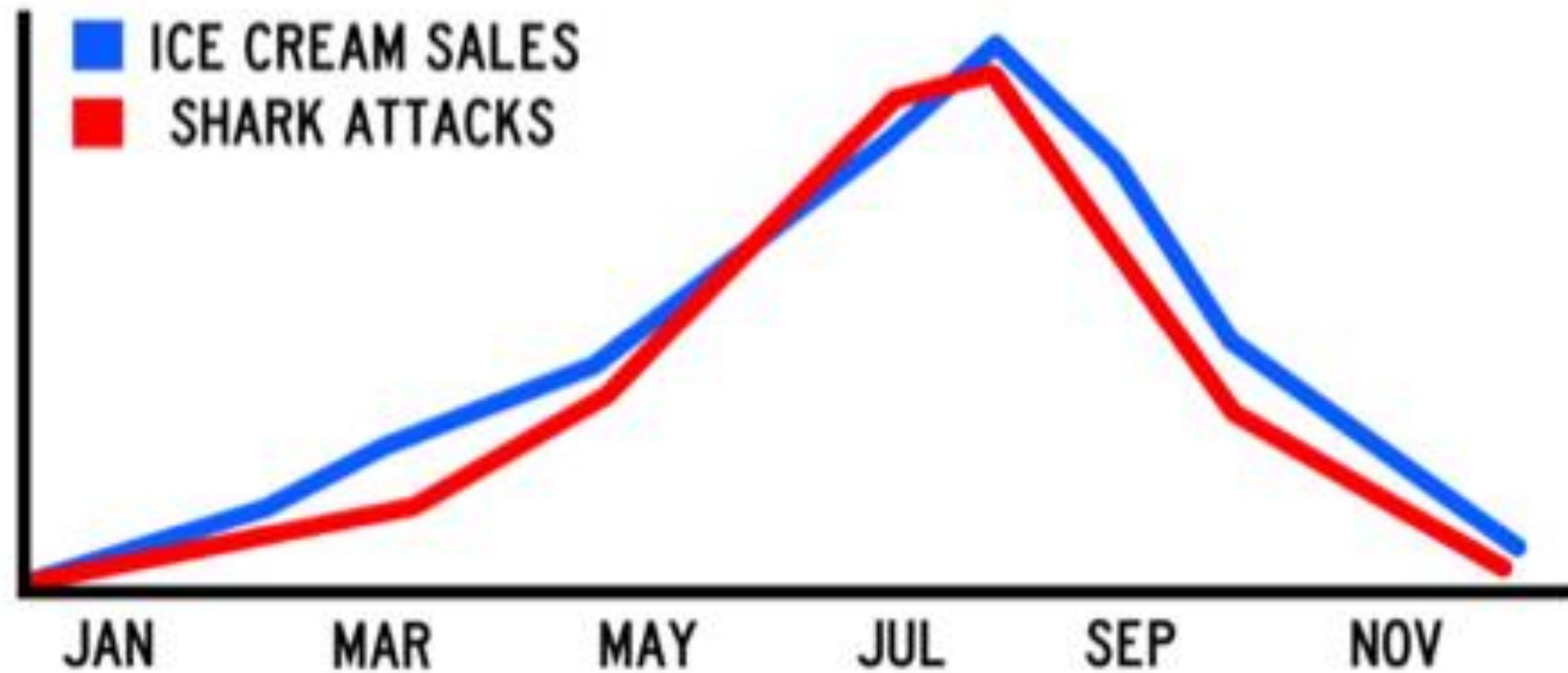


CORRELATION IS NOT CAUSATION!

