Durham University
Research Methods Centre

# MATH43515: Multilevel Modelling

## Lecture 1: Correlation, Simple Linear Regression

**Module Convenor/Tutor**
Andy Golightly

# Outline

1. Variables

2. Correlation

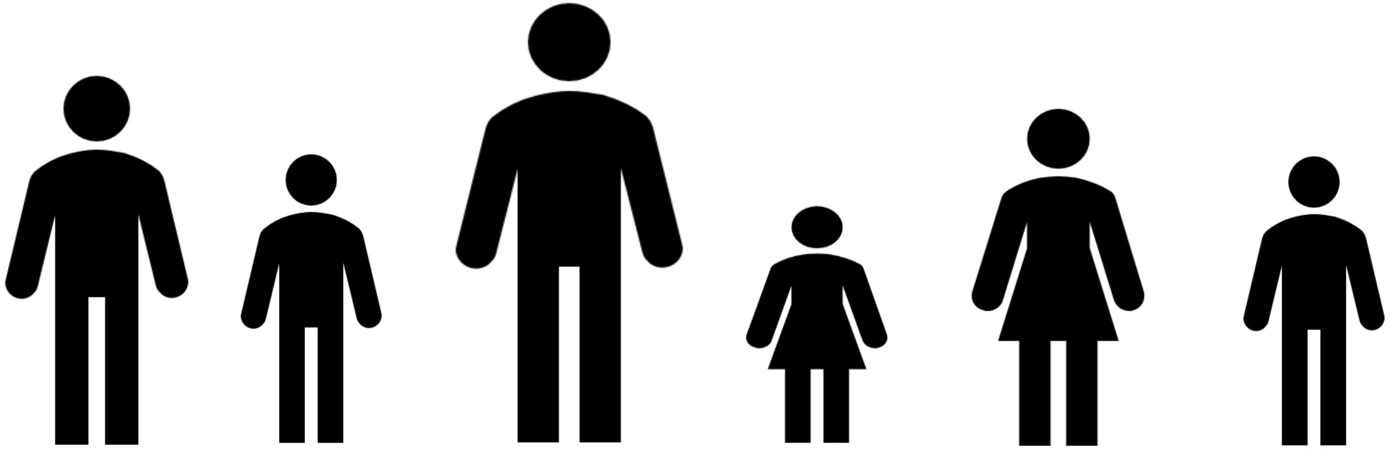
3. Simple linear regression

4. Regression assumptions and diagnostics

# Variables

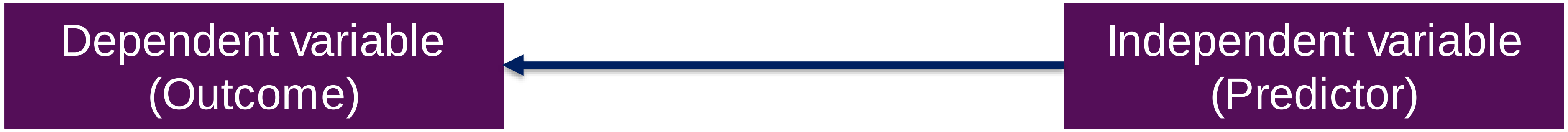
Variables are **measurable entities that can change or *vary***, for example*:*

- **Between people**, such as height
- **By location**, such as sunlight at different latitudes
- **Across time**, such as weight



# Variables

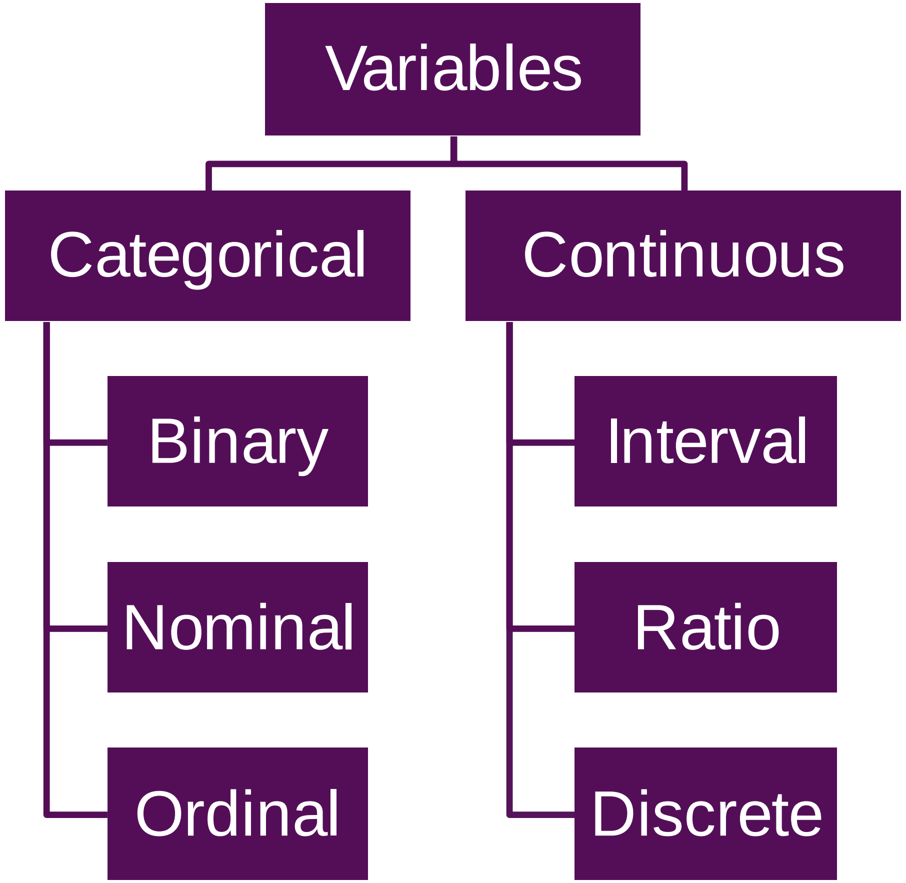
- Most hypotheses involve at least two variables: a proposed cause and a proposed outcome
- The **independent variable** is the variable we think might be associated with the outcome; its value does not depend on other variables (sometimes also called **predictor variable**)
- The **dependent variable** is the effect that depends on the value of the independent variable (sometimes also called **outcome variable**)



