increases the other decreases
- A value close to 0 means a very weak or no linear relationship.

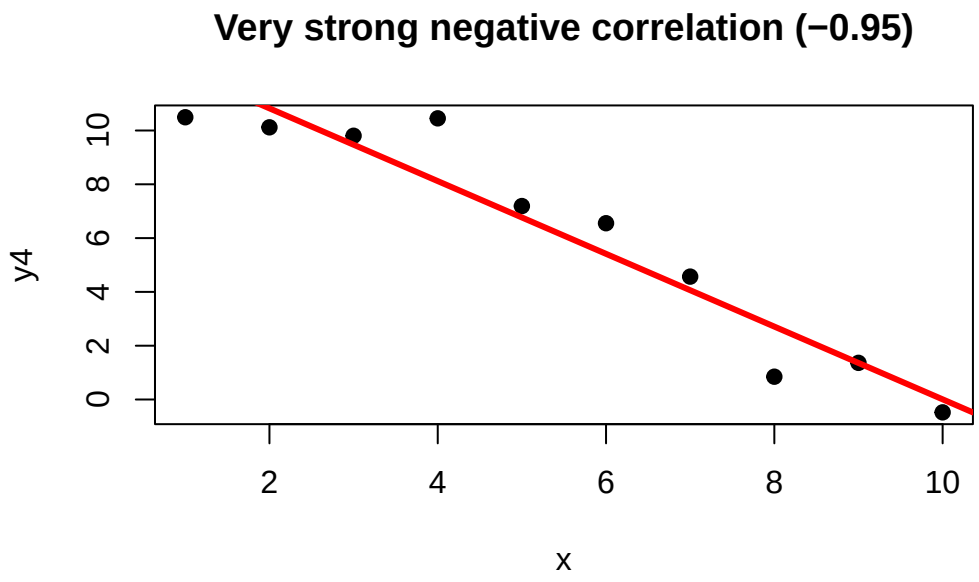
(Note that even if there is no linear relationship, there still may be a relationship which is non-linear)

Some examples of correlations




No correlation (0.064)
y2
2
0
−2
−4
2
4
6
8
10
x


Strong negative correlation (−0.57)
y3
15
5
0
−10
2
4
6
8
10
x



Consider the following example. The number of vehicles, millions, and the number of accidents, thousands, in 15 different countries were:

Country	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Vehicles (x) millions	8.6	13.4	12.8	9.3	1.3	9.4	13.1	4.9	13.5	9.6	7.5	9.8	23.3	21	19.4
Accidents (y) thousands	33	51	30	48	12	23	46	18	36	50	34	35	95	99	69

Calculate the product moment correlation coefficient for the number of vehicle and the number of accidents.

We will need the following

$$\sum x = 176.9 \sum y = 679$$
$$\sum x^2 = 2576.47 \sum y^2 = 39771$$
$$\sum xy = 9915.3$$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 9915.3 - \frac{176.9 * 679}{15} = 1907.62...$$

$$s_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 257.47 - \frac{(176.9)^2}{15} = 490.22...$$
$$s_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 39771 - \frac{(679)^2}{15} = 9034.93...$$
$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{1907.62...}{\sqrt{490.22... * 9034.93...}} = 0.906(3.d.p)$$

This suggests a very strong positive linear relationship between the number of vehicles and the number of accidents.

Consider the following example. The head circumference in cm (x) and gestation period (y) in weeks for new born babies at a certain clinic over a period of time is as follows.

Baby	A	B	C	D	E	F
Head circumference (x)	31.1	33.3	30.0	31.5	35.0	30.2
Gestation period (y)	36	37	38	38	40	40

Find the correlation between head circumference and gestation period.

We will need the following

$$\sum x = 191.1 \sum y = 229$$
$$\sum x^2 = 6105.39 \sum y^2 = 8753$$
$$\sum xy = 7296.7$$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 7296.7 - \frac{191.1 * 229}{6} = 3.05$$

$$s_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 6105.39 - \frac{(191.1)^2}{6} = 18.855$$
$$s_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 8753 - \frac{(229)^2}{6} = 12.833$$
$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{3.05}{\sqrt{18.855 * 12.833}} = 0.196$$

This is a low correlation which suggests very little evidence of a linear relationship between head circumference and gestation period

Some problems with interpreting correlations

A very strong correlation does not necessarily mean that a change in x causes a change in y

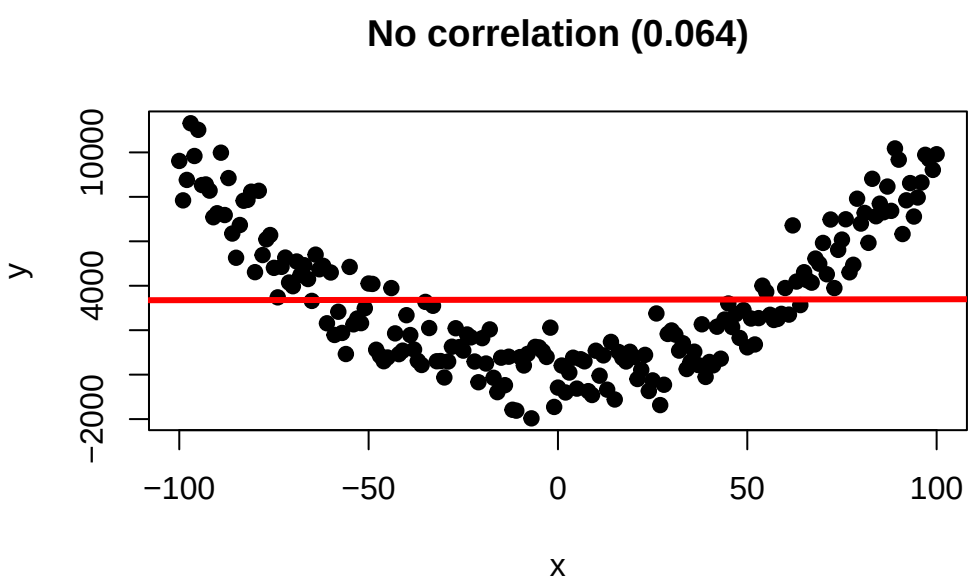
Pirate Population and Global Warming: An amusing example often cited is the inverse relationship between the number of pirates and global temperatures. As the number of pirates has decreased over the centuries, global temperatures have risen. Of course, the decline in piracy doesn't cause global warming; this highlights how unrelated variables can appear correlated if we ignore other factors.