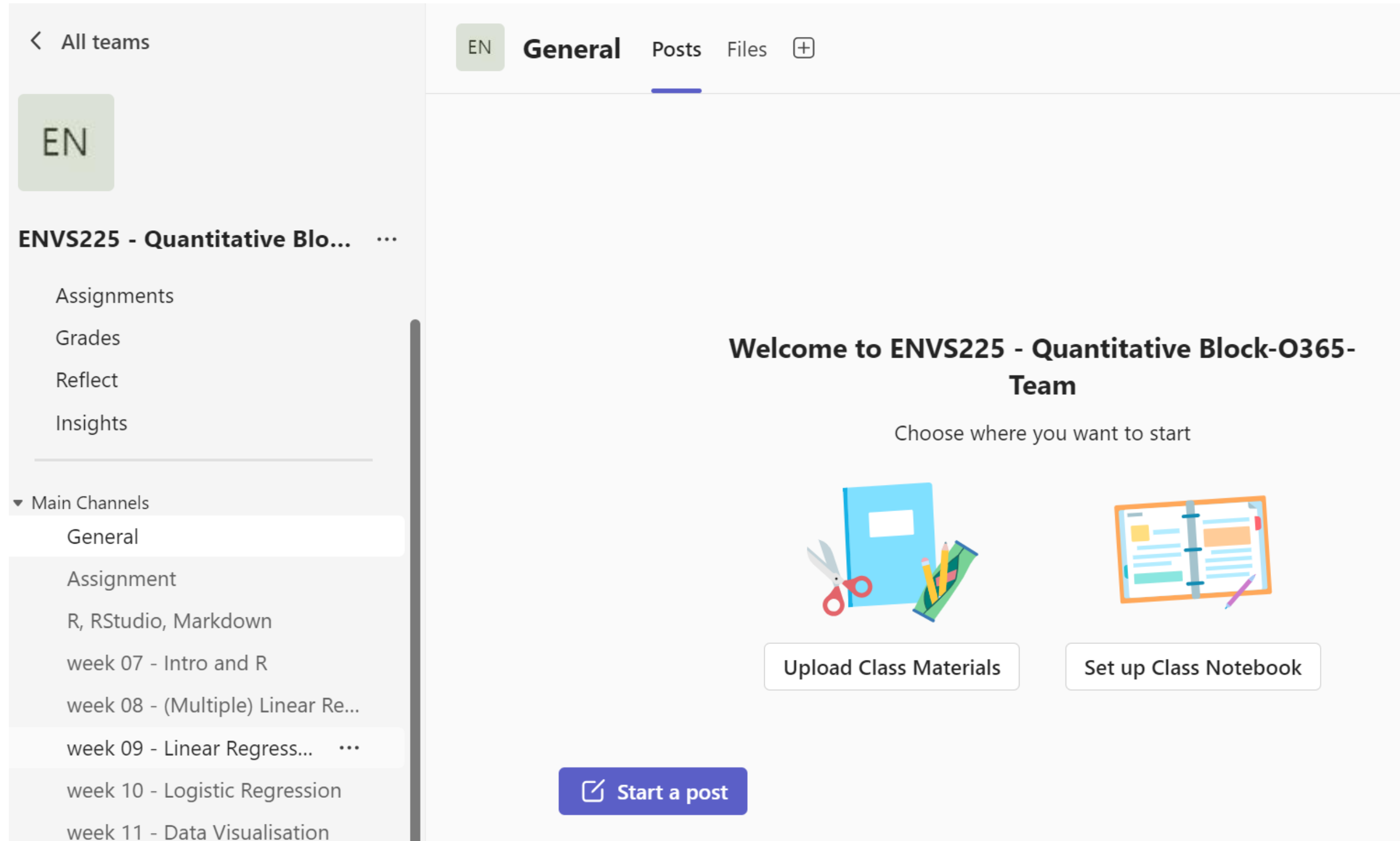


Intro & Review

ENVS225 Exploring the Social World

Gabriele Filomena
gfilo@liverpool.ac.uk

Quantitative Data Block: Questions -> MS Teams



The screenshot shows a Microsoft Teams interface. On the left is a sidebar with a back arrow and 'All teams'. Below that is a green tab labeled 'EN'. The main channel is 'ENVS225 - Quantitative Blo...' with a three-dot menu. Underneath are links for 'Assignments', 'Grades', 'Reflect', and 'Insights'. A 'Main Channels' section is expanded, showing a list of channels: 'General' (highlighted), 'Assignment', 'R, RStudio, Markdown', 'week 07 - Intro and R', 'week 08 - (Multiple) Linear Re...', 'week 09 - Linear Regress...' (with a three-dot menu), 'week 10 - Logistic Regression', and 'week 11 - Data Visualisation'. The main area of the channel has a header with 'EN', 'General' (selected), 'Posts', and 'Files' (with a plus icon). The content area displays a welcome message: 'Welcome to ENVS225 - Quantitative Block-O365-Team' followed by 'Choose where you want to start'. There are two large buttons: 'Upload Class Materials' (with an icon of a blue folder, scissors, and pencils) and 'Set up Class Notebook' (with an icon of a notebook and pencil). At the bottom left is a blue button labeled 'Start a post' with a speech bubble icon.



[Join the Ms Teams Class](#)

Quantitative Data Block: Webpage

📖

📄

⬇️

🔄

📅

🔍

🗑️

🔄

📄

✅

W

+

Exploring the Social World - Quantitative Block: Statistics

🔍

Welcome

Overview

Assessment

1 Lab: Introduction to R for Statistics

2 Lab: Single and Multiple Linear Regression

🔒

🌐

🔒

🌐

https://gdsl-ul.github.io/stats/

Table of contents

Welcome

Contact

🔄

 Edit this page

Exploring the Social World - Quantitative Block: Statistics

AUTHOR

Gabriele Filomena and Zi Ye

PUBLISHED

November 6, 2024

Welcome

This is the website for “Exploring the Social World - Quantitative Block: Statistics” (module **ENVS225**) at the University of Liverpool. This block of the module is designed and delivered by Dr. Gabriele Filomena and Dr. Zi Ye from the Geographic Data Science Lab at the University of Liverpool. The module seeks to provide hands-on experience and training in introductory statistics for human geographers.

The website is **free to use** and is licensed under the [Attribution-NonCommercial-NoDerivatives 4.0 International](#). A compilation of this web course is hosted as a GitHub repository that you can access:

- As an [html website](#).
- As a [GitHub repository](#).

Contact

Gabriele Filomena - gfilo [at] liverpool.ac.uk Lecturer in Geographic Data Science Office 1xx, Roxby Building, University of Liverpool - 74 Bedford St S, Liverpool, L69 7ZT, United Kingdom.

Zi Ye - zi.ye [at] liverpool.ac.uk Lecturer in Geographic Data Science Office 1xx, Roxby Building, University of Liverpool - 74 Bedford St S, Liverpool, L69 7ZT, United Kingdom.

Overview →

gdsl-ul.github.io/stats/

Material and Labs

Quantitative Data Block: **Assignment**

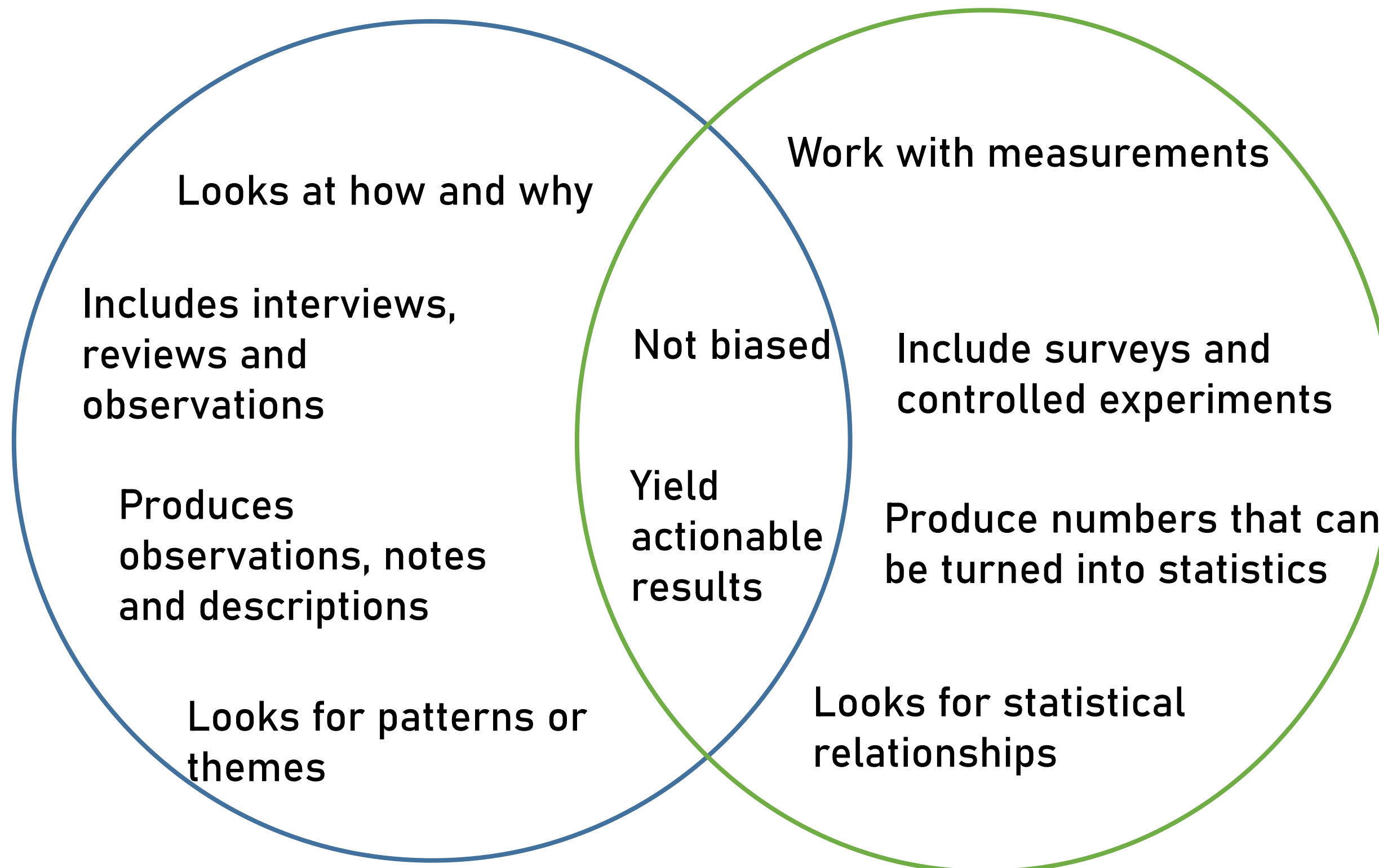
Writing a research report using one of the regression techniques learned during the module. The report should have the following sections:

- An introduction.
- A literature review.
- Methods and data.
- Results and Discussion.
- Conclusion.

See webpage. There will be a Template

Qualitative

Quantitative



ENVS225 Stats...?

- Ask questions
- Come to class & labs!

Average Grade students who came to class: 67.5

Average Grade students who missed classes: 46.75

- Keep up with practical sessions. Do not overestimate this module!
- Ask on Teams, stick to the instructions, do not make things harder for you

Learning goals

- Understand how to explore a dataset, containing a number of observations described by a set of variables.
- Demonstrate an understanding in the application and interpretation of commonly used quantitative research methods.
- Demonstrate an understanding of how to work with quantitative data to address real-world research questions.

Structure

1	Introduction & Review	GF
2	Single & Multiple Linear Regression	GF
3	Multiple Linear Regression with Categorical Variables	ZY
4	Logistic Regression	ZY
5	Data Visualisation	GF
6	Summary and Assessment Support	ZY

} + Support
Drop-in
sessions
(2h a week)

What are we going to work with?

nrKilled	nrWound	year	country	group	attack	target	weapon
0	0	2005	United Kingdom	Abu Hafs al-Masri Brigades	Bombing/Explosion	Transportation	Explosives/Bombs/Dynamite
0	0	2005	United Kingdom	Abu Hafs al-Masri Brigades	Bombing/Explosion	Transportation	Explosives/Bombs/Dynamite
0	0	2005	United Kingdom	Abu Hafs al-Masri Brigades	Bombing/Explosion	Transportation	Explosives/Bombs/Dynamite
0	0	2005	United Kingdom	Abu Hafs al-Masri Brigades	Bombing/Explosion	Transportation	Explosives/Bombs/Dynamite
0	1	1982	United Kingdom	Abu Nidal Organization (ANO)	Assassination	Government (Diplomatic)	Firearms
0	0	2014	United Kingdom	Anarchists	Facility/Infrastructure Attack	Business	Incendiary
0	0	2014	United Kingdom	Anarchists	Facility/Infrastructure Attack	Business	Incendiary
0	0	2014	United Kingdom	Anarchists	Facility/Infrastructure Attack	Business	Incendiary
0	0	2014	United Kingdom	Anarchists	Facility/Infrastructure Attack	Private Citizens & Property	Incendiary
0	0	2014	United Kingdom	Anarchists	Facility/Infrastructure Attack	Police	Incendiary
0	0	1984	United Kingdom	Animal Liberation Front (ALF)	Unarmed Assault	Private Citizens & Property	Chemical
0	0	1988	United Kingdom	Animal Liberation Front (ALF)	Facility/Infrastructure Attack	Business	Unknown
0	0	1988	United Kingdom	Animal Liberation Front (ALF)	Facility/Infrastructure Attack	Business	Unknown
0	0	1988	United Kingdom	Animal Liberation Front (ALF)	Facility/Infrastructure Attack	Business	Incendiary
0	0	1988	United Kingdom	Animal Liberation Front (ALF)	Facility/Infrastructure Attack	Business	Incendiary
0	0	1988	United Kingdom	Animal Liberation Front (ALF)	Facility/Infrastructure Attack	Business	Incendiary
0	0	1988	United Kingdom	Animal Liberation Front (ALF)	Facility/Infrastructure Attack	Business	Incendiary

Descriptive Statistics

Descriptive statistics

- Help us summarise the characteristics of our data/observations
- They form the initial step in a statistical analysis
- We get information about the centrality (e.g., mean, median)
- We get information about the variability (e.g., variance, standard deviation)

Mean

- *Mean* is another word for average.
- Most commonly used statistic that tells us about the centre of a data set.
- It is the average of the numbers in your dataset.

The sum of all values divided by the number of observations.

Median

- It is the middle point of your ordered data.
- If one orders the observations:

1,3,5,8,10

The median is the number in the middle

- When the observations are even:

1,3,5,6,8,10

Average the mid values -> 5.5

Difference between Mean and Median

- Both are single values that represent the center of a data set.
- But median is more resilient to **outliers**.

