IT 347 328

Department of Information Technology

# Probability and Statistics

## Lecture 5: Correlation and Simple Linear Regression

Salisu Ibrahim

Tishk International University

# Learning Outcomes

By the end of this lecture, you will:

- Understand the concept of correlation and its role in measuring the relationship between two variables.
- Learn to create and interpret scatterplots to visually assess the relationship between variables.
- Identify and differentiate between the types of correlation, including positive, negative, no correlation, and non-linear correlation based on calculations and based on scatterplot.
- Calculate the linear correlation coefficient (Pearson's r) to quantify the strength and direction of the linear relationship between variables.
- Master the concept of simple linear regression and its application in predicting one variable from another using a linear equation.
- Develop proficiency in fitting a regression line to a scatterplot and interpreting its slope and intercept in the context of the relationship between variables.

# Introduction

In many studies, we measure more than **one variable** for **each individuale**. For example, we measure

a) Time Spent Running vs. Body Fat
b) Time Spent Watching TV vs. Exam Scores
c) Height vs. Weight
d) Temperature vs. Ice Cream Sales
e) Precipitation and plant growth
f) Number of young with nesting habitat

We collect **pairs of data** and instead of examining each variable separately (**univariate data**), we want to find ways to describe **bivariate data**, in which two variables are measured on each subject in the sample.

- Given such data, we begin by determining if there is a **relationship between these two variables**.
- As the values of **one variable change**, do we see **corresponding changes in the other variable**?

We can describe the relationship between these two variables **graphically** and **numerically**. We begin by considering the concept of **correlation**.

**Correlation** is defined as the **statistical association** between **two** variables.

A **correlation** exists between **two variables** when one of them is **related** to the other in some way.

## What Does Correlation Measure?

- Correlation studies and measures the direction and extent of relationship among variables.
- So, the correlation measures co-variation, not causation.
- Therefore, we should never interpret correlation as implying cause and effect relation.
- For example, there exists a correlation between two variables X and Y, which means the value of one variable is found to change in one direction, the value of the other variable is found to change either in the same direction (i.e. positive change) or in the opposite direction (i.e. negative change).

A **scatterplot** (or scatter diagram) is a graph of the pairs $(x, y)$ sample data along with the way in which these two variables relate to each other. The values of variable $\boldsymbol{X}$ are given along the horizontal axis, with the values of the variable $\boldsymbol{Y}$ given on the vertical axis.

A scatterplot can identify several different types of relationships between two variables.

- A relationship has **no correlation** when the points on a scatterplot do not show any pattern.
- A relationship is **non-linear** when the points on a scatterplot follow a pattern but not a straight line.
- A relationship is **linear** when the points on a scatterplot follow a somewhat straight-line pattern. This is the relationship that we will examine.