# Variables

- Variables (whether independent or dependent) can generally be grouped into two categories: categorical and continuous

- Understanding these distinctions is important for making sense out of the data and to determine appropriate statistical methods



# CORRELATION

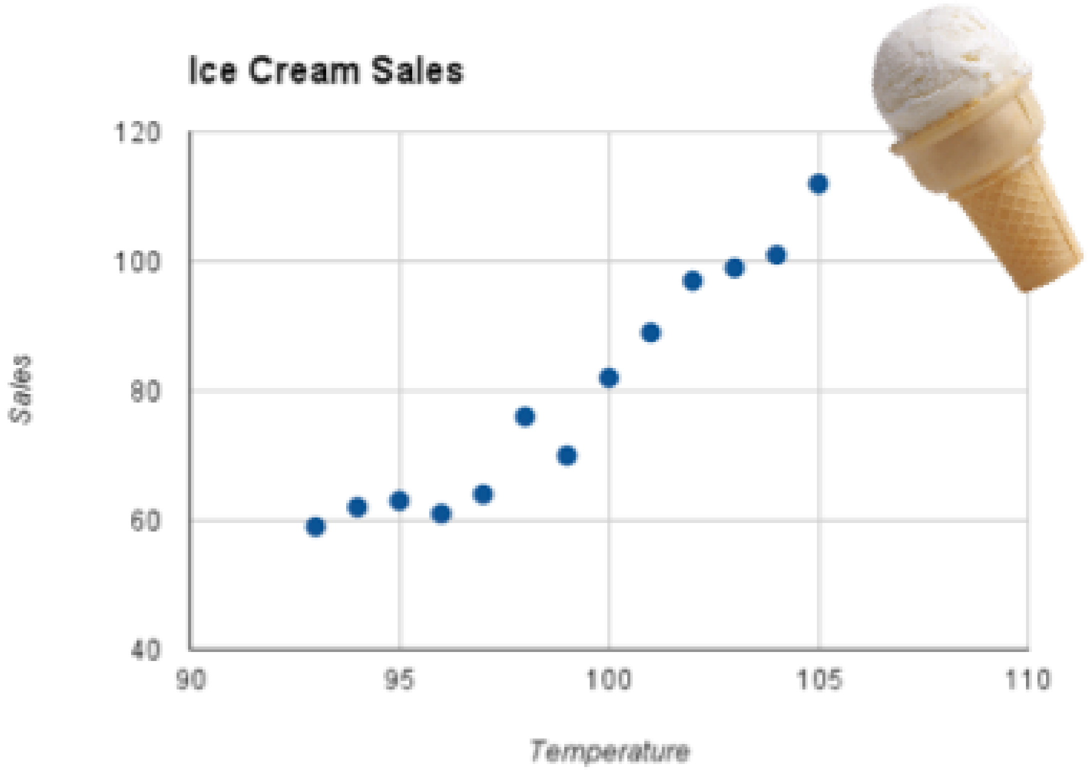
# Linear Association

- Before thinking of any analysis always see your data graphically
- Graphs are just a waste of your precious time? Nope!
- Graphs are a useful way to look at your data before you get to the nitty-gritty of analyzing them.

# Linear Association

**What Scatterplot says, you say:**

- Is there a relationship between the variables.
- What type of relationship ?
- (Linearly) **correlated**?
- Is there an outlier?



# Linear Association

A psychologist was interested in the *effects of exam stress* on *exam scores.*

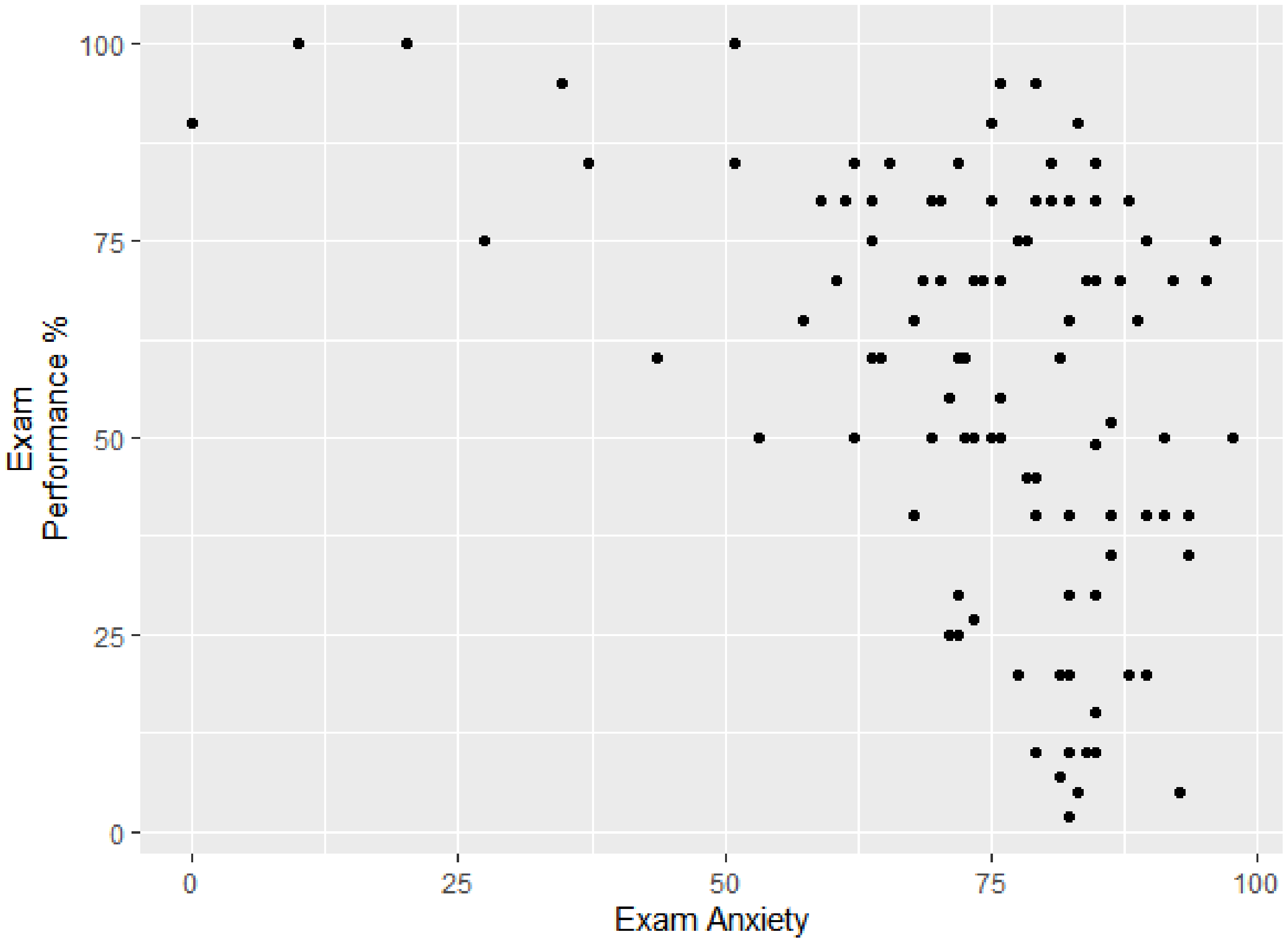
- Anxiety is measured by a questionnaire before an exam.
- Percentage mark of each student was used to assess the exam performance.
- The **first thing** that the psychologist should make scatterplot of the two variables.
- **We want <u>Anxiety (predictor or independent)</u> plotted on the *x*-axis and <u>Exam</u> (outcome or dependent ) on the *y*-axis.**

