Introduction to Data Science (11372 & G 11516)

Semester 1 2021

UNIVERSITY OF CANBERRA

INTRODUCTION TO DATA SCIENCE

Lecture 11

Dr. Ibrahim Radwan

DISTINCTIVE BY DESIGN

OUTLINE

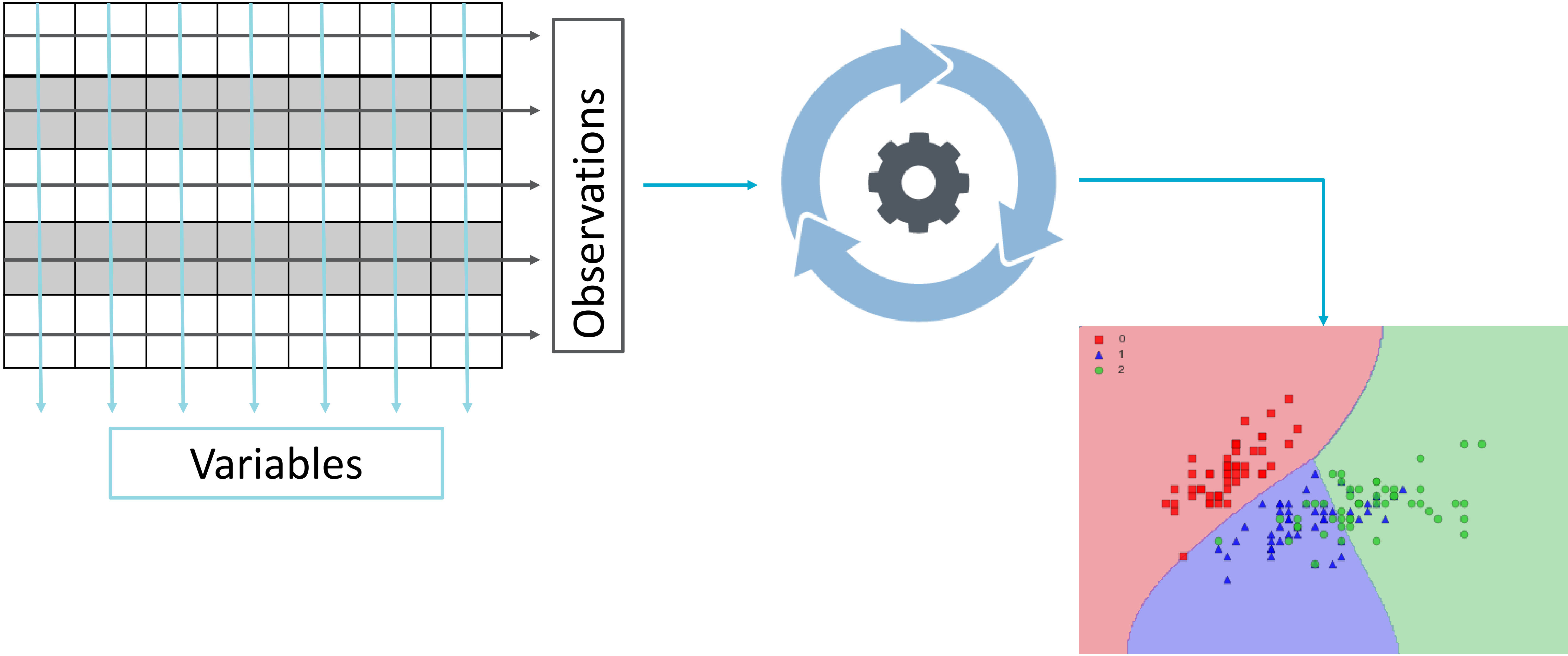
UNIVERSITY OF CANBERRA

- Data Modelling
  - Supervised
  - Unsupervised
- Correlation
- Linear Regression

DATA MODELLING

- Modelling is the process of teaching machines to learn relationships between the data variables for sake of delivering a business value such as predicting an outcome or discovering potential issues or trends
- Data Modelling makes assumptions about the data distributions & relationships
- A data model provides a simple, low-dimensional summary of the data and encodes the variations between the data variables

DATA MODELLING (2)



Observations

Variables

DATA MODELLING (3)