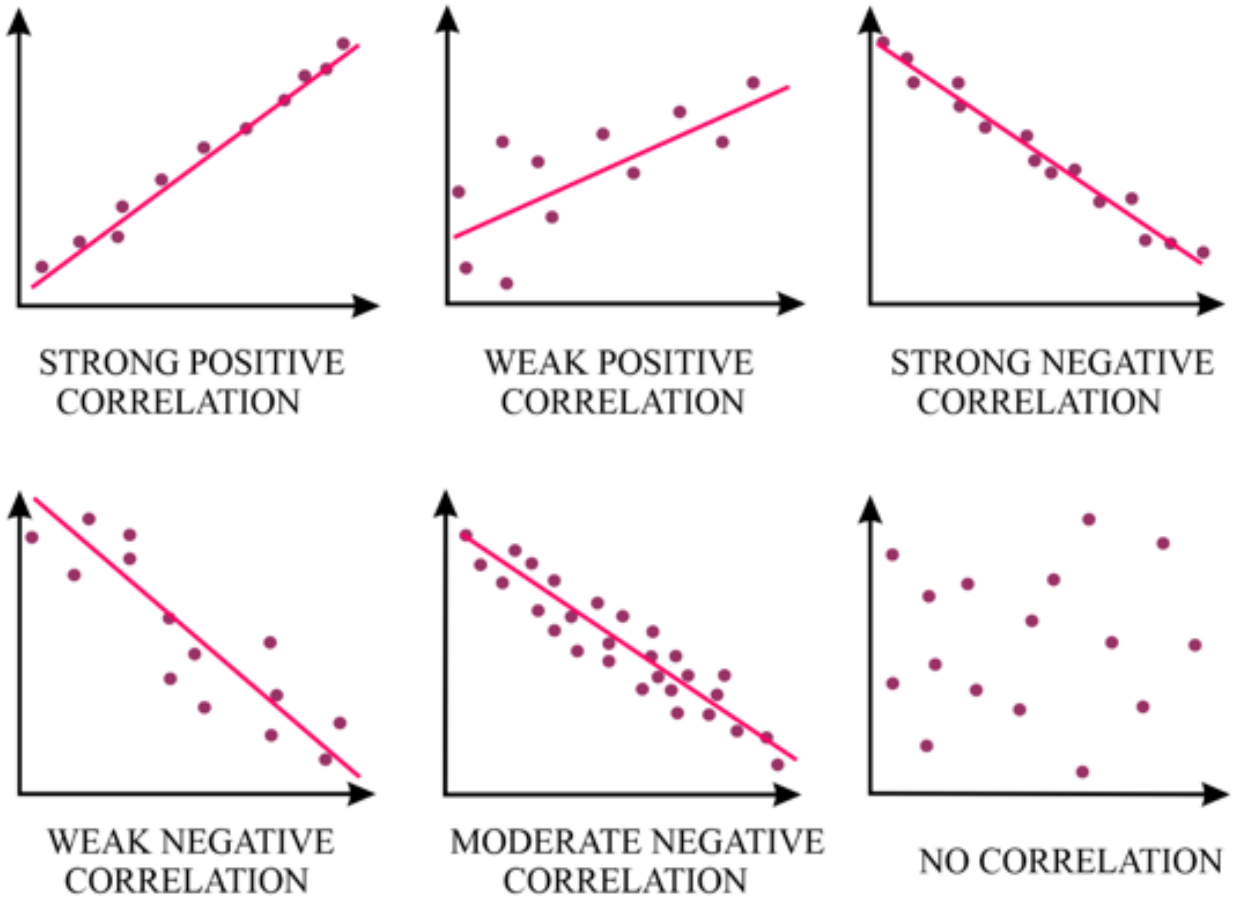
# Types of Correlation

The scatter plot explains the correlation between the two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables –

- **Positive Correlation:** when the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable.

- **Negative Correlation:** when the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable.

- **No Correlation:** when there is no linear dependence or no relation between the two variables.



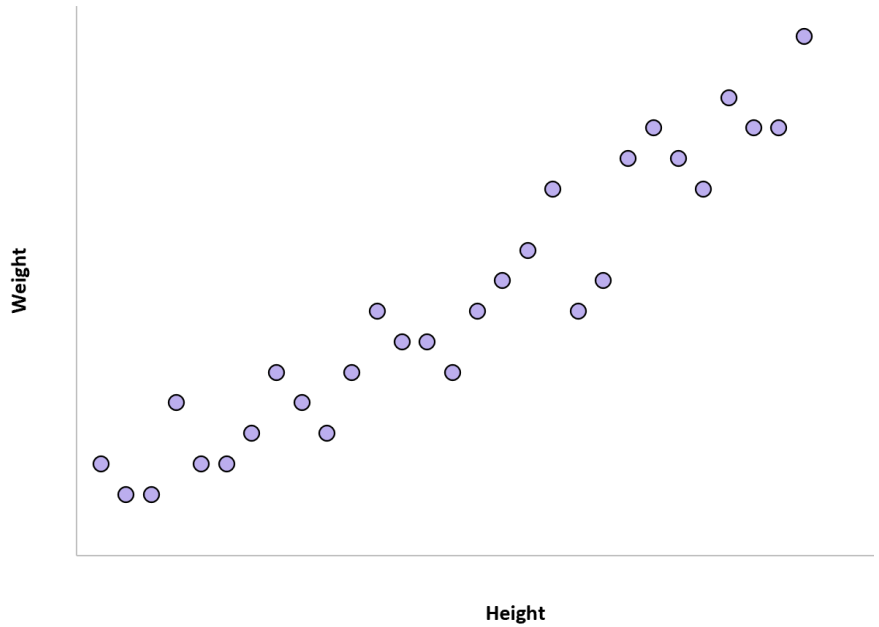
# Example: Positive Correlation

## Height vs. Weight

The correlation between the height of an individual and their weight tends to be positive.

In other words, individuals who are taller also tend to weigh more.

If we created a scatterplot of height vs. weight, it may look something like this:

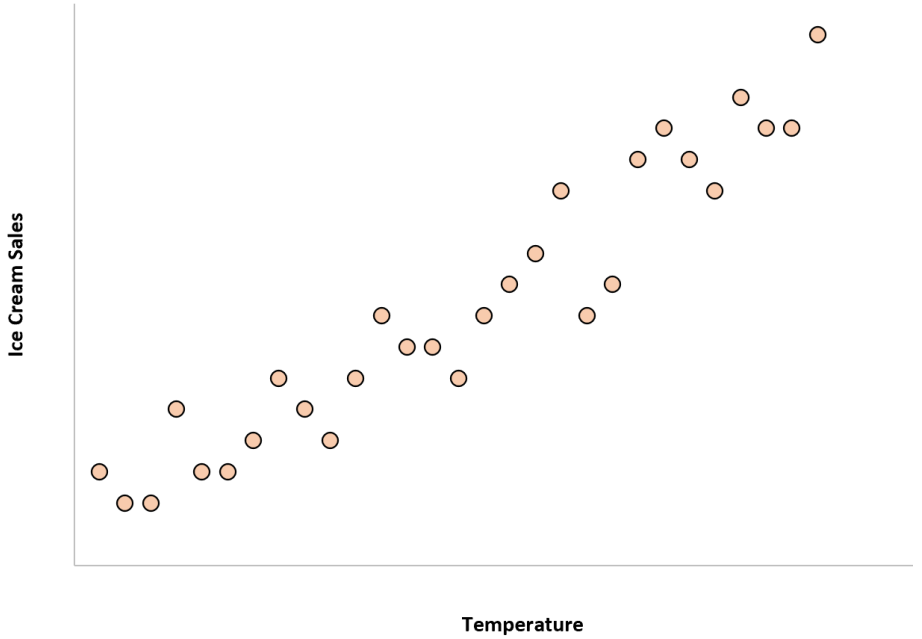



## Temperature vs. Ice Cream Sales

The correlation between the temperature and total ice cream sales is positive.

In other words, when it's hotter outside the total ice cream sales of companies tends to be higher since more people buy ice cream when it's hot out.

If we created a scatterplot of temperature vs. ice cream sales, it may look something like this:



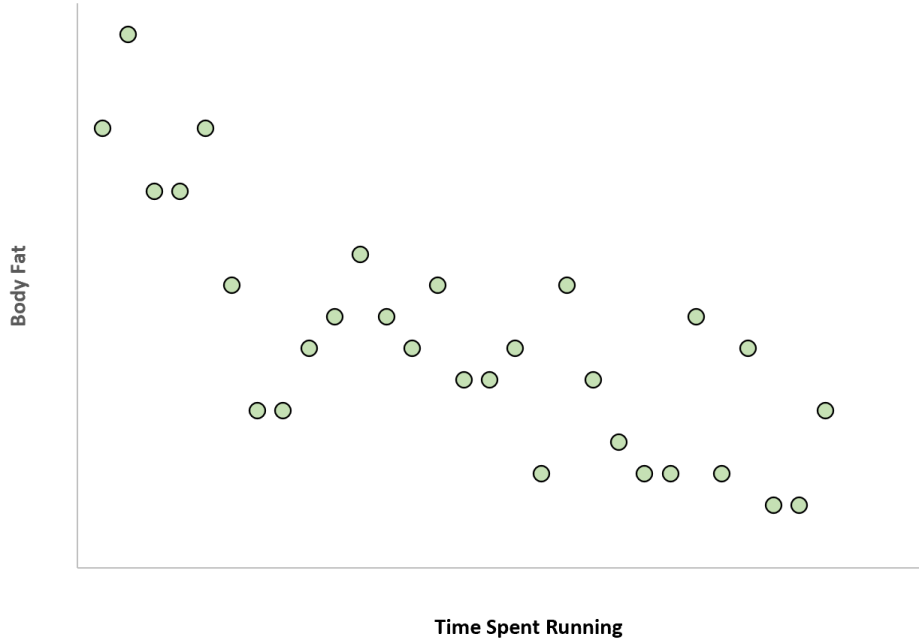
# Example: Negative Correlation

## Time Spent Running vs. Body Fat

The more time an individual spends running, the lower their body fat tends to be.

In other words, the variable running time and the variable body fat have a negative correlation.

As time spent running increases, body fat decreases. If we created a scatterplot of time spent running vs. body fat, it may look something like this:

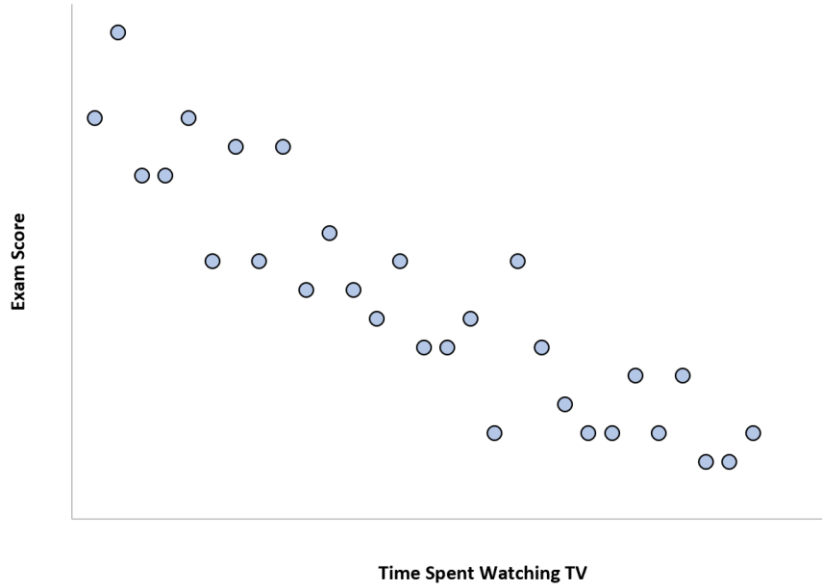



## Time Spent Watching TV vs. Exam Scores

The more time a student spends watching TV, the lower their exam scores tend to be.