# Linear Association

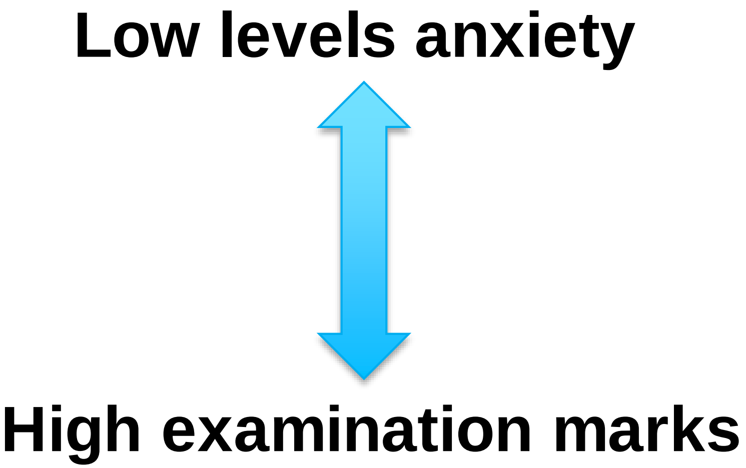
- **What can we understand…?**

- Is there an association?

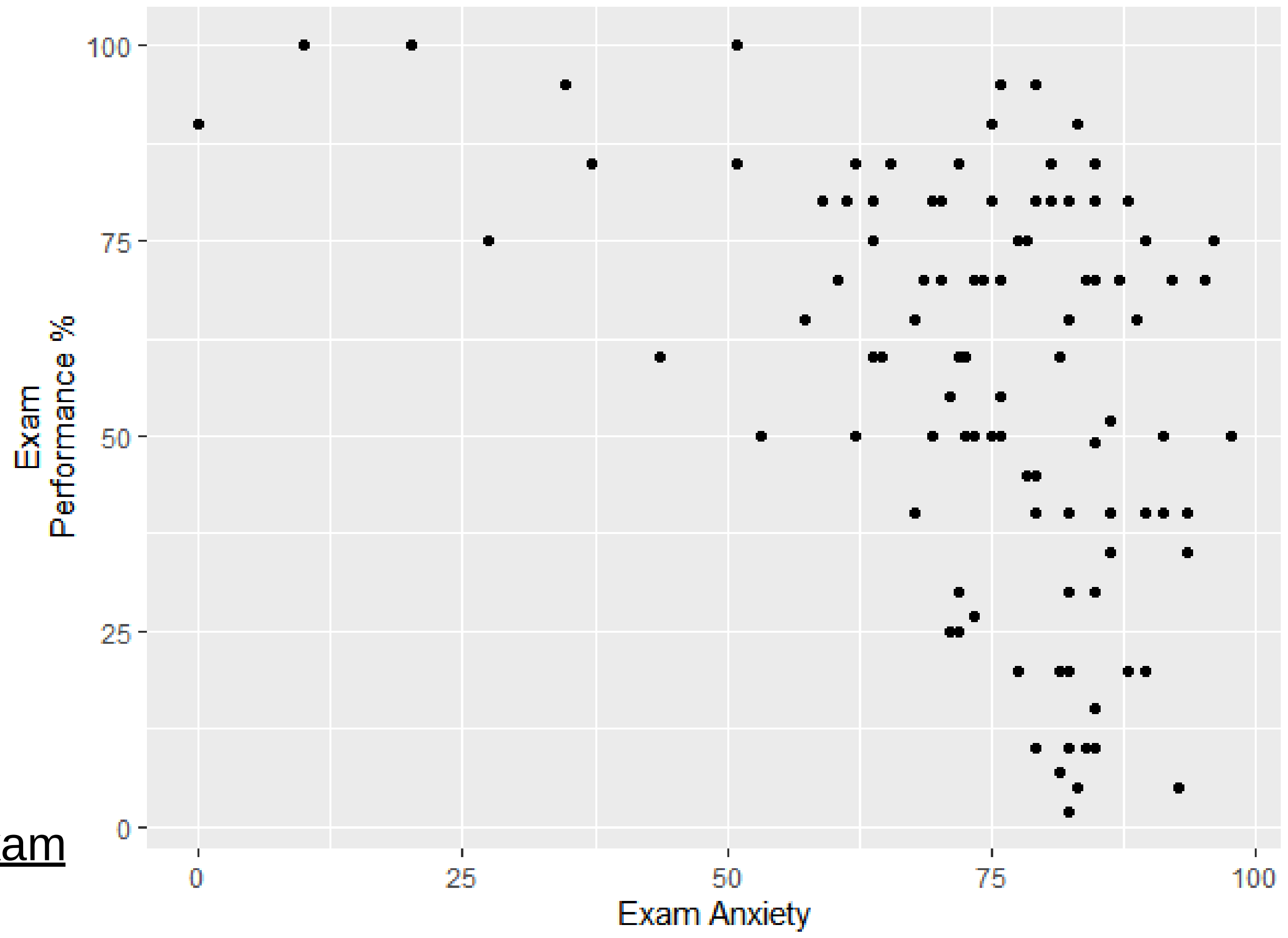
- What type of association?