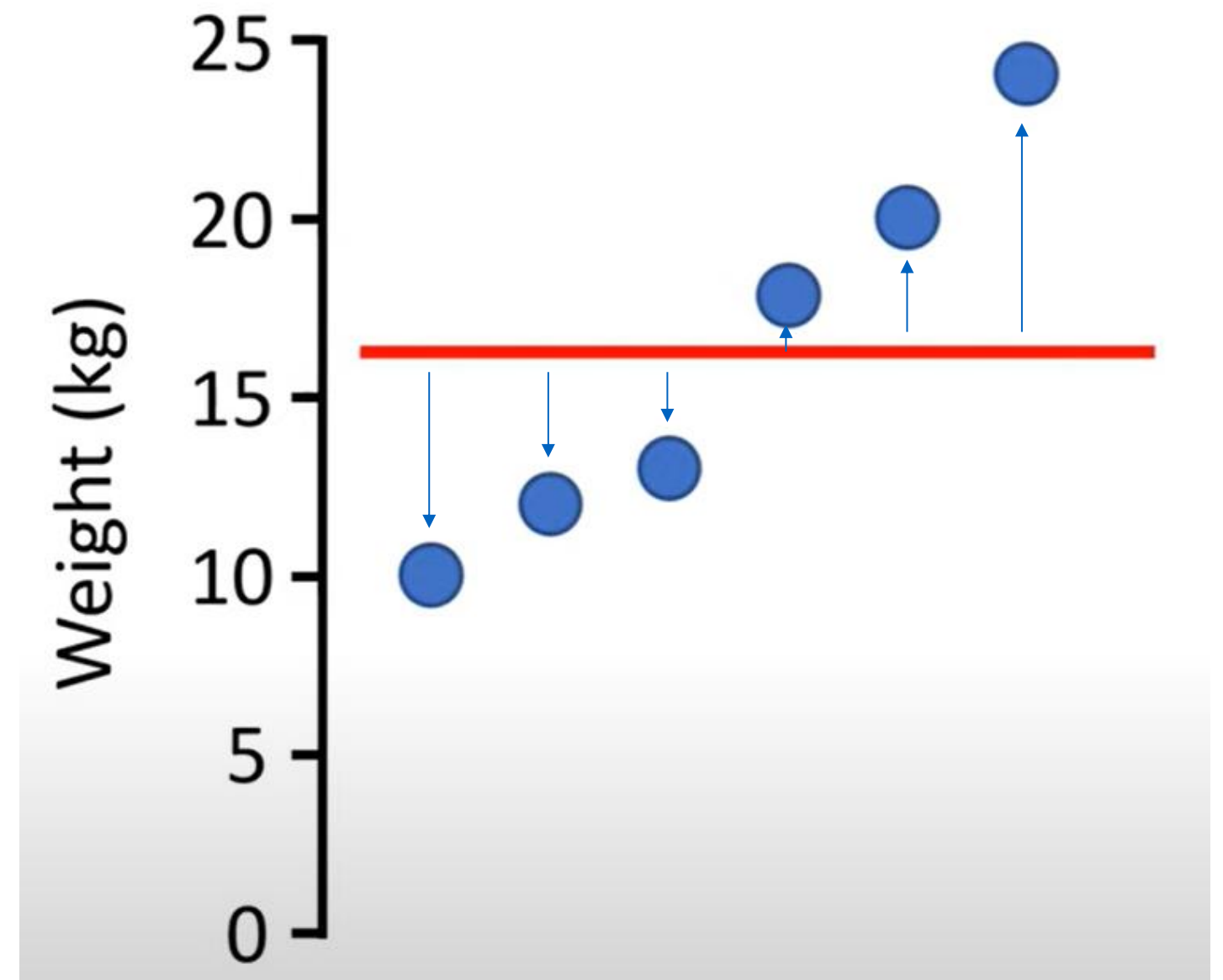
E.g.

3, 1, 4, 2.5, 50 (n=5)

- *Mean:* $\frac{3+1+4+2.5+50}{5}=12.1$
- *Median:* 1, 2.5, 3, 4, 50 =3

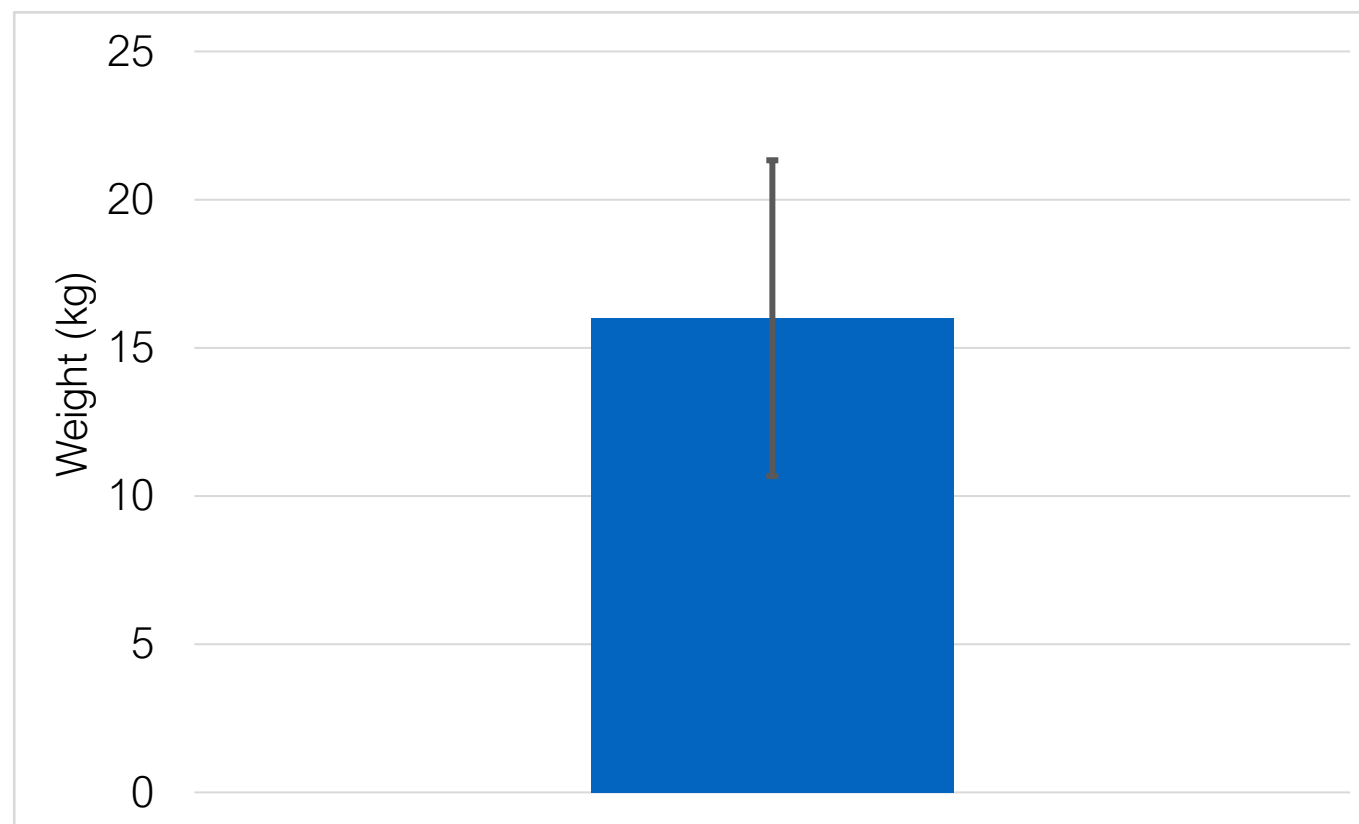
Variance (s^2)

- A measure to show how spread out our observations are
- Based on the distance of each observation from the mean
- The average of the squared differences from the mean.



Standard deviation (s)

- The square root of *variance*
- The result is in the same measurement units as our data (e.g., kg).
- This makes it easier to interpret and understand the spread of the data than *variance* alone.



- *Small value indicates data are gathered close to the mean*
- *Large value indicates data are gathered far from the mean*

Activity:

Form groups of 5/6.