In other words, the variable time spent watching TV and the variable exam score have a negative correlation.

As time spent watching TV increases, exam scores decrease. If we created a scatterplot of time spent watching TV vs. exam scores, it may look something like this:



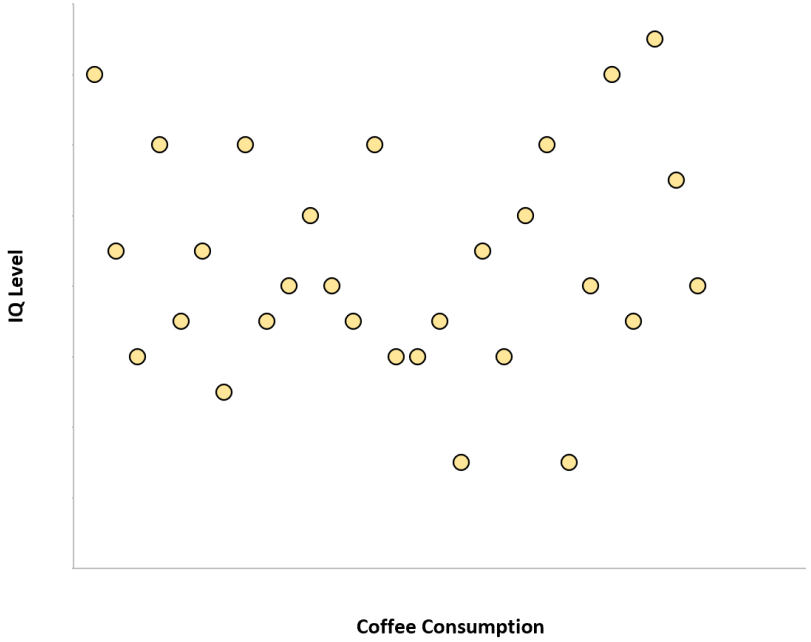
# Example: No Correlation

## Coffee Consumption vs. Intelligence

The amount of coffee that individuals consume and their IQ level has a correlation of zero.

In other words, knowing how much coffee an individual drinks doesn't give us an idea of what their IQ level might be.

If we created a scatterplot of daily coffee consumption vs. IQ level, it may look something like this:

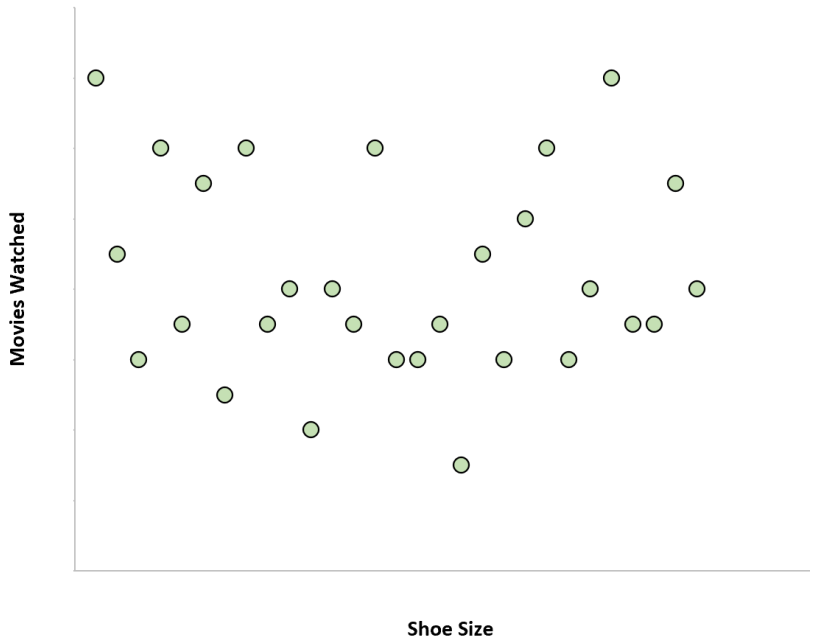



## Shoe Size vs. Movies Watched

The shoe size of individuals and the number of movies they watch per year has a correlation of zero.

In other words, knowing the shoe size of an individual doesn't give us an idea of how many movies they watch per year.