# Linear Association


Low levels anxiety

High examination marks



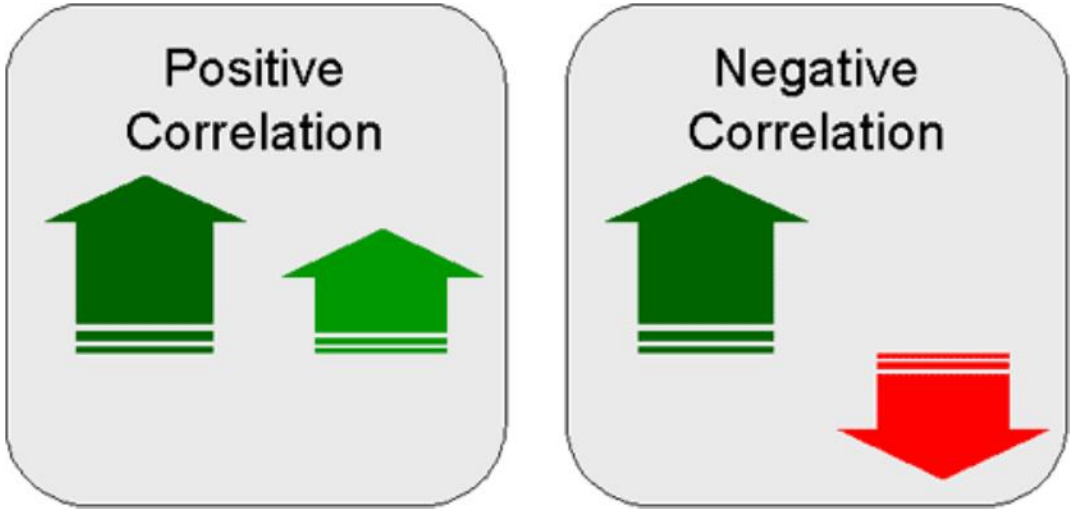
No cases having low anxiety and low exam performance.

Durham University

# Linear Association

Association between two continuous variables:

- Linear or non-linear
- Positive or negative
- No association



# Covariance

**Basic Idea of Covariance:**

- If we are interested in whether two variables are related, then we are interested in whether:

changes in one variable are met with similar changes in the other variable?

# Covariance

**Toffee Sweets Example:**

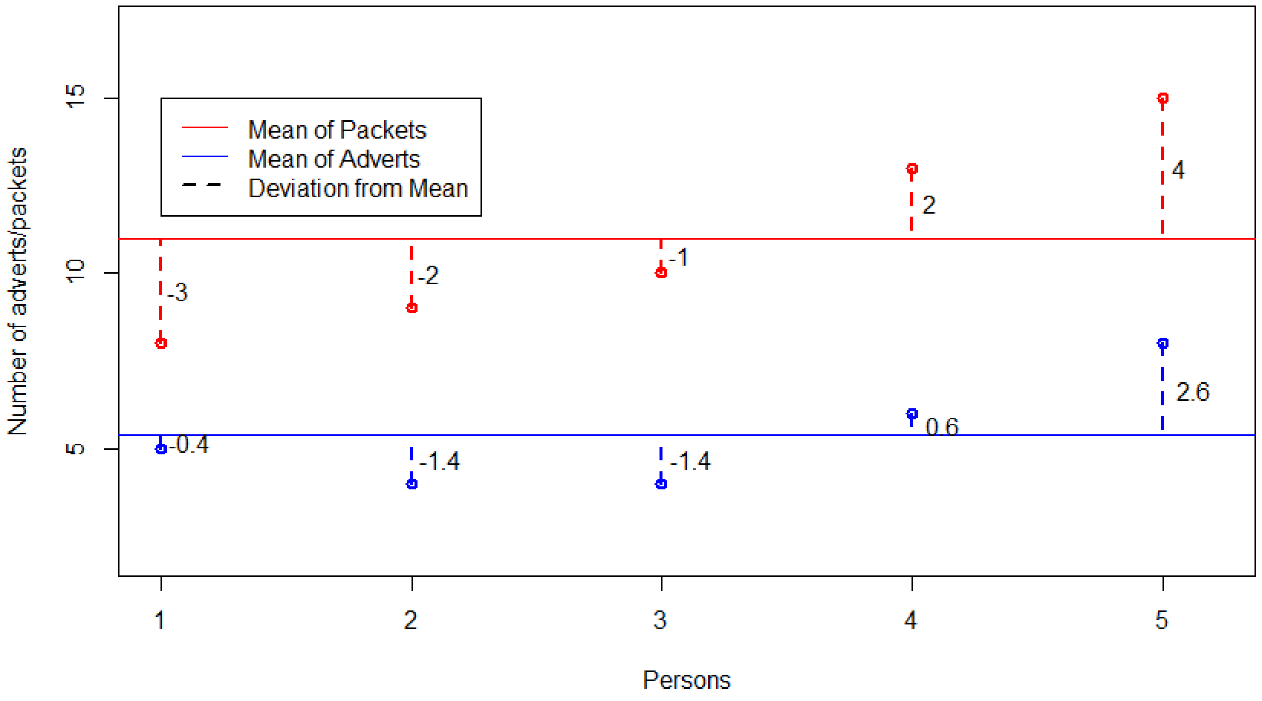
Imagine we took five people and subjected them to a certain number of advertisements promoting toffee sweets, and then measured how many packets of those sweets each person bought during the next week.

| Persons | 1 | 2 | 3 | 4 | 5 | Mean | SD |
|---|---|---|---|---|---|---|---|
| Adverts watched | 5 | 4 | 4 | 6 | 8 | 5.4 | 1.67 |
| Packets bought | 8 | 9 | 10 | 13 | 15 | 11 | 2.92 |

Is there an association?