Supervised

Classification

Regression

Ranking

Scoring

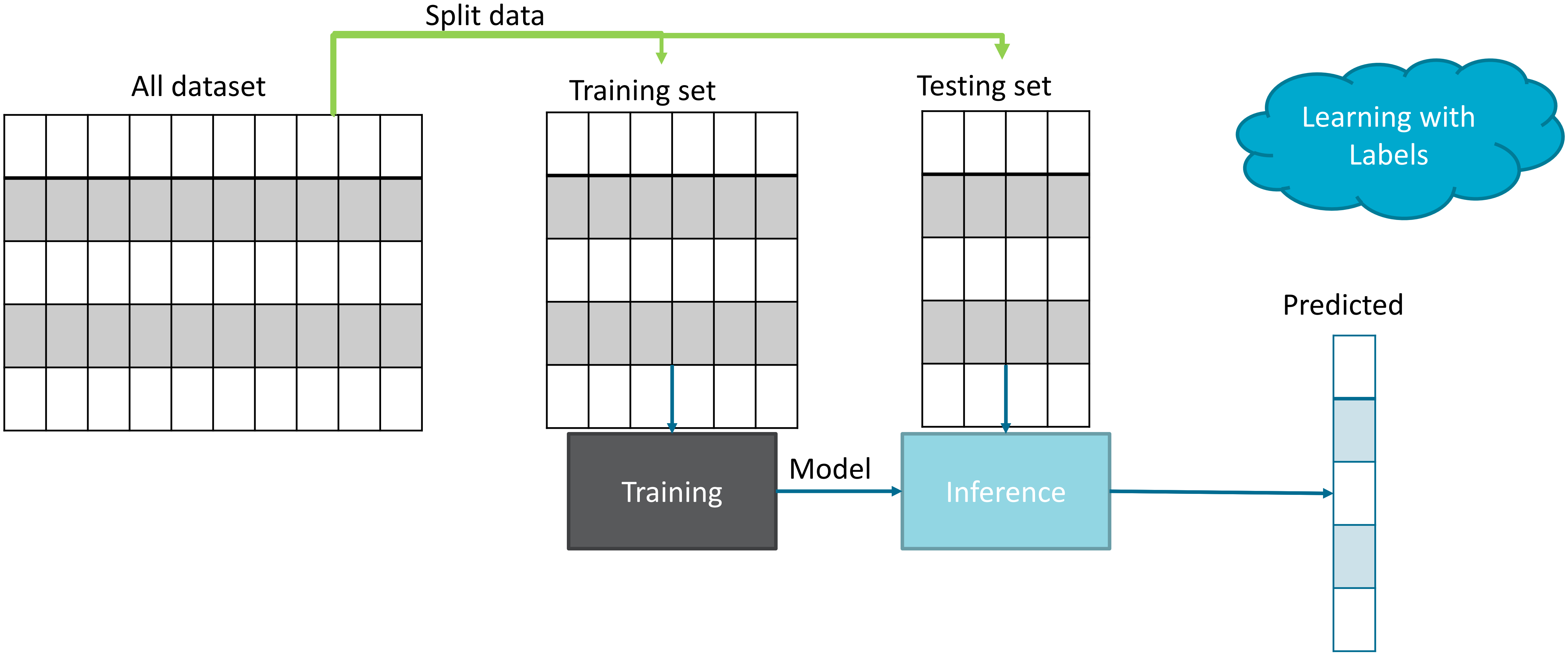
Un-supervised

Clustering

Dimensionality Reduction

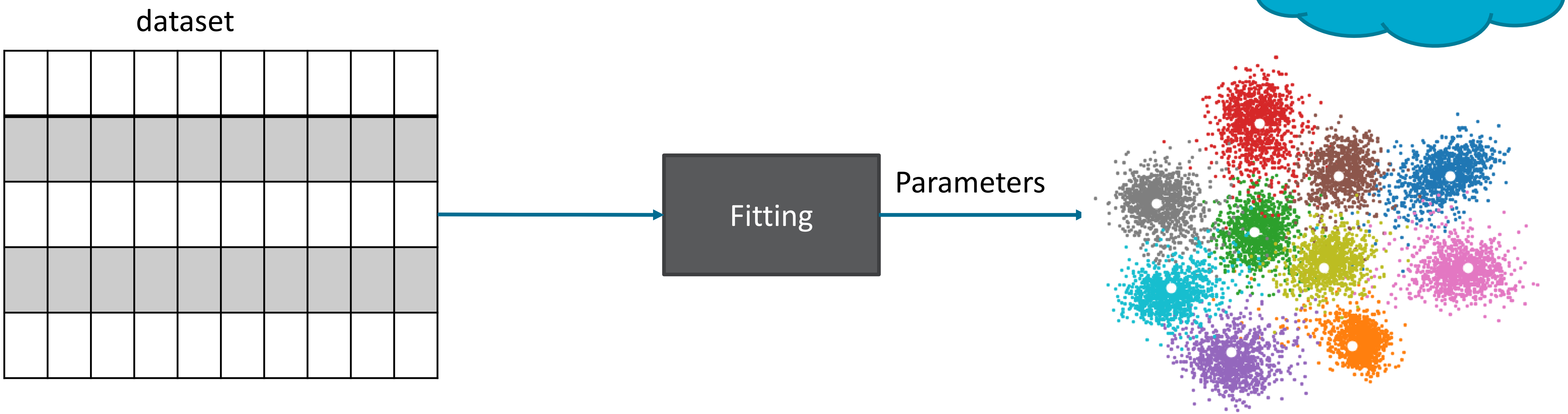
Reinforcement

DATA MODELLING – SUPERVISED



DATA MODELLING - UN-SUPERVISED

Learning without Labels



SUPERVISED MODELLING

Generative modelling

- Naive Bayes
- Bayesian networks
- Markov random fields
- Hidden Markov Models (HMM)

Discriminative modelling

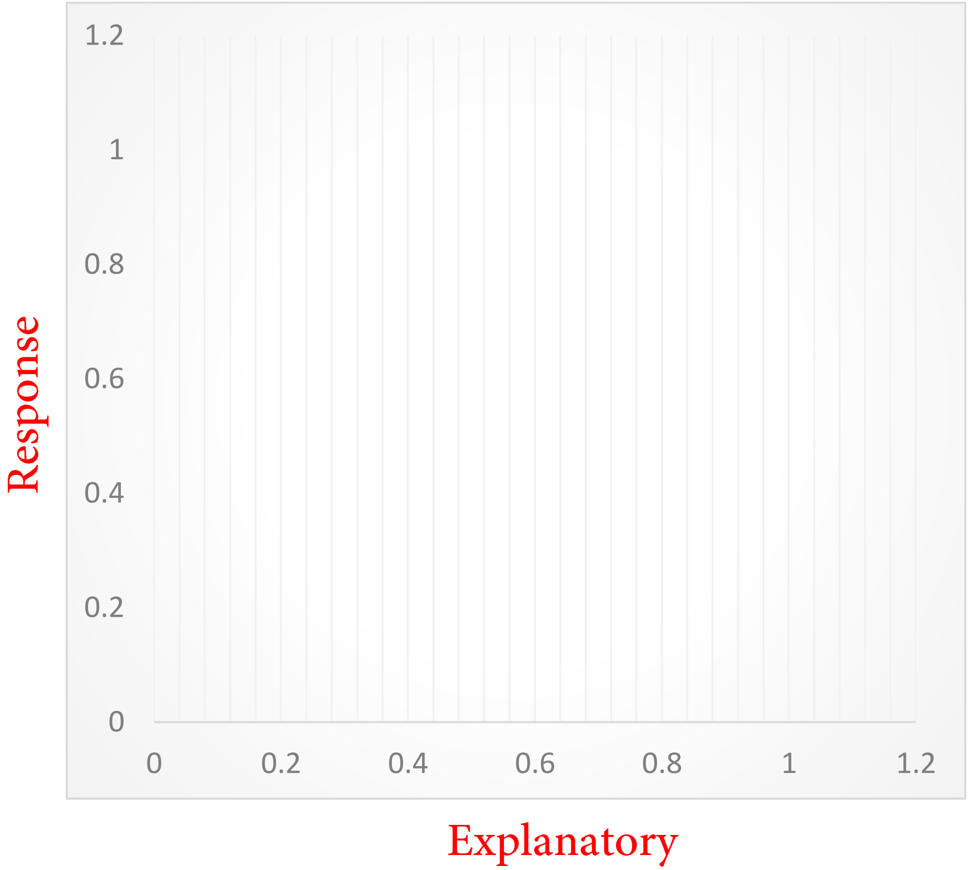
- Linear regression
- Support Vector Machine
- Neural networks
- Nearest neighbour
- Conditional Random Fields (CRF)

Features

Differences

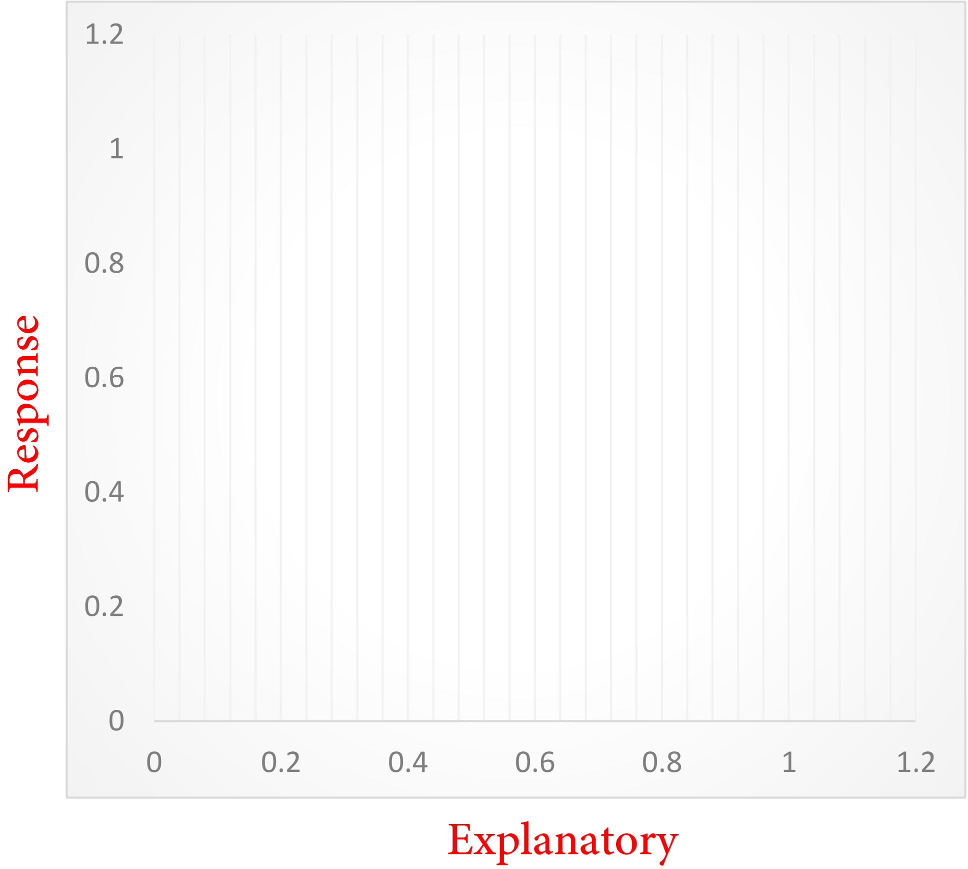
LINEAR REGRESSION

- A supervised learning method that enables us to assess and predict the relationship between explanatory variables and response variables.
- A response variable is the dependent variable that we want to predict, while, explanatory variables, some times refer to independent variables, are the variables that explain the changes in the response variable.



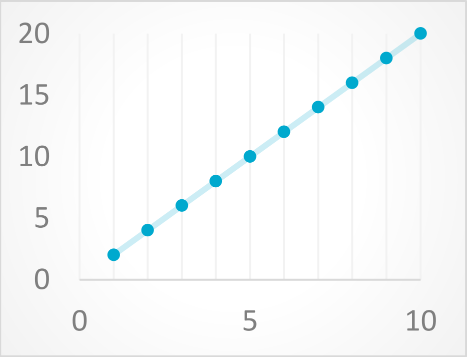
LINEAR REGRESSION (2)

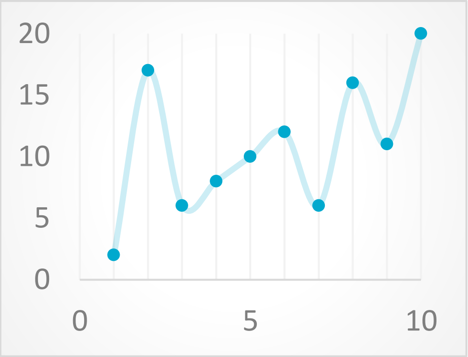
- For example, what is the relationship between:
  - The share price of xyz fuel company and the news on the social media about the middle east?
  - Soda and/or fast food and the gain in weights?
  - The ads in the YouTube videos and contents of the videos?
  - Etc.