If we created a scatterplot of shoe size vs. number of movies watched, it may look like this:

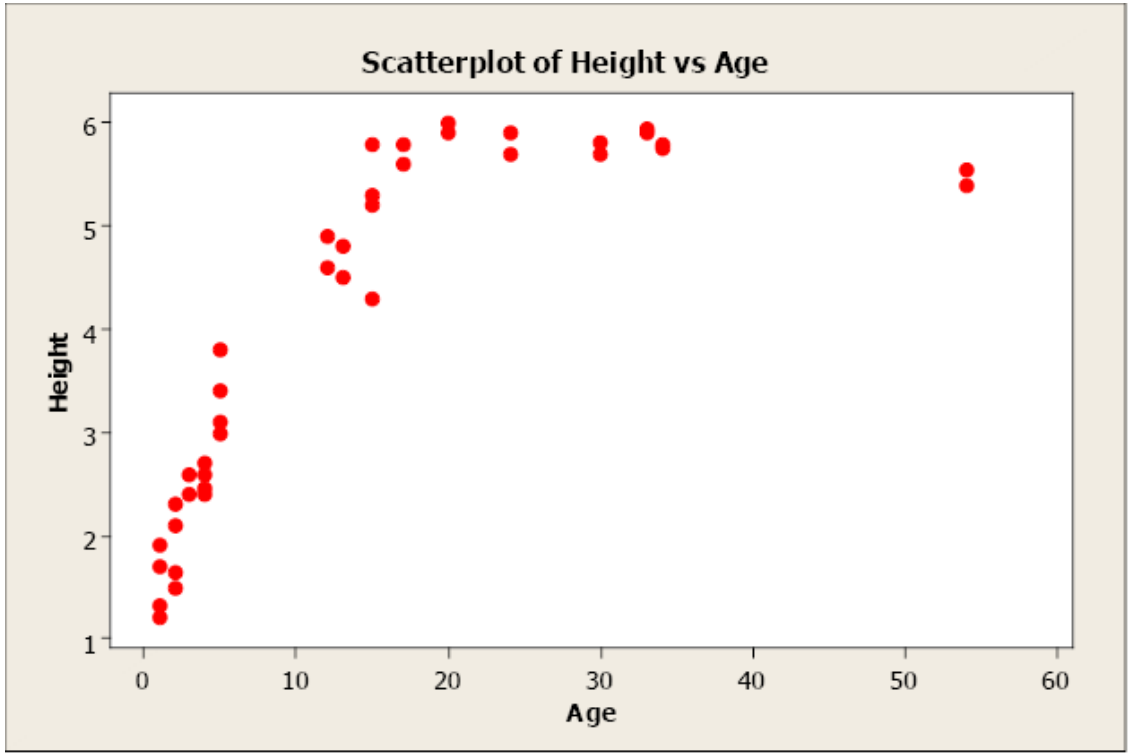


# Example: Non-linear Correlation

## Height vs. Age

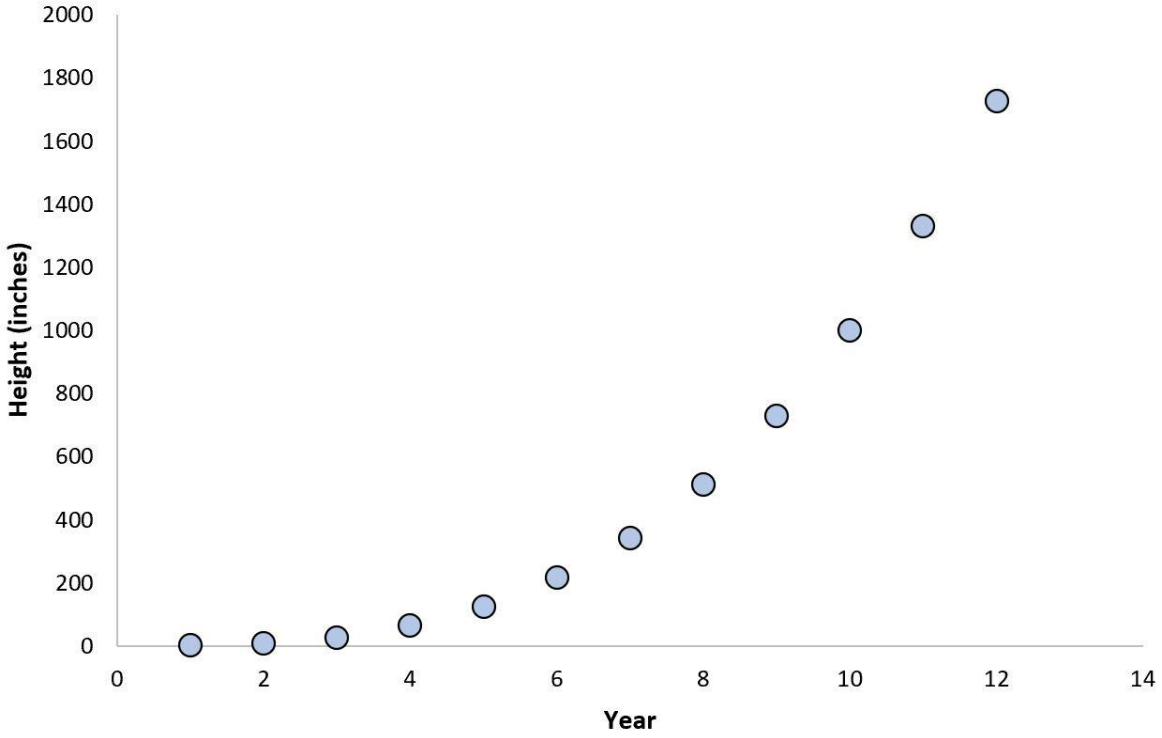
Non-linear relationships have an apparent pattern, just not linear.