# Covariance

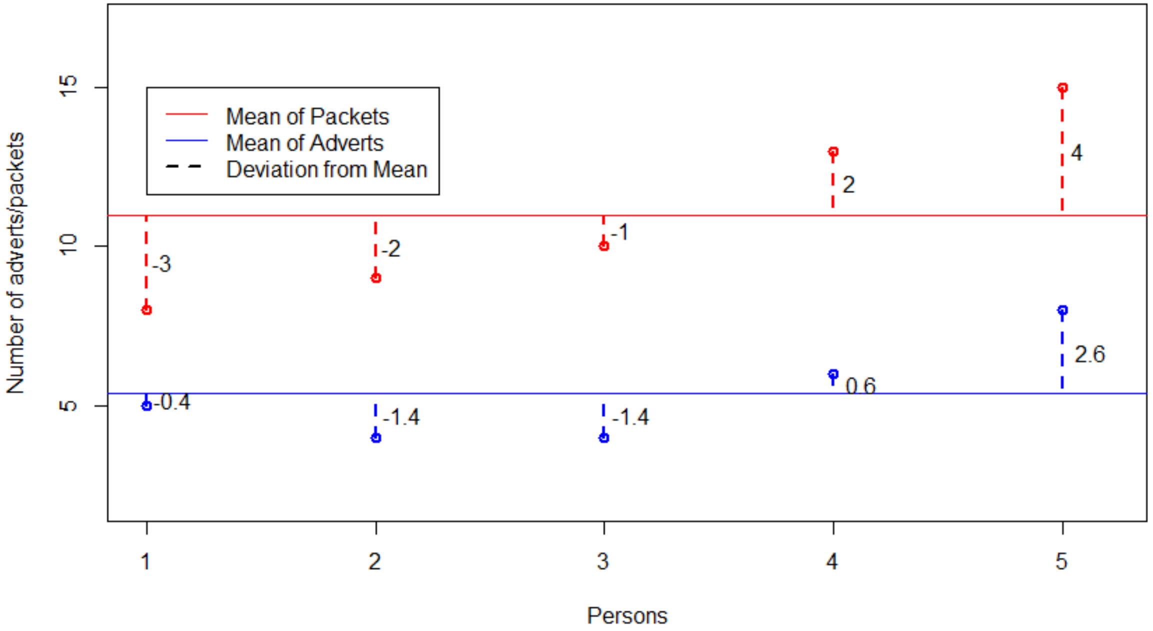
As one variable deviates from its mean, the other variable should deviate from its mean in the same or opposite way.



# Covariance

**Calculating similarity between the patterns:**

- Calculate mean product of deviations per individual or items



$$cov(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

# Covariance

- A positive covariance means a positive (linear) association
- A negative covariance means a negative (linear) association
- A covariance of zero means no (linear) association

**A problem with covariance?**

- It is difficult to interpret. Which value represents strong or weak association?
- It depends on the scale of the data.

# Correlation

**Basic idea**:

- We need to convert the covariance into standardised units.
- We need a unit that works for any scale of measurement!
- We use the **standard deviation** unit.

# Correlation

- Correlation analysis is used to measure strength of the association (linear relationship) between two variables. Consider:
  - strength of relationship,
  - direction (+/-) of relationship,
  - statistical significance of relationship.
- No causal effect is implied (Correlation ≠ Causation (doesn't mean change in one causing change in other)).
- It is unit-free.

# Correlation

**Toffee Sweets Example:**

The standard deviation of Adverts watched ($SD_y$) is 1.67 and the standard deviation for Packets bought ($SD_x$) is 2.92. If the covariance between Adverts watched and Packets bought is 4.25, what is the correlation between Adverts watched and Packets bought?

- Hint: $r = \frac{Cov_{x,y}}{SD_x SD_y}$

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

**r is known as Pearson correlation**

# Correlation: methods

- Pearson correlation coefficient is a ratio between the covariance of two variables and the product of the standard deviation. It is only meaningful for continuous data. It also assumes normality.

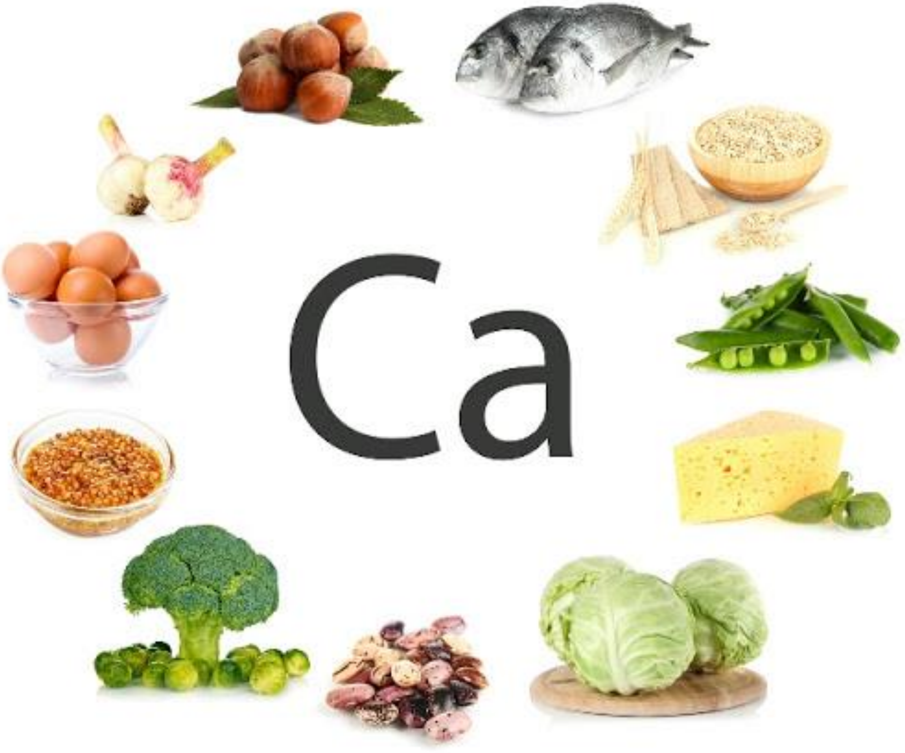
- Spearman rank correlation coefficient is a nonparametric method. It replaces the actual data by ranks (rank done separately for the variables) and then applies Pearson correlation on the ranks.

- Kendal tau is another non-parametric method preferred when there are many ties in the data.