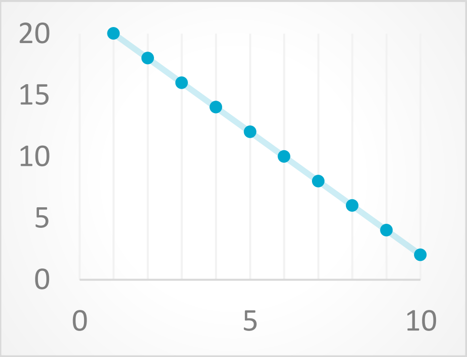


LINEAR REGRESSION (3)

- The relationships between variables have the following properties:
  - **Linearity** (i.e. linear or not linear)
  - **Direction** (i.e. positive or negative)
  - **Strength** (i.e. weak -- strong)





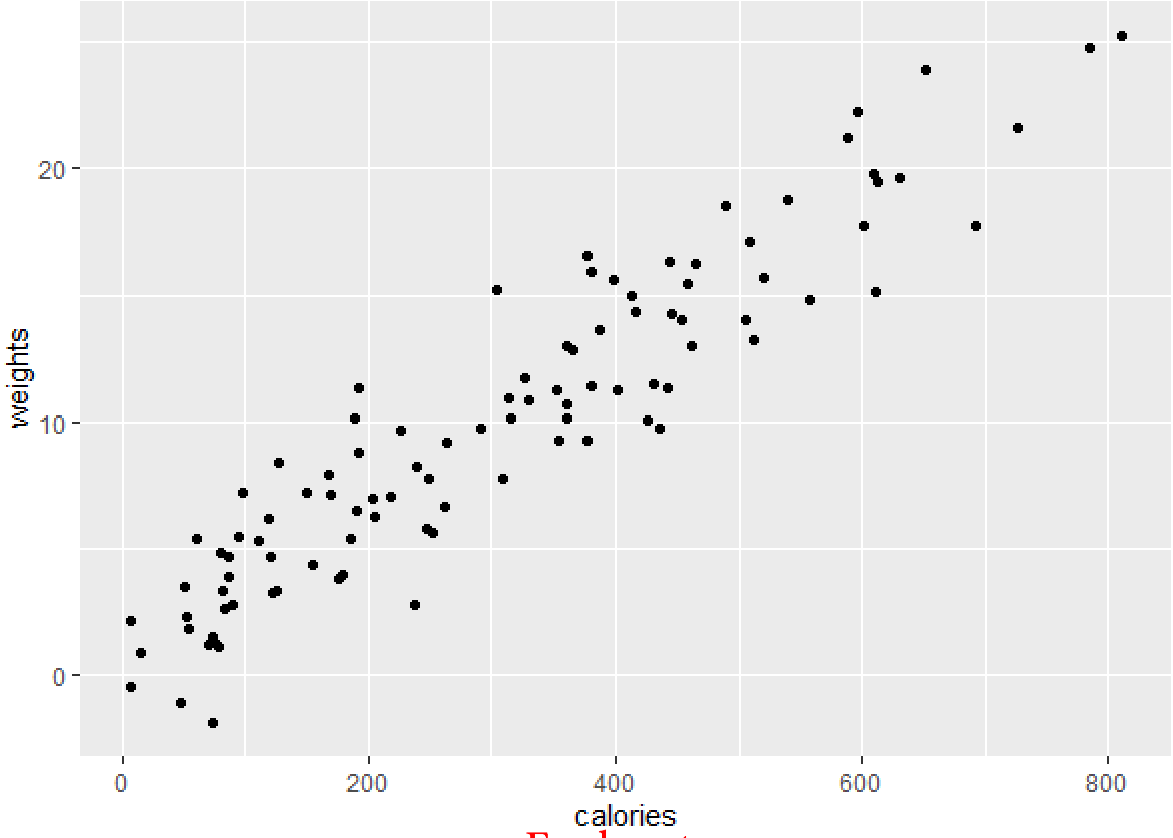


CORRELATION

- It is used to measure the relationship between two numerical variables.
- It is actually a measurement for the strength of the linear association between numerical variables.

Relationships:

- Linear
- Positive
- Strong



Response

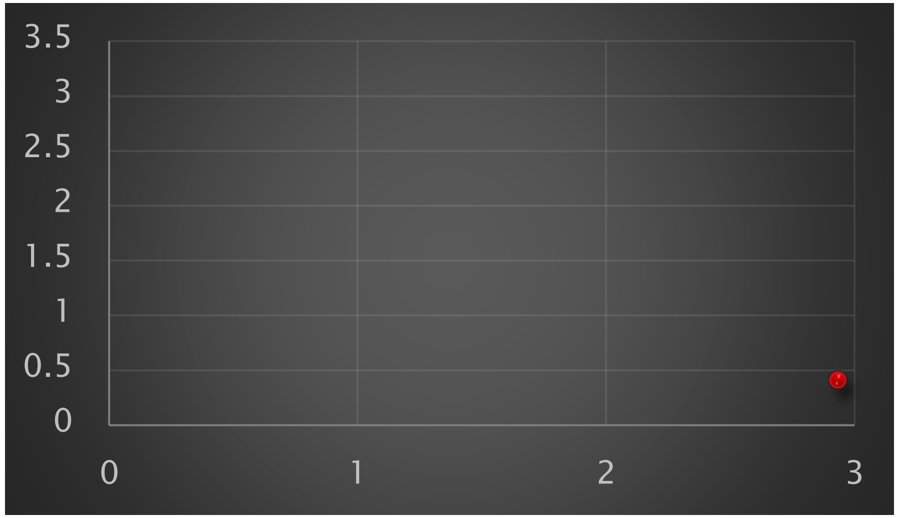
Explanatory

CORRELATION (2)

- A correlation between variables indicates that as one variable changes, the other variable tends to change in a specific direction.
- Understanding this relationship is useful because we can use the value of one variable to predict the value of the other variable.
- …, for example, height and weight of people are correlated in average, *i.e.* as height increases, weight also tends to increase.
  - Consequently, if we observe an individual who is unusually tall, we may predict that his weight is also above the average.

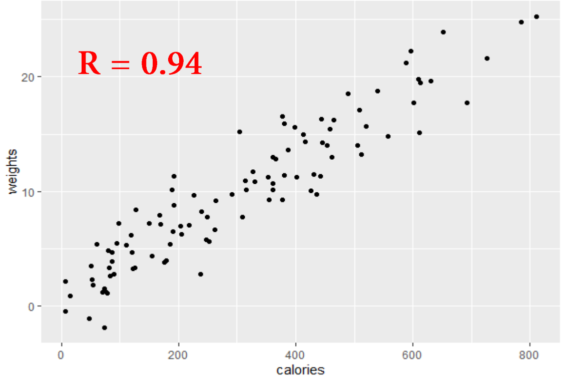
CORRELATION (3)

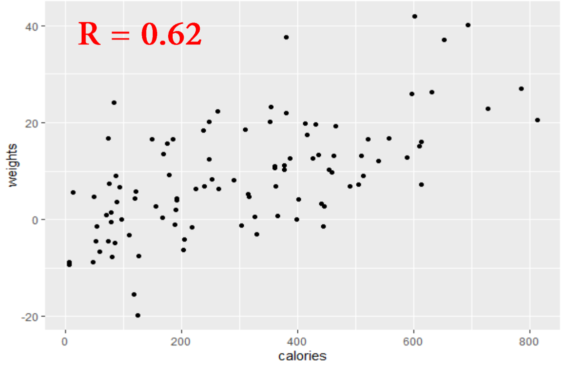
- A correlation coefficient is a quantitative measurement, which assesses the direction and the strength of tendency of the variables to vary.
- Scatterplot is a great way to check for the relationship between pairs of continuous variables, visually.
  - Each dot on the graph represents the relationship between the observational value of the two variables on the x axis and y axis.