Find a dataset in an excel format (.csv, .xls)

Download, Open and Explore the dataset in Ms Excel.

Be prepared to:

- describe the dataset, two variables. One qual, one quant.
- Find the mean and the standard deviation of the quant variable.

Associations between Variables

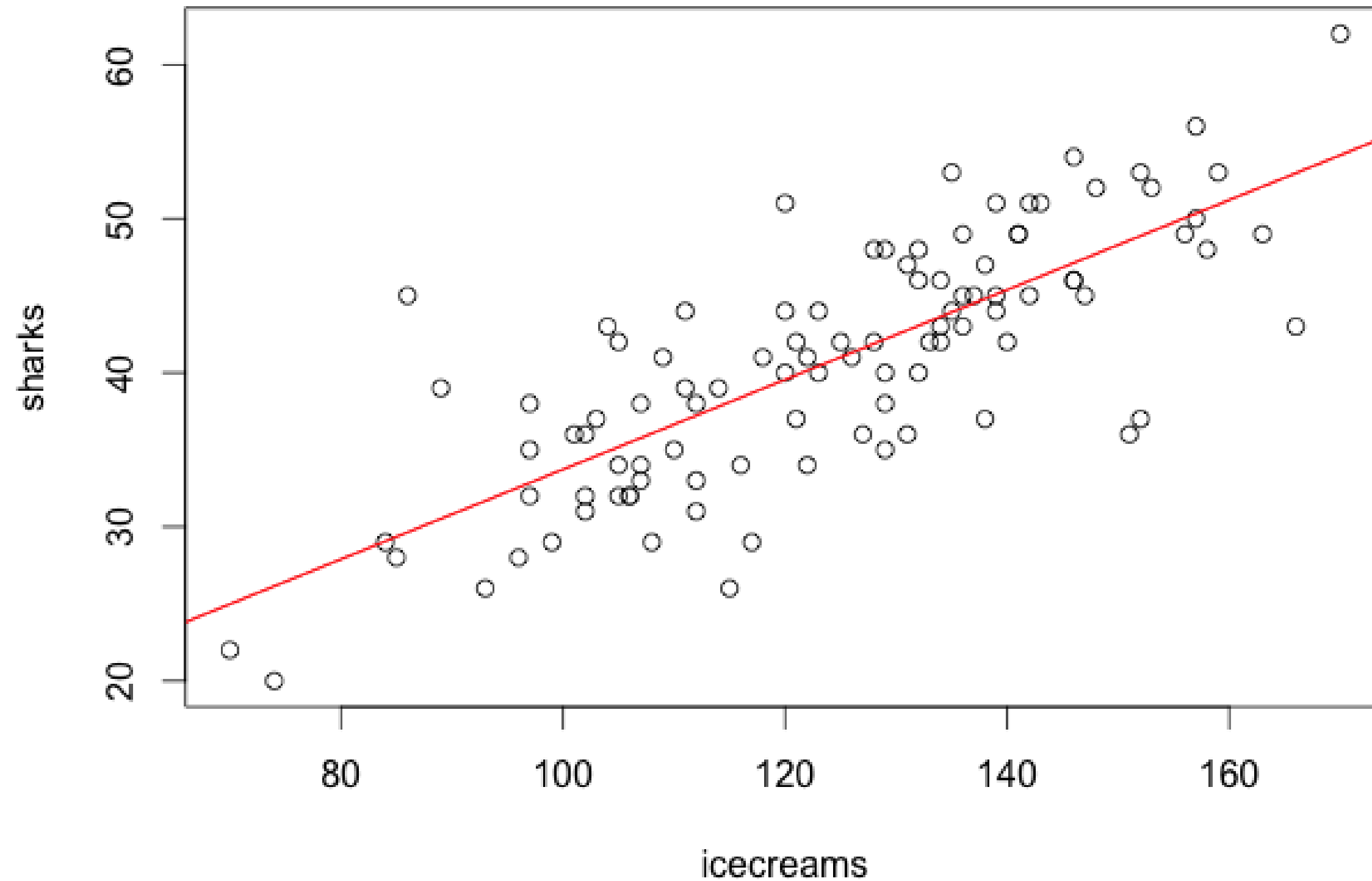
Correlation

- Describes an association between variables.
- When one variable changes, so does the other.
- Covariation, rather than a direct causal link.

Examples:

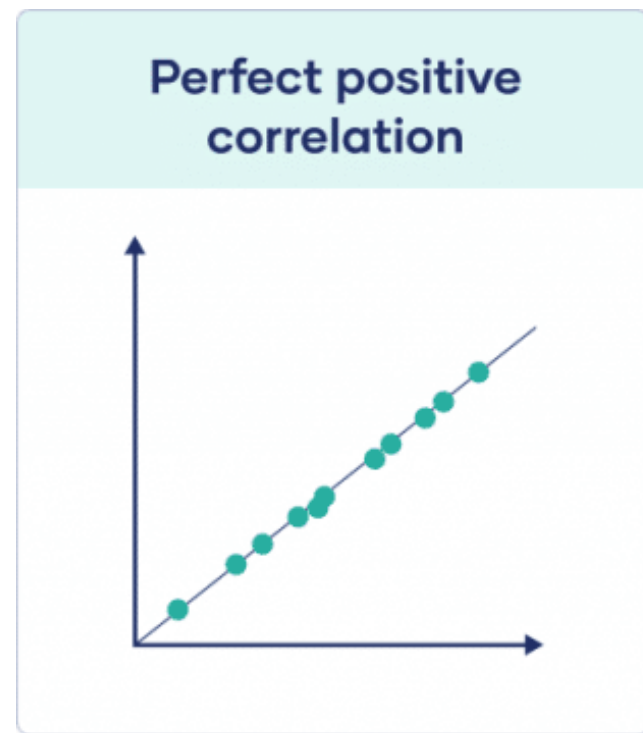
- Ice cream sales and shark attacks.
- Time spent in shops and temperature.

What is Correlation?