- Pearson correlation is rather more common.

# Correlation

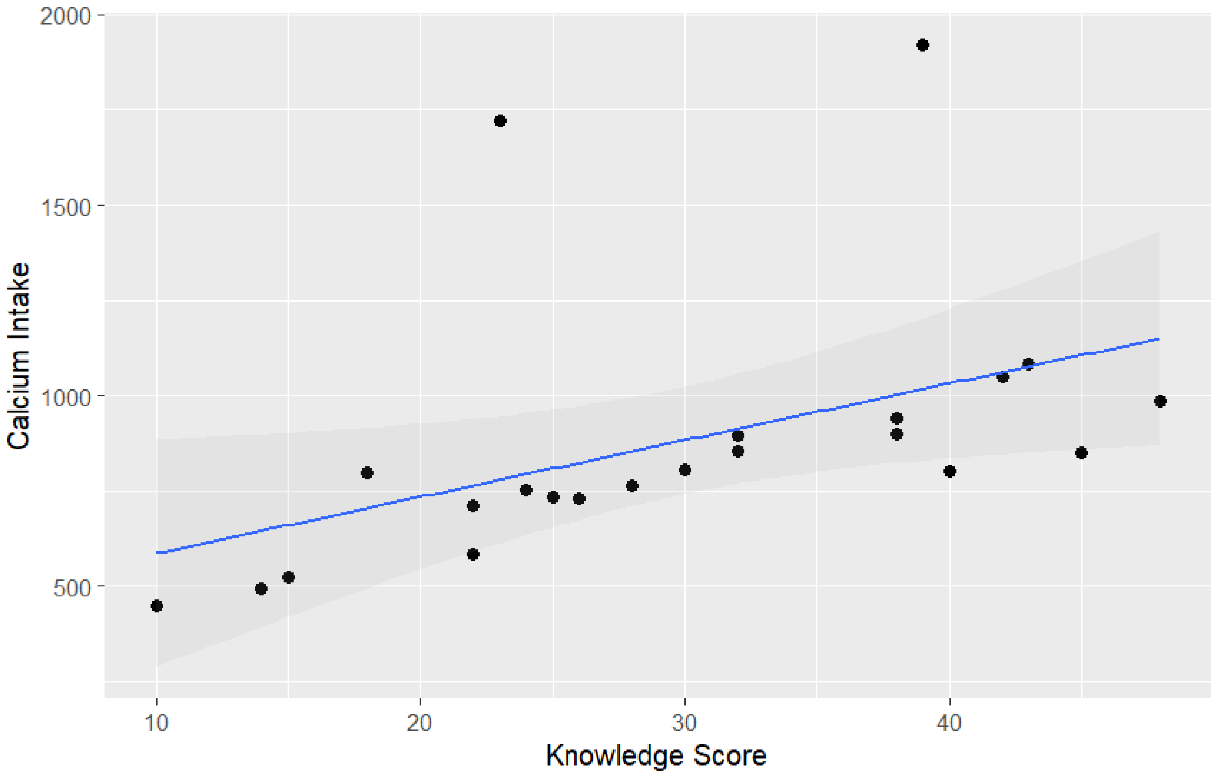
A student wanted to look at the relationship between calcium intake and knowledge about calcium in sports science students. Logically, there must be a correlation! Let’s see



# Correlation

**What we can see?**

- Outliers! Not in scores but in calcium.
- Middle scores (not so high not so less) have clearer relationship with calcium intake?
- A positive linear relationship!



Durham University

# Correlation

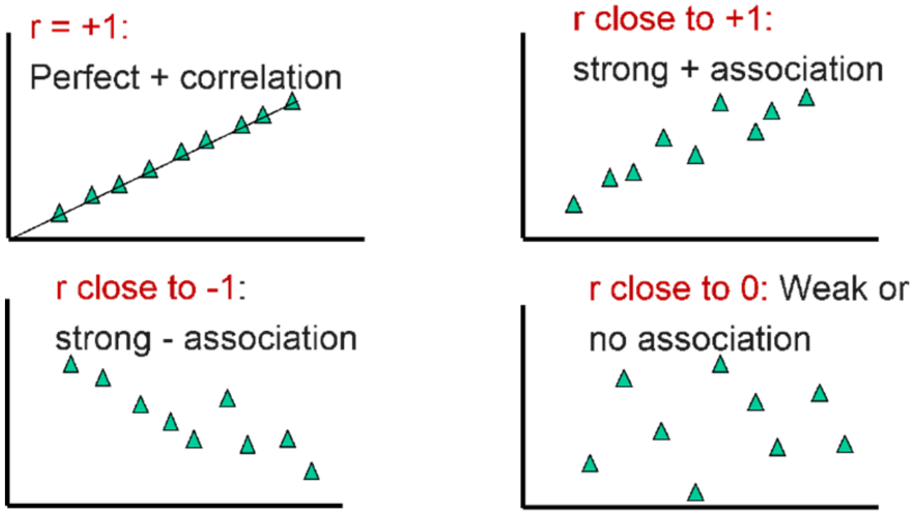
| | Covariance | SD(Scores) | SD(calcium) | Correlation |
|---|---|---|---|---|
| With outlier | 6967.2 | 43.3 | 348.5 | 0.5 |
| Without outlier | 6813.4 | 44.3 | 174.4 | 0.9 |