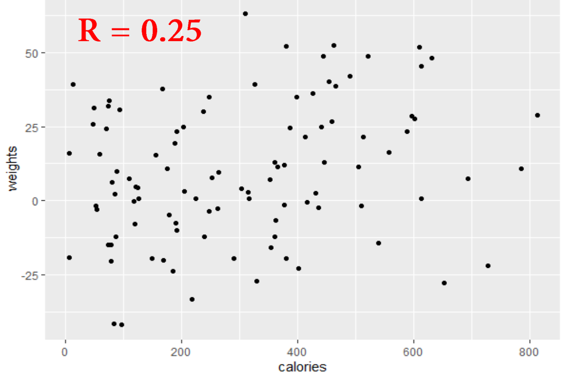


CORRELATION (4)

- **Properties of the correlation coefficient:**
  - We refer to the correlation coefficient with the symbol R.
  - The **magnitude** (absolute value) of the correlation coefficient represents the strength of the linear association between the explanatory and the response variable.
  - The higher the magnitude, the stronger the relationship between variables.





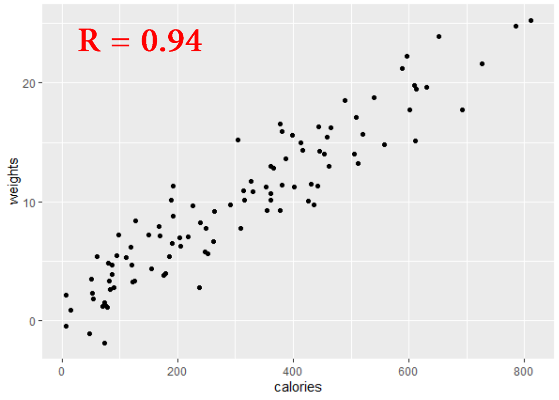


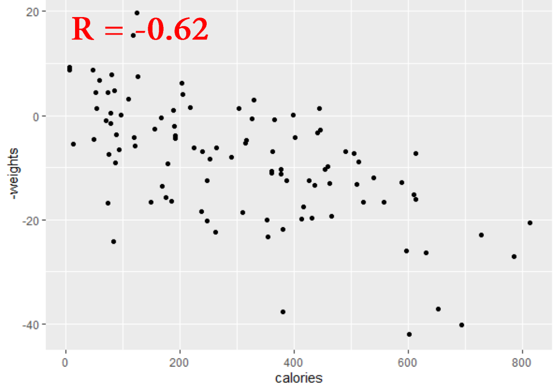
CORRELATION (5)

- **Properties of the correlation coefficient:**
  - The magnitude of the correlation coefficient (R) represents the strength of the relationship:
  - R is always between +1 and -1:
    - R = +1 means the y increases with the same amount as x increases.
    - R = -1 means the y decreases with the same amount as x increases.
    - R = 0 means y never changes when x does, *i.e.* the relationship is represented by a horizontal line.

CORRELATION (6)

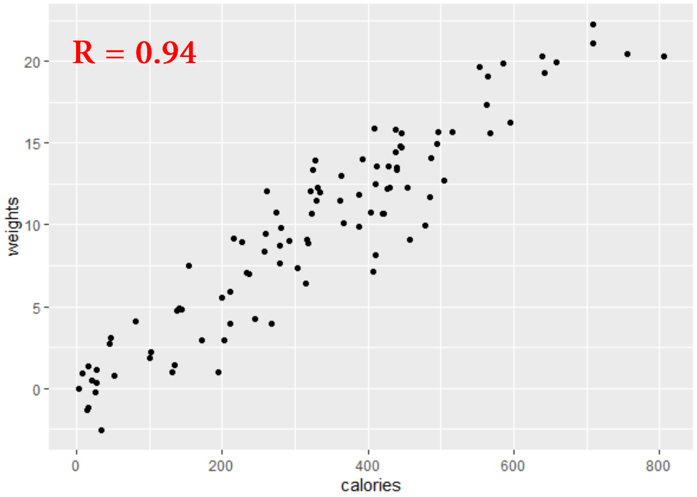
- Properties of the correlation coefficient :
  - The sign of the correlation coefficient indicates the **direction** of the relationship between the variables
  - For example:
    - a positive correlation is the relationship between the speed of a wind turbine and the amount of energy it produces.
    - a negative correlation is the relationship between outdoor temperature and heating costs.

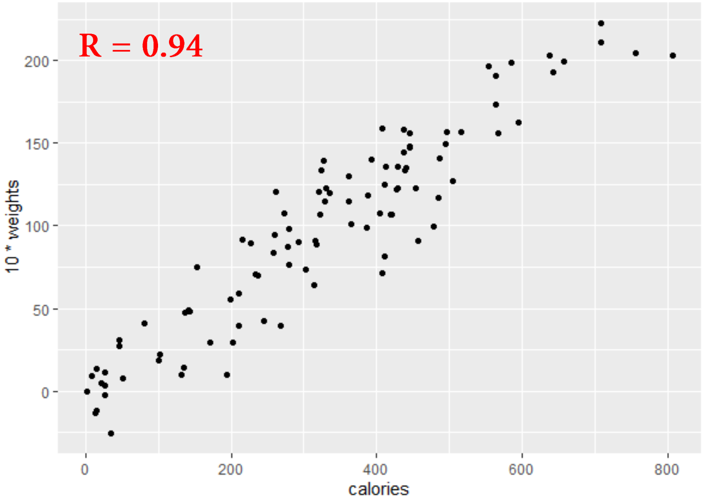




CORRELATION (7)

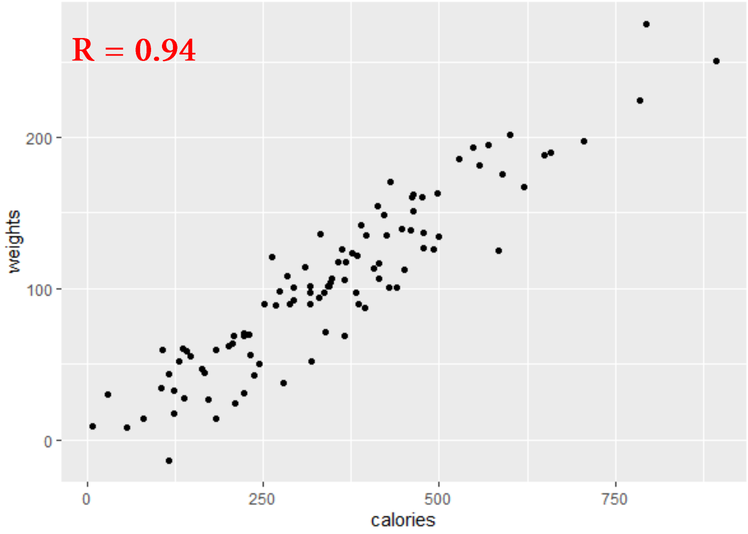
- **Properties of the correlation coefficient :**
  - The correlation coefficient is **unit-less**, *i.e.* the value will stay very similar even with scaling or normalising any of the variables.
  - Correlation coefficient between x and y is the same as the one between y and x.

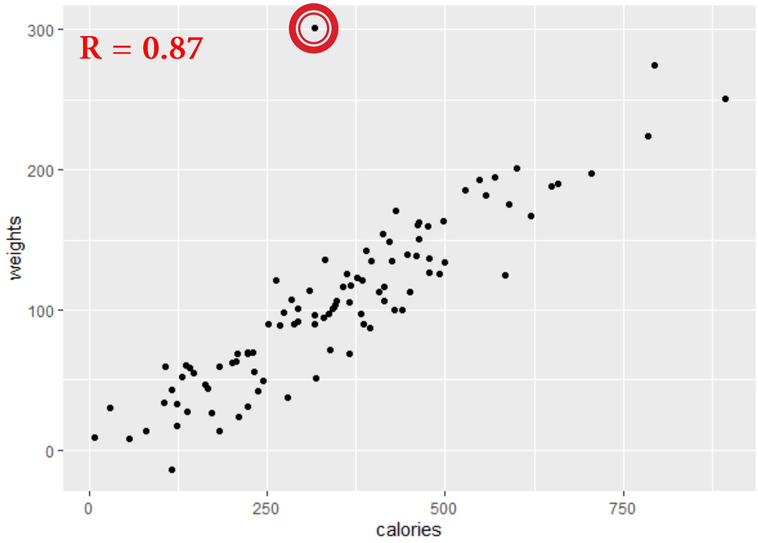




CORRELATION (8)