For example, as age increases height increases up to a point then levels off after reaching a maximum height.




## Lifespan of Bamboo Tree

During the first few years of growth, a bamboo plant grows very slowly but once it reaches a certain age it explodes in height and grows at a rapid pace.



# Linear Correlation Coefficient

Because visual examinations are largely subjective, we need a more precise and objective measure to define the correlation between the two variables.

To quantify the strength and direction of the relationship between two variables, we use the linear correlation coefficient:

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

where

$$S_{xx} = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

$$S_{yy} = \frac{\sum(Y_i - \bar{Y})^2}{n-1}$$

$$S_{xy} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{n-1}$$

**The properties of "*r*":**

- It is always between -1 and +1.
- It is a unitless measure so "r" would be the same value whether you measured the two variables in pounds and inches or in grams and centimeters.
- Positive values of "r" are associated with positive relationships.
- Negative values of "r" are associated with negative relationships

This statistic numerically describes how strong the straight-line or linear relationship is between the two variables and the direction, positive or negative.

# Example

A statistics instructor at a large western university would like to examine the relationship between the number of optional homework problems students do during the semester and their final course grade. A sample of 12 students selected randomly for the study. The number of these problems completed during the semester and their final grade is given below. Find the correlation coefficient between the course grade and the number of optional homework problems

| Problems Course x | 51 | 58 | 62 | 65 | 68 | 76 | 77 | 78 | 78 | 84 | 85 | 91 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade y | 62 | 68 | 66 | 66 | 67 | 72 | 73 | 72 | 78 | 73 | 76 | 75 |

Solution:

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | n =12.00 |
| X | 51.00 | 58.00 | 62.00 | 65.00 | 68.00 | 76.00 | 77.00 | 78.00 | 78.00 | 84.00 | 85.00 | 91.00 | | | =72.75 |
| Y | 62.00 | 68.00 | 66.00 | 66.00 | 67.00 | 72.00 | 73.00 | 72.00 | 78.00 | 73.00 | 76.00 | 75.00 | | | =70.67 |
| | | | | | | | | | | | | | | | |
| $(X_i - \bar{X})^2$ | 473.06 | 217.56 | 115.56 | 60.06 | 22.56 | 10.56 | 18.06 | 27.56 | 27.56 | 126.56 | 150.06 | 333.06 | | | =143.84 |
| $(Y_i - \bar{Y})^2$ | 75.11 | 7.11 | 21.78 | 21.78 | 13.44 | 1.78 | 5.44 | 1.78 | 53.78 | 5.44 | 28.44 | 18.78 | | | =23.15 |
| $X_i \times Y_i$ | 3162.00 | 3944.00 | 4092.00 | 4290.00 | 4556.00 | 5472.00 | 5621.00 | 5616.00 | 6084.00 | 6132.00 | 6460.00 | 6825.00 | | | =51.09 |