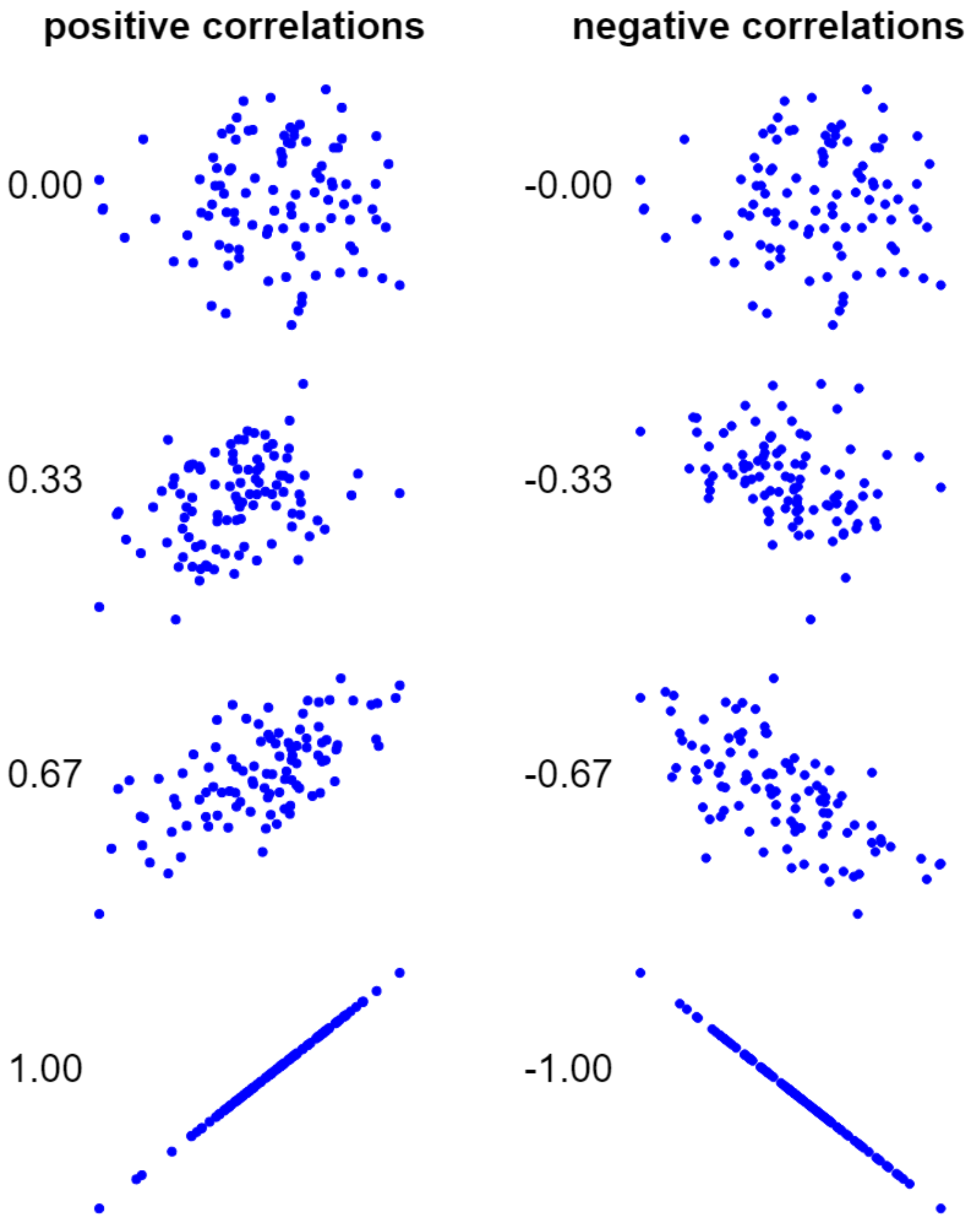
**Do outliers matter when calculating correlation between variables?**

# Correlation: summary

Applicable to two continuous variables

- It can take values between -1 to +1
- Assumes normally distributed data (Pearson)
- Strong correlation (r > 0.7)
- Moderate correlation (between 0.3 & 0.7)
- Weak correlation (r < 0.3)





# Correlation to regression

- Correlation can be very useful, but we can take this process a step further and predict one variable from another.

- Correlation tells you if there is an association between x and y but it doesn’t describe the relationship or allow you to predict one variable from the other.

- To do this we need REGRESSION! - **A simple regression analysis is a way of predicting an outcome variable (Y) from one predictor variable (X).**

# SIMPLE LINEAR REGRESSION

# Regression Model Formulation

The simple linear regression model is

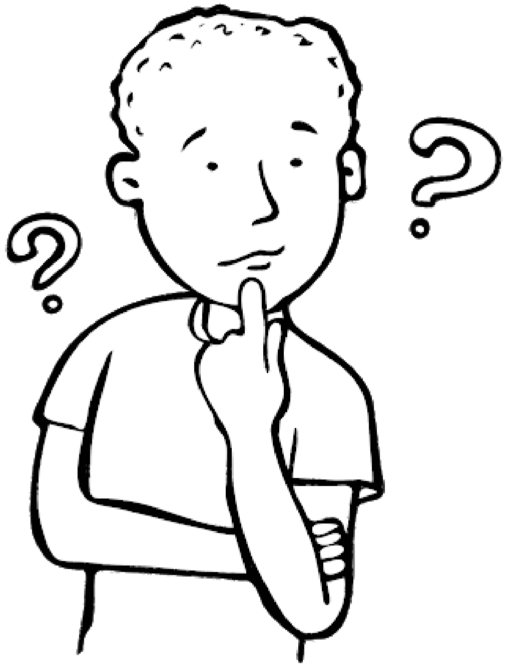
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where

- $Y_i$ is the outcome/response of $i^{th}$ person/unit that we want to predict
- $x_i$ is the $i^{th}$ person/unit value on the predictor variable (also known as explanatory or independent variable)
- $\beta_1$ is the slope; $\beta_0$ is the intercept. In practice, these are estimated by values $b_1$ and $b_0$ respectively. They are also known as regression coefficients
- The error term $\varepsilon_i$ represents the discrepancy between the outcome and line for the $i^{th}$ unit. It is also called a residual.