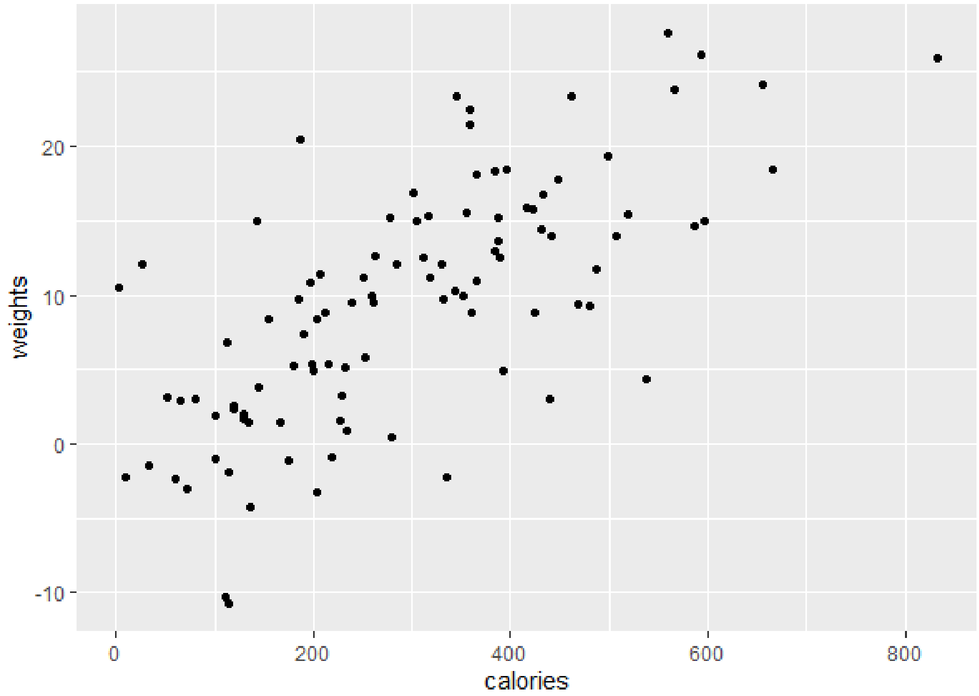
- Properties of the correlation coefficient :
  - The correlation coefficient is very **sensitive to outliers.**





CORRELATION (9)

- Question?
  - Which of these values is the best guess for the correlation coefficient between calories consumption and gained weights?



a) 0.1

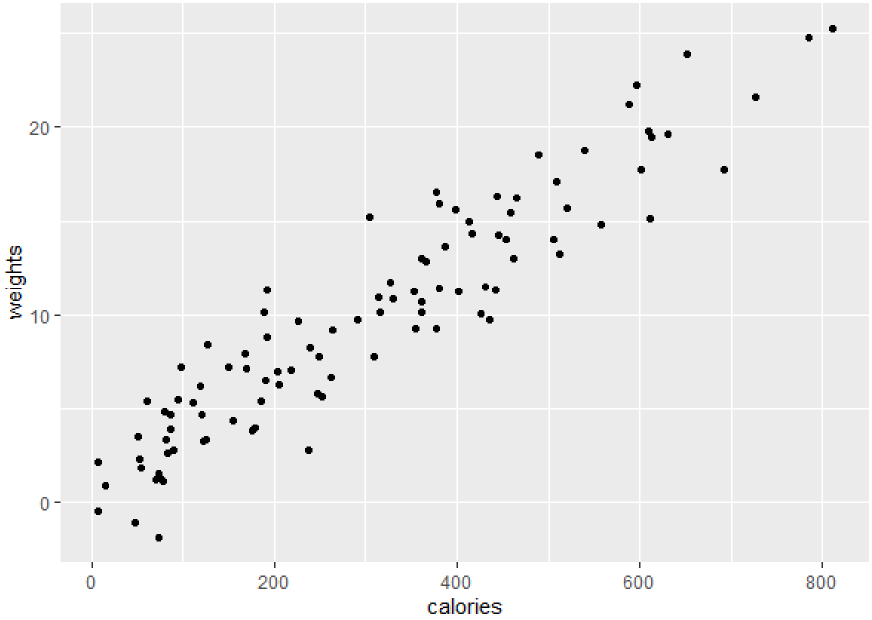
b) 0.7

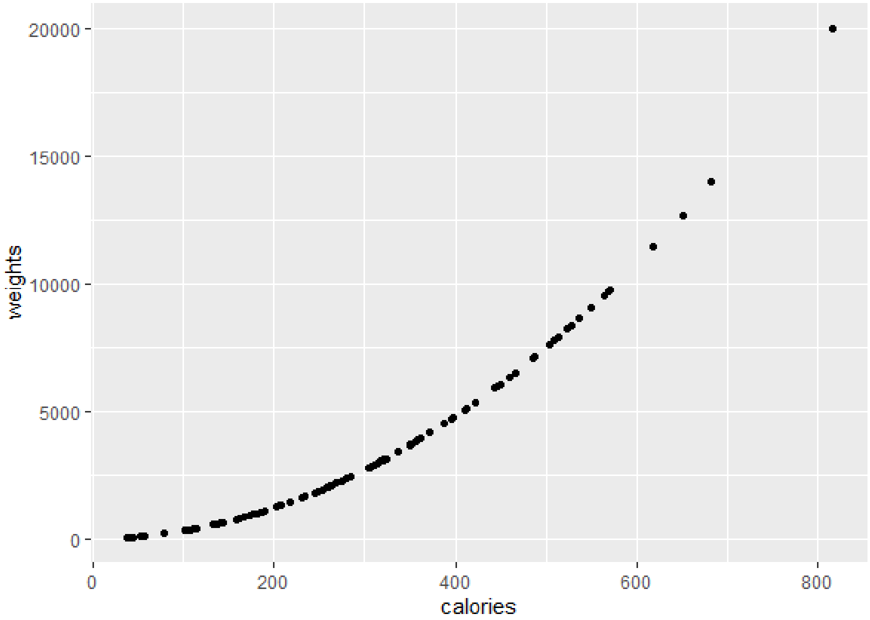
c) -0.5

d) 1.2

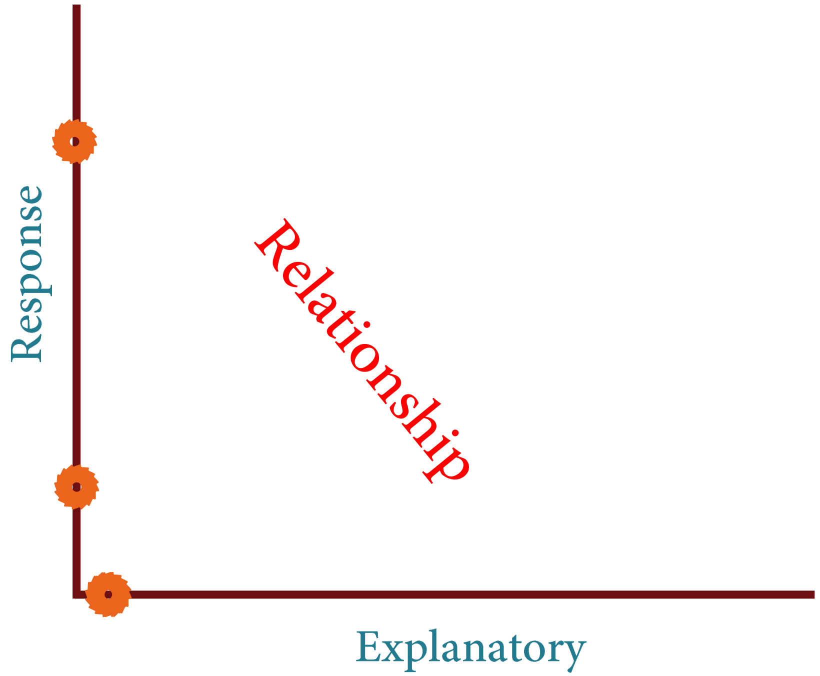
CORRELATION (10)

- Question?
  - Which of these has the strongest ‘linear’ correlation?