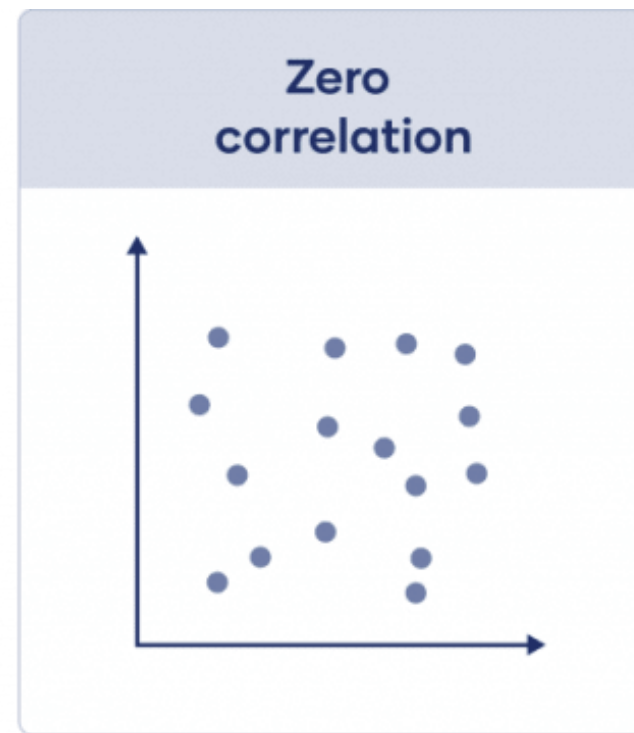
Pearson Correlation

The Pearson correlation coefficient (r) measures the strength and direction of the relationship between **Continuous** variables.

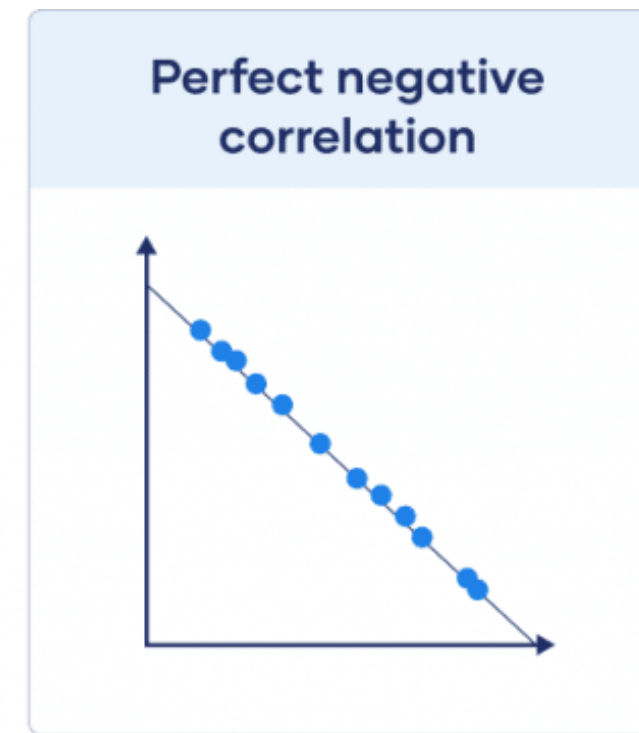


$$r = 1$$

Same direction



$$r = 0$$



$$r = -1$$

*Opposite
direction*

Single Linear Regression

- Regression is used to estimate the relationship between two variables.
- How a change in one variable (e.g., someone's exercise habits) can predict the outcome of another variable (e.g., general health).

DEPENDENT ◀ INDEPENDENT variables

=

RESPONSE ◀ EXPLANATORY

Examples

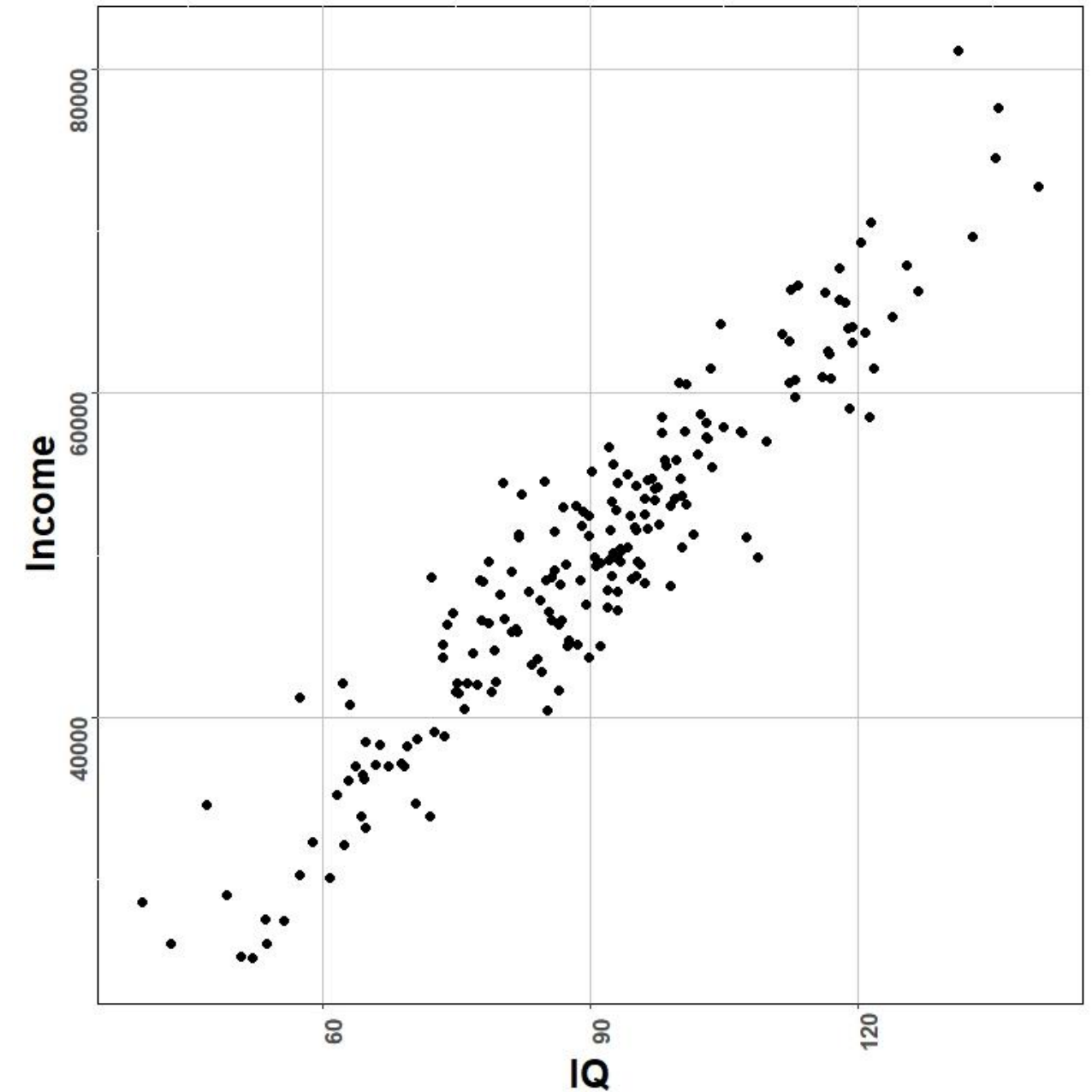
- IQ affects income (IQ is the [REDACTED] and income is the [REDACTED])?
- Study time affects grades (hours of study time is the [REDACTED] and average grade is the [REDACTED])?
- Exercise affects blood pressure (hours of exercise is the [REDACTED] and blood pressure is the [REDACTED])?