So, $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.881$

The scatter plot

# Example

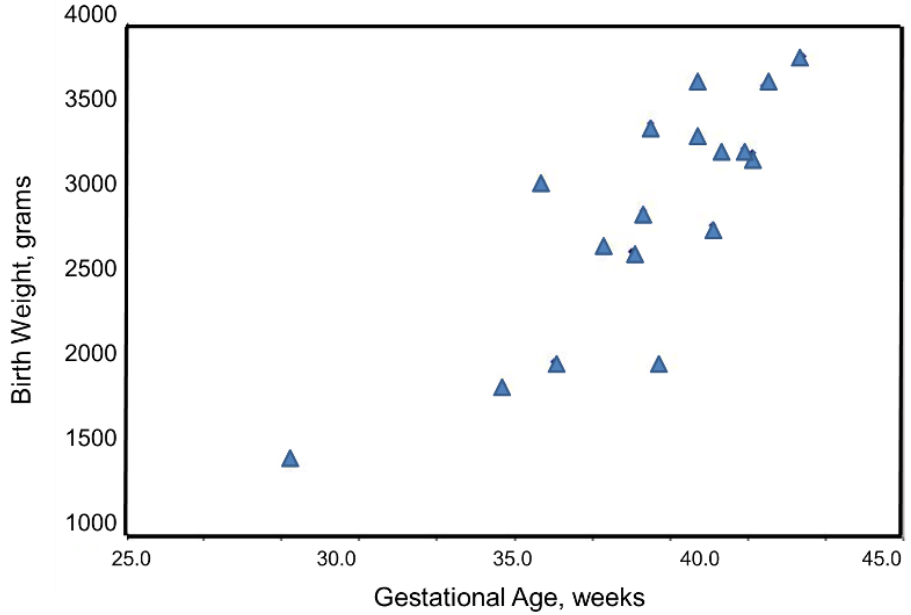
A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams. We wish to estimate the association between gestational age and infant birth weight. In this example, birth weight is the dependent variable and gestational age is the independent variable. Thus y=birth weight and x=gestational age.

Solution: $S_{xx} = 10.0$

$S_{yy} = 485478.8$

$S_{xy} = 1798.0$

So, $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.82$




| Infant ID # | Gestational Age (weeks) | Birth Weight (grams) |
|---|---|---|
| 1 | 34.7 | 1895 |
| 2 | 36.0 | 2030 |
| 3 | 29.3 | 1440 |
| 4 | 40.1 | 2835 |
| 5 | 35.7 | 3090 |
| 6 | 42.4 | 3827 |
| 7 | 40.3 | 3260 |
| 8 | 37.3 | 2690 |
| 9 | 40.9 | 3285 |
| 10 | 38.3 | 2920 |
| 11 | 38.5 | 3430 |
| 12 | 41.4 | 3657 |
| 13 | 39.7 | 3685 |
| 14 | 39.7 | 3345 |
| 15 | 41.1 | 3260 |
| 16 | 38.0 | 2680 |
| 17 | 38.7 | 2005 |

# Simple Linear Regression

- Once we have identified two variables that are **correlated**, we would like to **model this relationship**.
- We want to use **one variable** as a **predictor** or explanatory variable to explain the other variable, the response or dependent variable.
- In order to do this, we need a good relationship between our two variables.
- The model can then be used to predict changes in our response variable.
- A strong relationship between the predictor variable and the response variable leads to a good model.

**"A simple linear regression model is a mathematical equation that allows us to predict a response for a given predictor value."**

The formula for simple linear regression is:

$$y_i = \alpha + \beta x_i + e_i, \qquad i = 1, 2, 3, \ldots, n$$

- $y_i$ is the predicted value of the dependent variable $y_i$ for any given value of independent variable $x_i$.
- $x_i$ is the independent variable (the variable we expect is influencing $y_i$).
- $\alpha$ is the constant or intercept, the predicted value of $y_i$ when the $x_i$ is $0$.
- β is the regression coefficient-how much we expect $y_i$ to change as $x_i$ increases.
- $e_i$ is the error of the estimate, or how much variation there is in our estimate of the regression coefficient?

# Least Square Method

The least square method is used to obtain the values of $\alpha$ and $\beta$. The equation to find the best fitting line is:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}x$$

where

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

# Example

For the last example, find the regression line of the gestational Age on the birth weight.

Solution:

$$\hat{Y} = \hat{\alpha} + \hat{\beta} x$$

Given from the example $\bar{x} = 38.4$ , $\bar{y} = 2902$ , $S_{xx} = 10$ , $S_{yy} = 485478.8$ , $Sxy = 1798.0$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}}$$

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

$$= \frac{1798}{10}$$

$$= 179.8$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$= 2902 - 179.8 \times 38.4$$

$$= -4002.32$$

$$\hat{Y} = -4002.3 + 179.8\, x$$

# Example

For the other example above, find the regression line of the course grade and the number of optional homework problems.

Solution:

$$\hat{Y} = \hat{\alpha} + \hat{\beta} x$$

Given from the example $\bar{x} = 72.75$ , $\bar{y} = 70.67$ , $S_{xx} = 143.84$ , $S_{yy} = 23.15$, $Sxy = 50.82$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}}$$