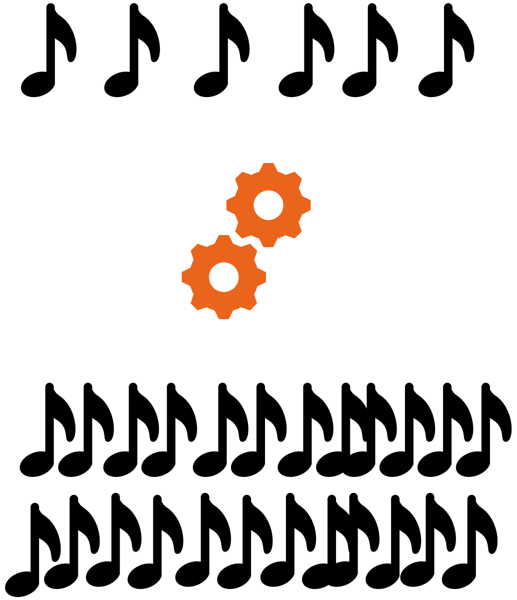


LINEAR REGRESSION AND CORRELATION



LINEAR REGRESSION

- Linear regression is a linear approximation of a causal relationship between two or more variables
- Process goes in the following steps:
  1. Get Sample Data
  2. Design a model that works for that sample
  3. Make predictions for the whole population
  4. Measure performance



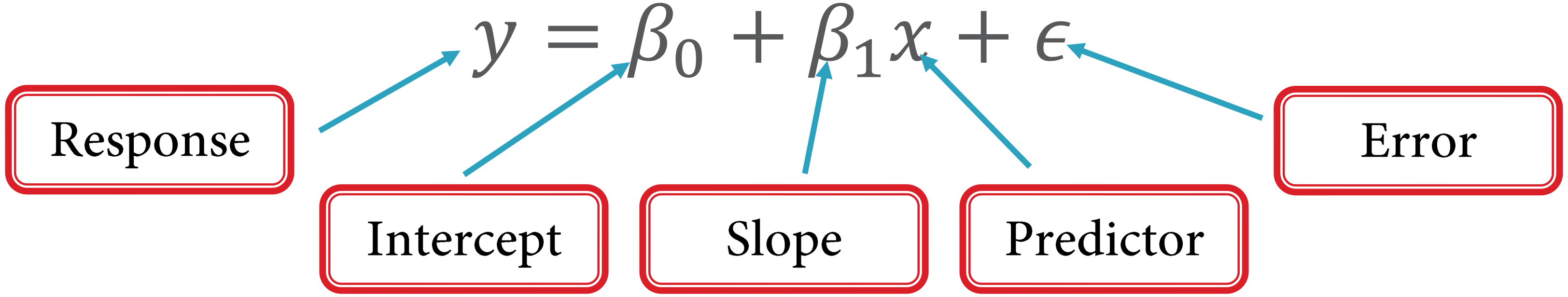
LINEAR REGRESSION (2)

- Linear regression aims to establish a linear relationship between predictor (i.e. explanatory) variable(s) and predicted (i.e. response) variable.
- …, so we can estimate the values of the response variable when only the values of the predictor variable are available.
- The predicted variable is a continuous variable, while the predictor can be categorical or numeric variable.

LINEAR REGRESSION (3)

- The linear regression can be estimated by following this formula:

$$y = \beta_0 + \beta_1 x + \epsilon$$



Response

Intercept

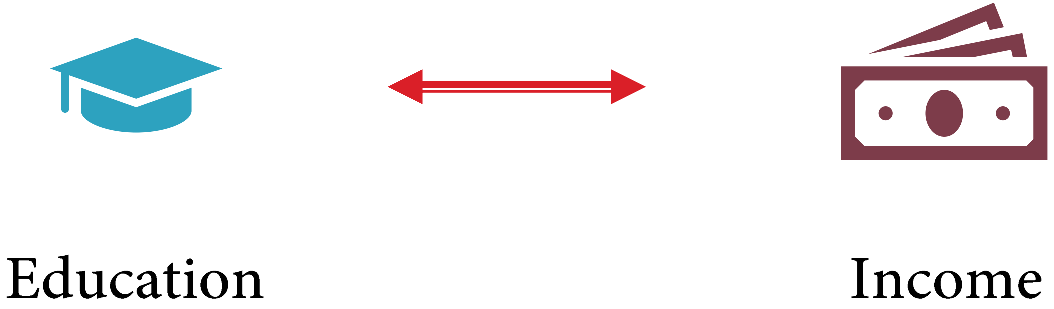
Slope

Predictor

Error

LINEAR REGRESSION (4)

- For example:



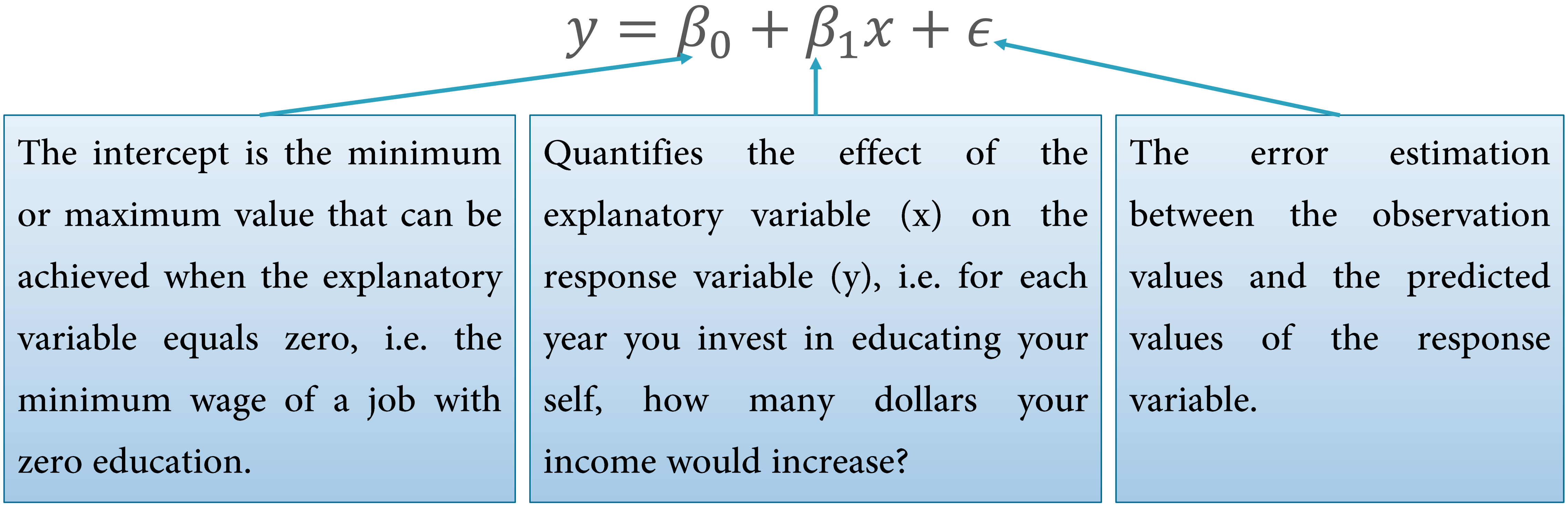
Education

Income

- More education you will get, is it more likely that your income will increase?

LINEAR REGRESSION (5)

- Interpreting the parameters of the following equation with the previous example as following:



$$y = \beta_0 + \beta_1 x + \epsilon$$

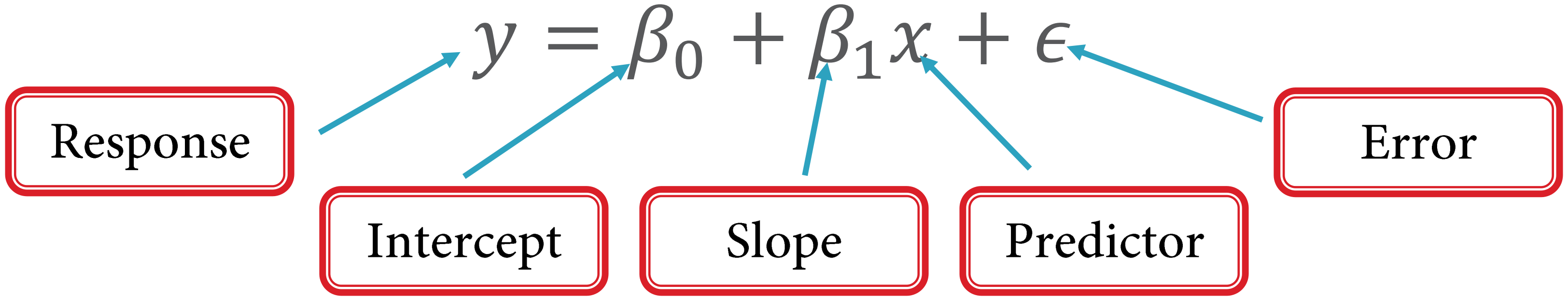
The intercept is the minimum or maximum value that can be achieved when the explanatory variable equals zero, i.e. the minimum wage of a job with zero education.