Examples

- IQ affects income (IQ is the **INDEPENDENT** and income is the **DEPENDENT**)?
- Study time affects grades (hours of study time is the **INDEPENDENT** and average grade is the **DEPENDENT**)?
- Exercise affects blood pressure (hours of exercise is the **INDEPENDENT** and blood pressure is the **DEPENDENT**)?

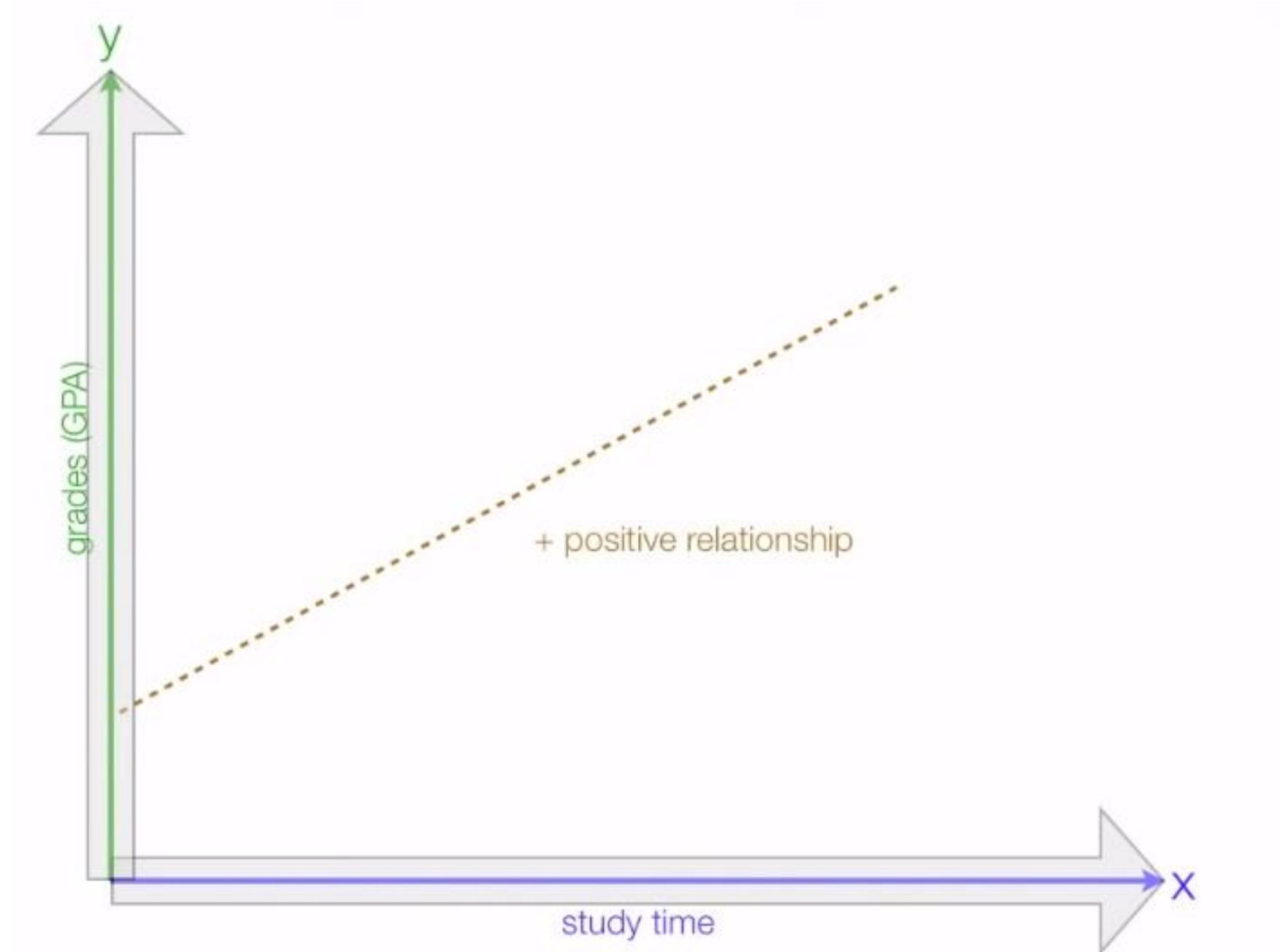
Displaying the data

When both the **DEPENDENT** and **INDEPENDENT** are numerical, we can represent data in the form of a scatterplot.