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

$$= 0.353$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$= 70.67 - 0.353 \times 72.75$$

$$= 44.96$$

$$\hat{Y} = 44.69 + 0.353\ x$$

# Exercises

1. Given the following data:

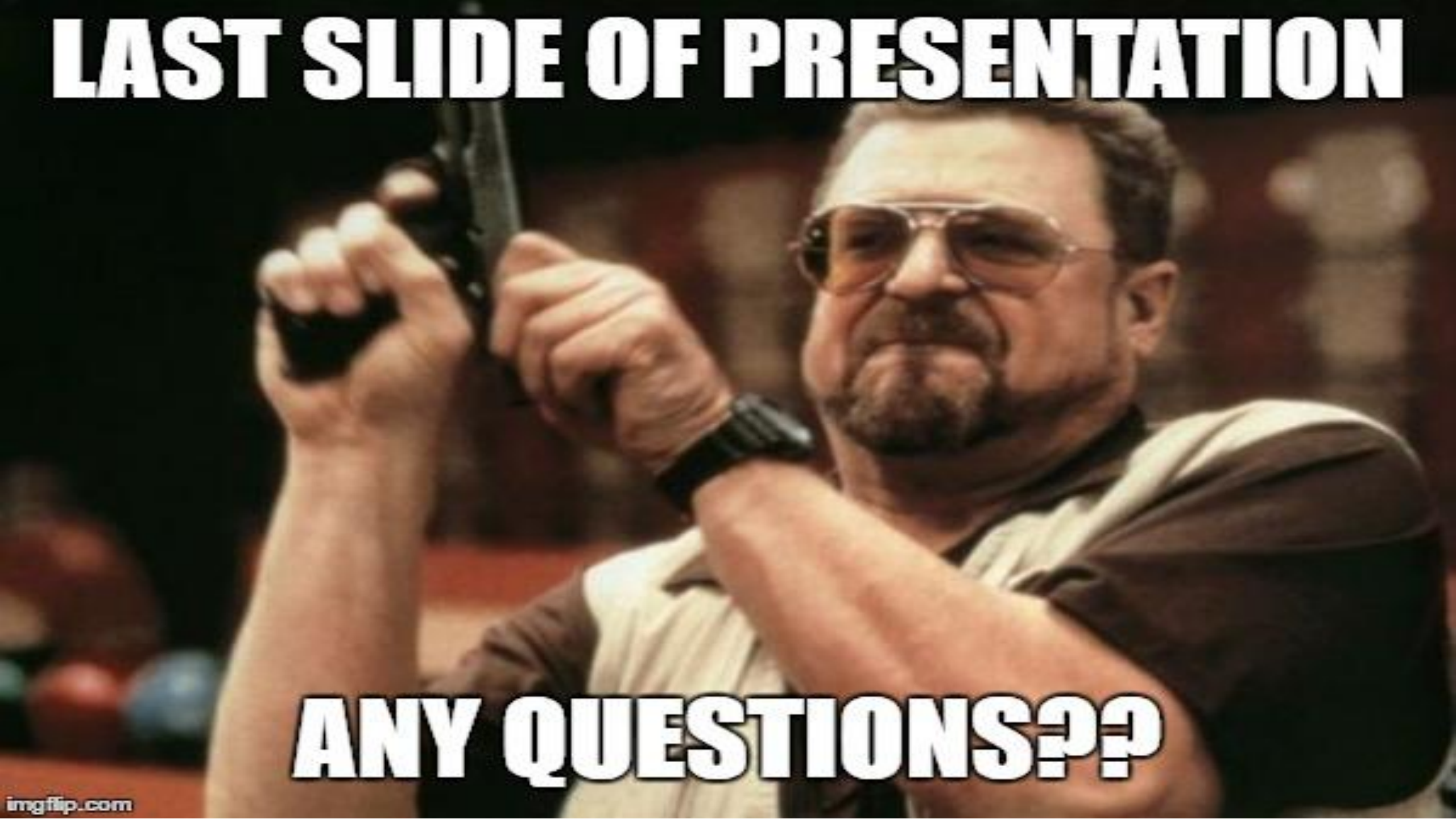
| x | 1 | 3 | 4 | 6 | 8 | 9 | 11 | 14 |
|---|---|---|---|---|---|---|---|---|
| y | 1 | 2 | 4 | 4 | 5 | 7 | 8 | 9 |

a. Plot the scatter diagram.
b. Find the correlation coefficient between x and y.
c. Find the regression equation of y on x.
d. What is the predicted value of y when x=10?

2. The following data represent the level of dose and the blood pressure of five Patients:

| Dose level | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Blood pressure | 278 | 240 | 198 | 132 | 111 |

a. Plot the scatter diagram.
b. Find the correlation coefficient between the dose level and the blood pressure.
c. Find the regression equation of the blood pressure on the dose level.
d. What is the predicted value of the blood pressure when the dose level is 8?


LAST SLIDE OF PRESENTATION
ANY QUESTIONS??
imgflip.com

# Thank you!

End of slides

# References


- Introduction to Probability, Statistics, and Random Processes
  Textbook by Hossein Pishro-Nik

- A Modern Introduction to Probability and Statistics: Understanding Why and How
  Book by Frederik Michel Dekking

- D. C. Montgomery and G.C. Runger, “Applied Statistics and Probability for Engineers”, 5th edition, John Wiley & Sons, (2009).Online lecture notes,