Quantifies the effect of the explanatory variable (x) on the response variable (y), i.e. for each year you invest in educating your self, how many dollars your income would increase?

The error estimation between the observation values and the predicted values of the response variable.

LINEAR REGRESSION (6)

- The linear regression can be estimated by following this formula:



$$y = \beta_0 + \beta_1 x + \epsilon$$

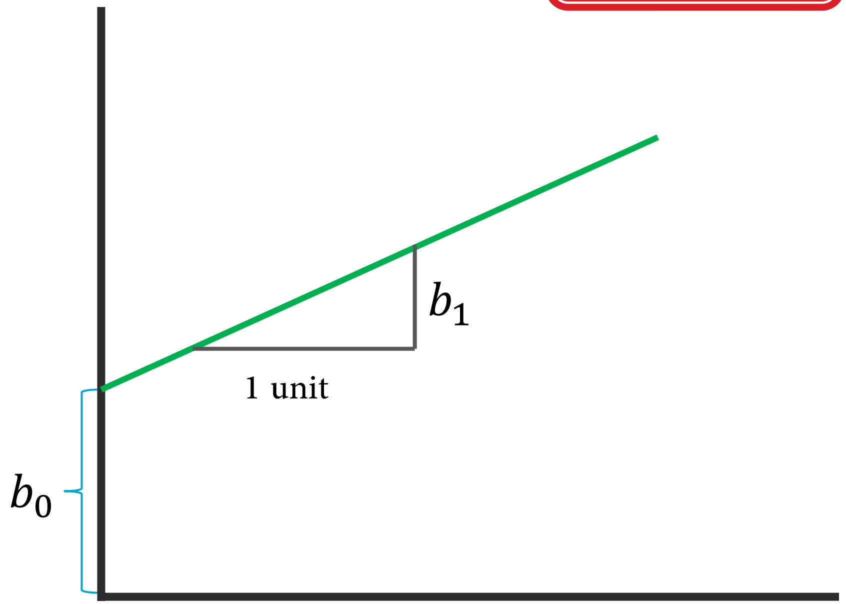
Response

Intercept

Slope

Predictor

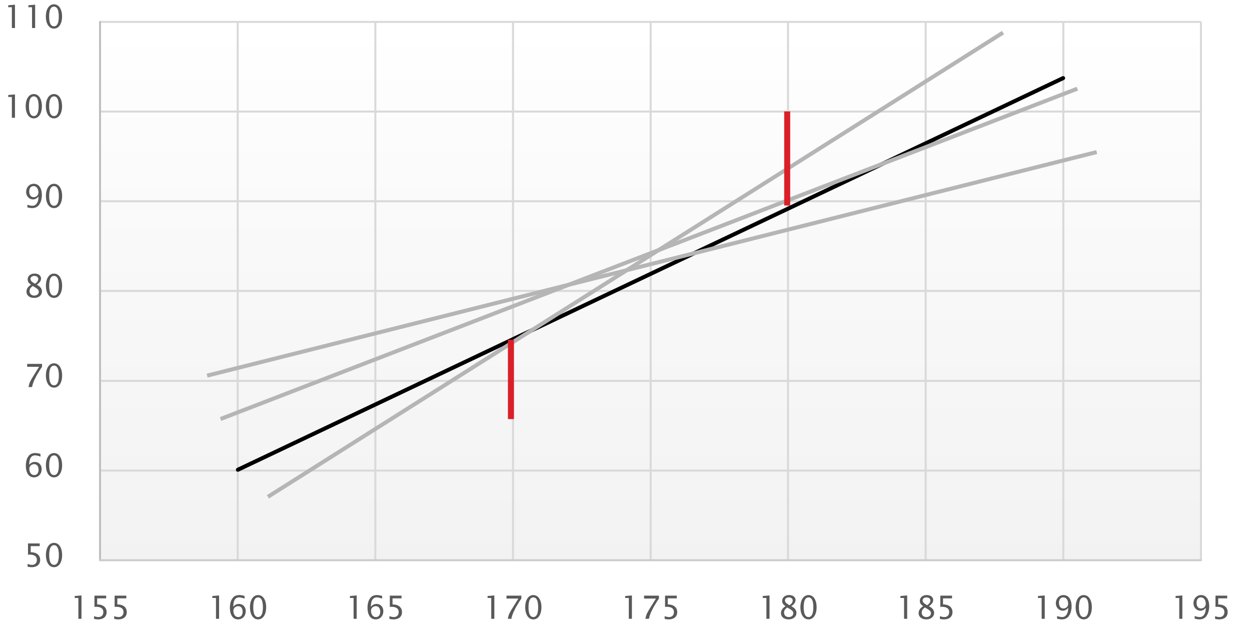
Error



…, so now given (x,y) pairs of data samples, can we estimate b_0 and b_1 ?

RESIDUALS

- To estimate the regression parameters (b_0 and b_1), we *minimise* the difference between the observed and the predicted values.
- This difference is called *residuals.*



RESIDUALS (2)

- To find the regression line that best fits the linear relationship between two variables, we minimise either:
  - *Sum of absolute error, or*
  - *Sum of square errors*

The sum of square error is more common and efficient as minimizing the large error values would look after the details of the residuals and would converge slowly, which decreases the likelihood of falling into local minimum.

EXAMPLE

- Here, we demonstrate how to build a linear regression model using “cars” dataset, which comes with R by default. cars is a standard built-in dataset
- This dataset consists of 50 observations and only 2 variables – dist and speed.

```
> head(cars)
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
```

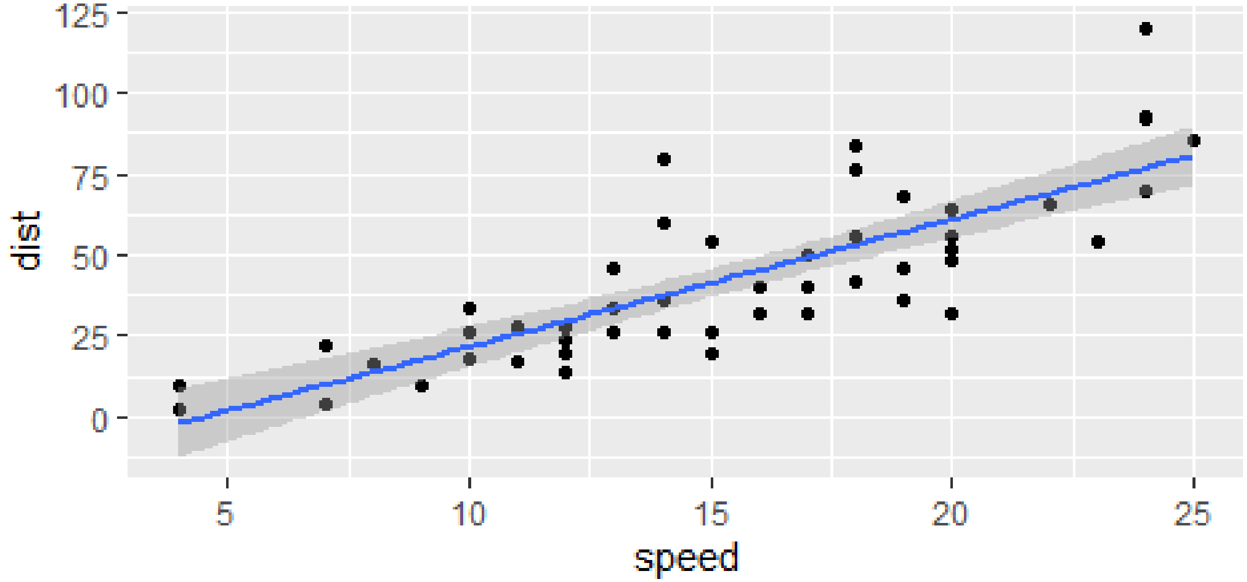
We aim to build a linear regression that predicts the distance values from the speed of the car

EXAMPLE (2)

- First, we need to understand the relationship between these two variables
  - Is it linear? Is it strong enough? Is there any outlier?
- To answer these question, we have two options:
  1. Visually, inspecting the relationships between variables, or
  2. Numerically, testing the linearity and normality hypotheses.