Storks and Birth Rates: In some European countries, statistical analyses have shown a positive correlation between stork populations and human birth rates. This doesn't mean storks deliver babies, but rather that both are influenced by third factors, such as rural settings, which might support larger stork populations and also have higher human birth rates due to different socio-economic factors.

Internet Explorer Use and Murder Rates: This is a facetious example showing that the declining use of Internet Explorer correlates with a decrease in murder rates. The implication that discontinuing Internet Explorer use reduces murder rates is absurd, illustrating how correlation alone is not sufficient for establishing a causal relationship.

A very weak correlation does not necessarily mean that x and y are unrelated. The relationship could be non-linear.

Consider the following plot



Given the line of best fit, and the correlation of 0.0038. We would assume that there is no relationship between x and y. However, looking at the graph we can see non-linear relationship.

High correlations can arise purely by chance, particularly if the sample size is small. The correlation may not be significantly different from zero.

Diet and Health Outcomes in Small Communities: Suppose a study looks at the health outcomes of people following a specific diet in a very small community. If this community happens to have unusually high or low rates of a certain health condition, it might falsely appear that the diet is responsible for these health outcomes. However, with such a small sample size, the findings might not be generalizable to larger populations.

Simple Linear Regression:

In simple linear regression, there is only one independent variable (X) and one dependent variable (Y). The relationship is expressed as $y = a + bx$. The goal is to find the best-fitting line that minimizes the sum of squared errors (residuals) between the predicted and actual values of Y.

a is the intercept of the line.
b is the slope of the line.

In order to calculate a and b, you need the following

$$a = \bar{y} - b\bar{x}$$

and

$$b = \frac{S_{xy}}{S_{xx}}$$

Wait, we've seen the terms S_{xy} and S_{xx} before when working out correlation!

Example: The results from an experiment in which different masses were placed on a spring and the resulting length of the spring measured, are shown below

Mass, x (kg)	20	40	60	80	100
Length, y (cm)	48	55.1	56.3	61.2	68

We will need the following

$$\sum x = 300, \sum y = 288.6$$

$$\sum x^2 = 22000$$

$$\bar{x} = 60, \bar{y} = 57.72$$

$$\sum xy = 18238$$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 18238 - \frac{300 * 288.6}{5} = 922$$

$$s_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 22000 - \frac{(300)^2}{5} = 4000$$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{922}{4000}$$

$$a = \bar{y} - b\bar{x} = 57.72 - 0.2305 * 60 = 43.89$$

$$y = 43.89 + 0.2305x$$

Predict and interpret:

a. the value of y when the mass x = 35 kg
b. the value of y when the mass x = 120 kg

a)

$$y = 43.89 + 0.2305 * 35 = 52.0cm$$

If the mass is 35 kg, then the length of the spring will be 52 cm on average

b)

$$y = 43.89 + 0.2305 * 120 = 71.55cm$$

If the mass is 120 kg, then the length of the spring will be 71.6 cm on average

In terms of predicting new values for the above example. The range of observed values of x was from 20 to 100. If we predict the value of Y for 20 100, this is called INTERPOLATION If the prediction is made outside this range, this is called EXTRAPOLATION

Extrapolation is generally unreliable since we have no way of knowing whether model assumptions remain valid outside the range of our observations

Hence the earlier prediction 'If the mass is 120 kg, then the length of the spring will be 71.6 cm on average' should be treated with caution

Applications of Linear Regression:

- Linear regression is widely used in various fields, including:
- Economics: Modeling economic factors and predicting outcomes.
- Finance: Predicting stock prices and risk assessment.
- Medicine: Predicting patient outcomes based on medical variables.
- Engineering: Predicting performance and optimizing processes.
- Social Sciences: Analyzing social and behavioral data.

Model Evaluation:

Evaluating a linear regression model is essential to assess its quality and predictive power. Common evaluation metrics include:

- R-squared (R^2): Measures the proportion of variance explained by the model.
- Mean Squared Error (MSE) and Root Mean Squared Error (RMSE): Measure the average squared error of predictions.
- Residual plots: Visualize the distribution of residuals to check for patterns or anomalies.

Conclusion

Linear regression is a powerful mathematical modeling technique that models relationships between independent and dependent variables using linear equations. It is widely applicable in various fields for prediction, analysis, and decision-making. Understanding the assumptions and techniques of linear regression is crucial for effective modeling and interpretation.