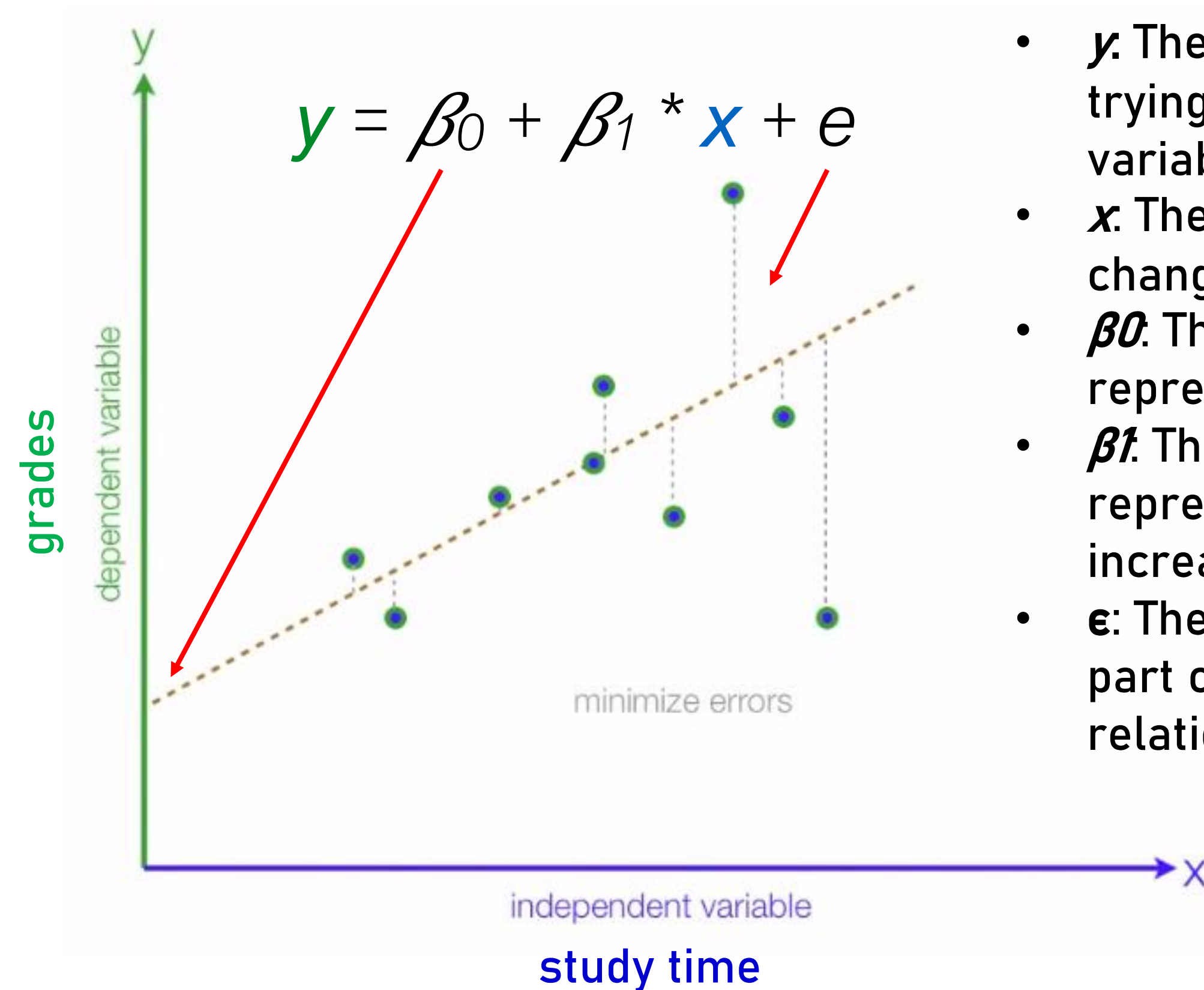
Single Linear Regression

- Straight line prediction model.
- As an **INDEPENDENT** variable changes, what happens to the **DEPENDENT** variable?

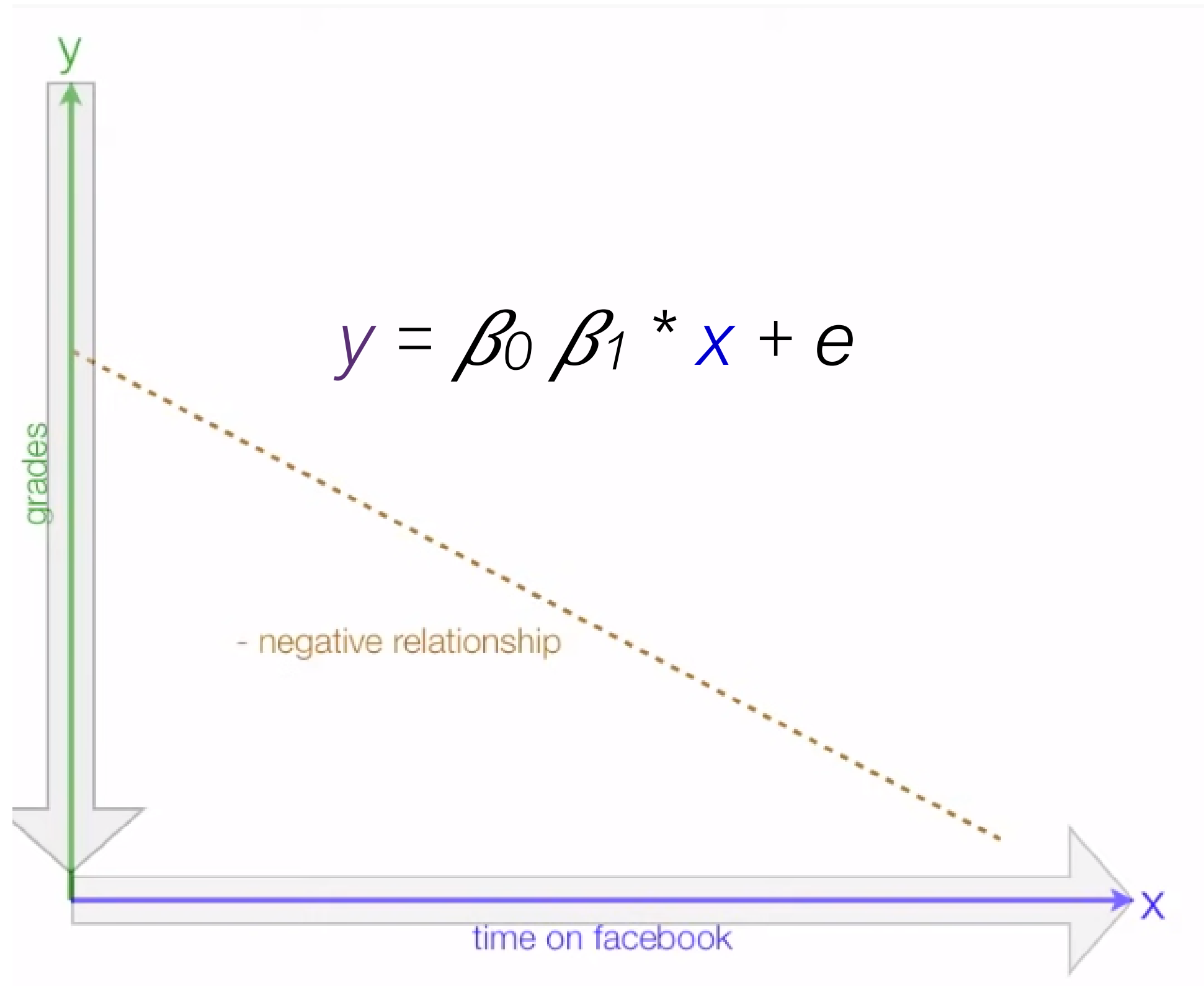


Correlation vs Single Linear Regression

- Correlation quantifies the strength and direction of the linear relationship between two variables but doesn't give information about the equation of the line.
- Single Linear Regression goes further by providing an equation that describes the linear relationship.
- Single Linear Regression models the relationship between two variables.



- y : The dependent variable. The outcome we are trying to explain based on the independent variable.
- x : The independent variable used to explain changes in the dependent variable.
- β_0 : The intercept of the regression line. It represents the expected value of y when x is 0.
- β_1 : The slope of the regression line. It represents the change in y for a one-unit increase in x .
- ϵ : The error term (or residual). It represents the part of y that cannot be explained by the linear relationship with x .



Week 7 Task

- Read the paper [*Defining Research Questions*](#) on Canvas.
- Complete today's practice.
- Familiarise yourself with RStudio.