EXAMPLE (3)

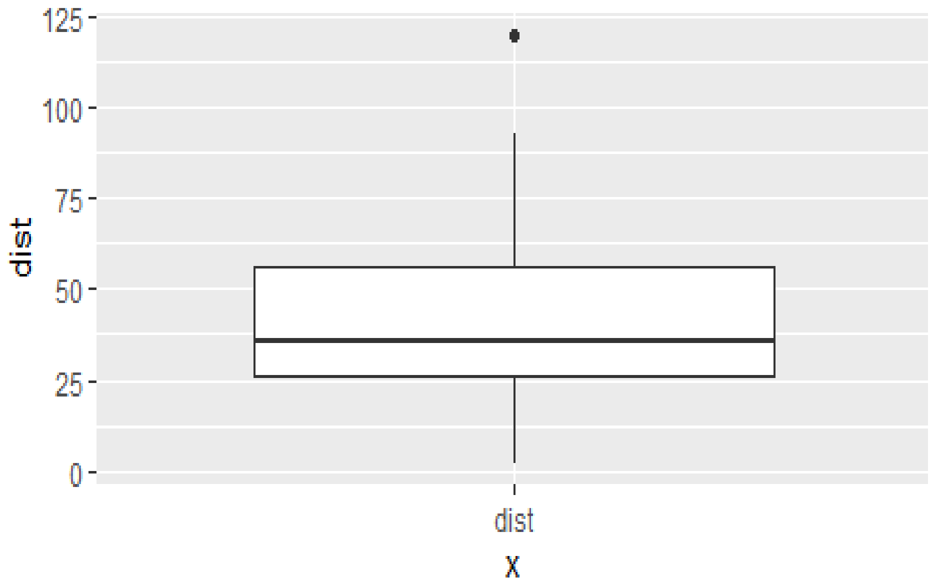
- Visually, checking the linearity:
  - <u>Scatter plots</u> can help visualise any linear relationships between the response variable and predictor variables.

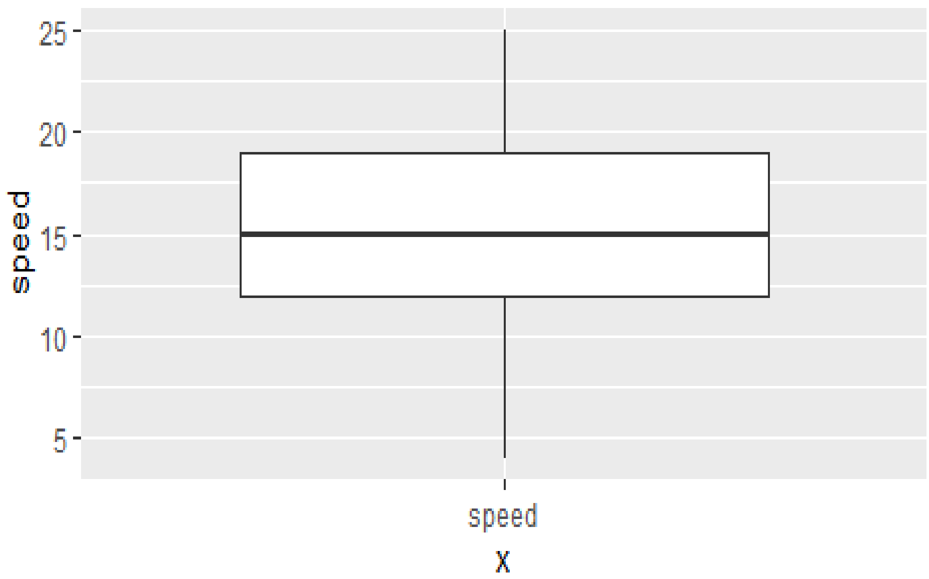


The scatter plot along with the smoothing line suggests a linearly increasing relationship between the variables.

EXAMPLE (4)

- Visually, we may use Box Plot to check for the outliers:

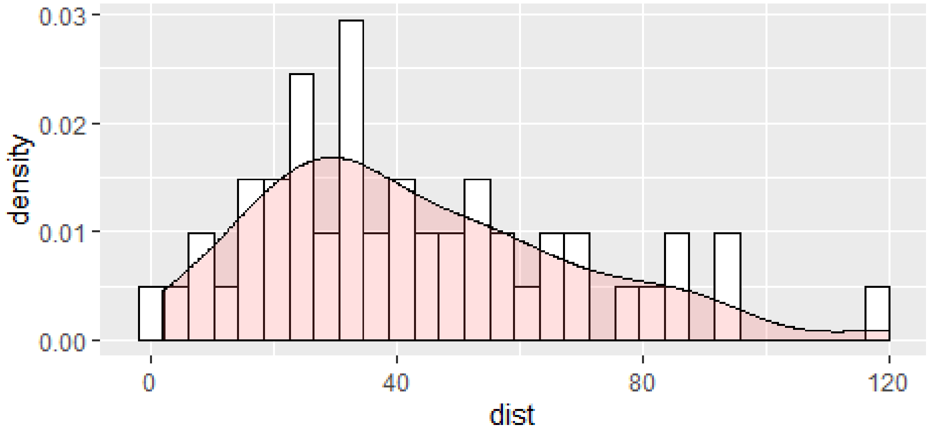


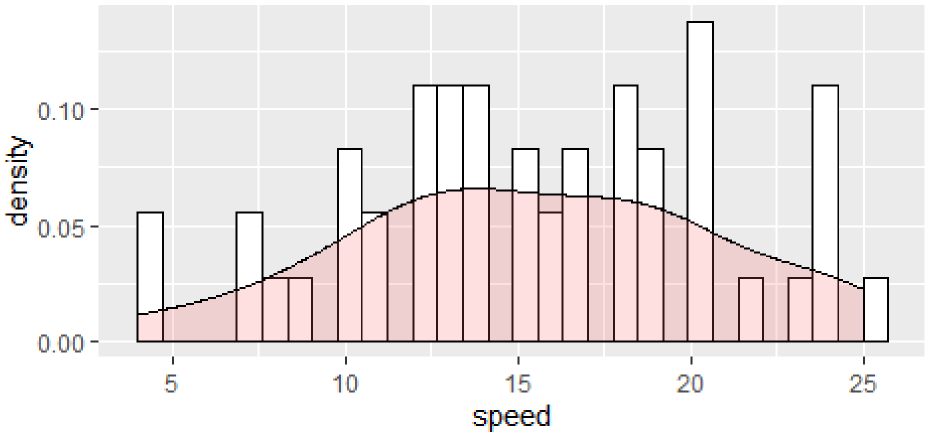


- datapoints that lie outside the (1.5 * IQR) is considered as outliers, where, IQR is calculated as the distance between the 25th percentile and 75th percentile values.

EXAMPLE (5)

- Inspecting Density Plot for normality:
- Linear regression modelling requires the two variables to be bi-variate normal. This assumption can be checked with a histogram with density plot for the two variables or by a Q-Q-Plot of the fitted model.





A regression line is the best prediction of Y given the value of X, when X and Y follow a bivariate normal distribution.

EXAMPLE (6)

- Inspecting linearity numerically:
  - This can be done using Correlation.
  - Correlation is a statistical measure that checks the level of linear dependence between the two variables.
  - Correlation can take values between -1 to +1, where a value closer to 0 means a weak relationship between the variables.
  - A low correlation (-0.2 < r < 0.2) probably suggests that much of variation of the response variable (Y) is unexplained by the predictor (X).

```
cor(cars$speed, cars$dist)
#> [1] 0.8068949
```

BUILDING MODEL

- In R, we use **lm** function to build the linear model
- When calling the lm function, the variable that we want to predict is put to the left of the ~ symbol, and the variables that we use to predict is put to the right of the ~ symbol.

```
# fit regression line to predict cars’s stopping distances
from cars’s speed
model <- lm(dist~ speed, data = cars)
```

RECOMMENDED READING

- You are recommended to read chapter 19 from the "Introduction to Data Science -Data Analysis and Prediction Algorithms with R" book:
- https://rafalab.github.io/dsbook/linear-models.html