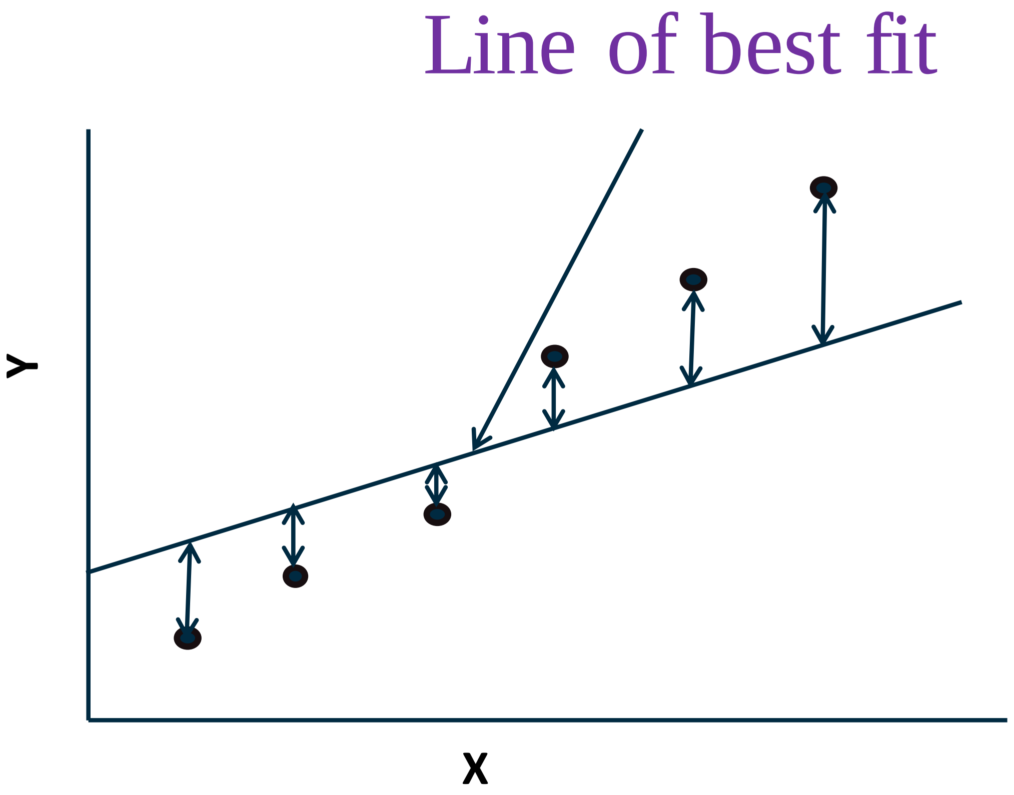
# FITTING REGRESSION MODEL

How do I fit a straight line to my data?



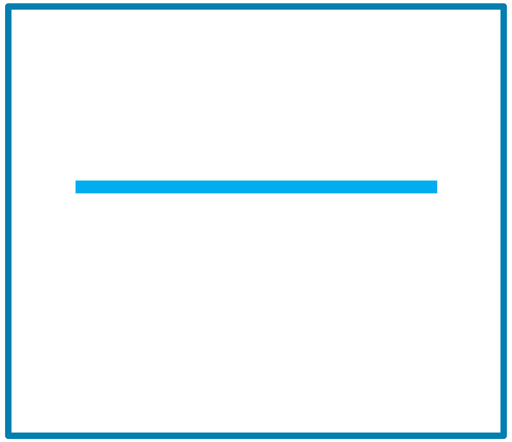
# Method of least squares

- Method of least squares is a way of finding the line that best fits the data
- We minimise the sum of squared differences between the line and the outcomes
- That is, we minimise the Error Sum of Squares (SSE)



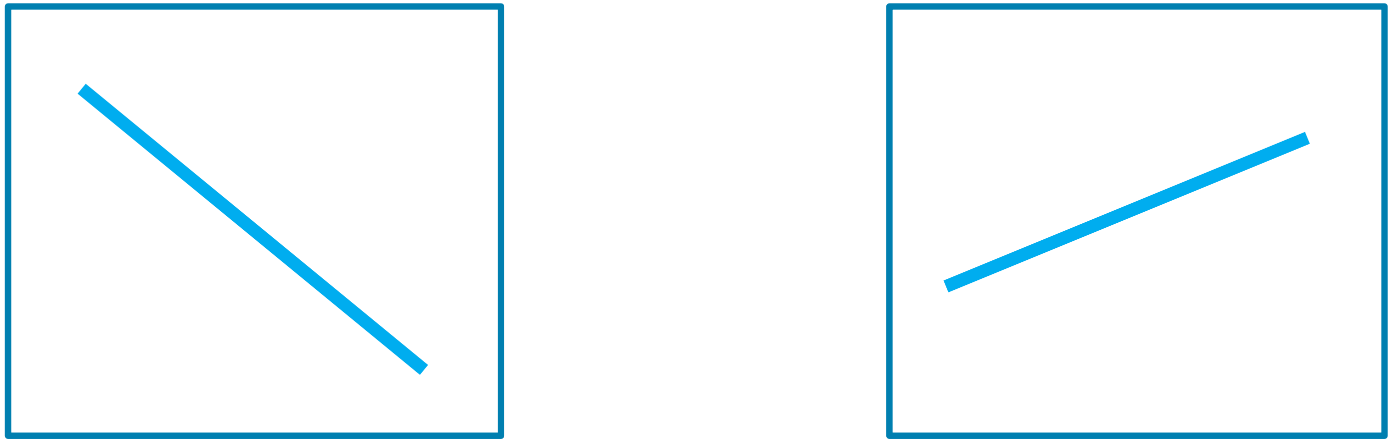
# Hypothesis Testing

**Null hypothesis:** there is no association between the outcome and the predictor. i.e., regression (slope) coefficient equals to zero



# Hypothesis Testing

**Alternative Hypothesis:** There is an association between the outcome variable and the predictor. i.e., regression (slope) coefficient not equal to zero.



# t-statistic and p-values

- The t-statistic (and its associated p-value) tests whether or not there is a statistically significant relationship between a given predictor and the outcome variable
- In other words, whether or not the regression coefficient is significantly different from zero.

# t-statistic and p-values

- Mathematically, for a given regression coefficient (b), the t-statistic is computed as t = (b - 0)/SE(b), where SE(b) is the standard error of the estimator associated with b.

- A bigger t-statistic will produce a small p-value.

- A small p-value implies that observed test statistic is not likely to occur by chance.

# INTERPRETATION

# Example 1 – Album sales data

A company believes that there is a linear relationship between sales of record albums (in thousands) and the amount spent in advertisements (in thousands of pounds).

| Brief snapshot of data – Total sample is 200 observations | |
|---|---|
| Advert spending (thousands of £) | sales (thousands) |
| 10 | 330 |
| 986 | 120 |
| 1446 | 360 |
| 1188 | 270 |
| 575 | 220 |
| 569 | 170 |
| 472 | 70 |
| 537 | 210 |
| 514 | 200 |

# Example 1 – Album sales data

The statistical hypotheses in this example:

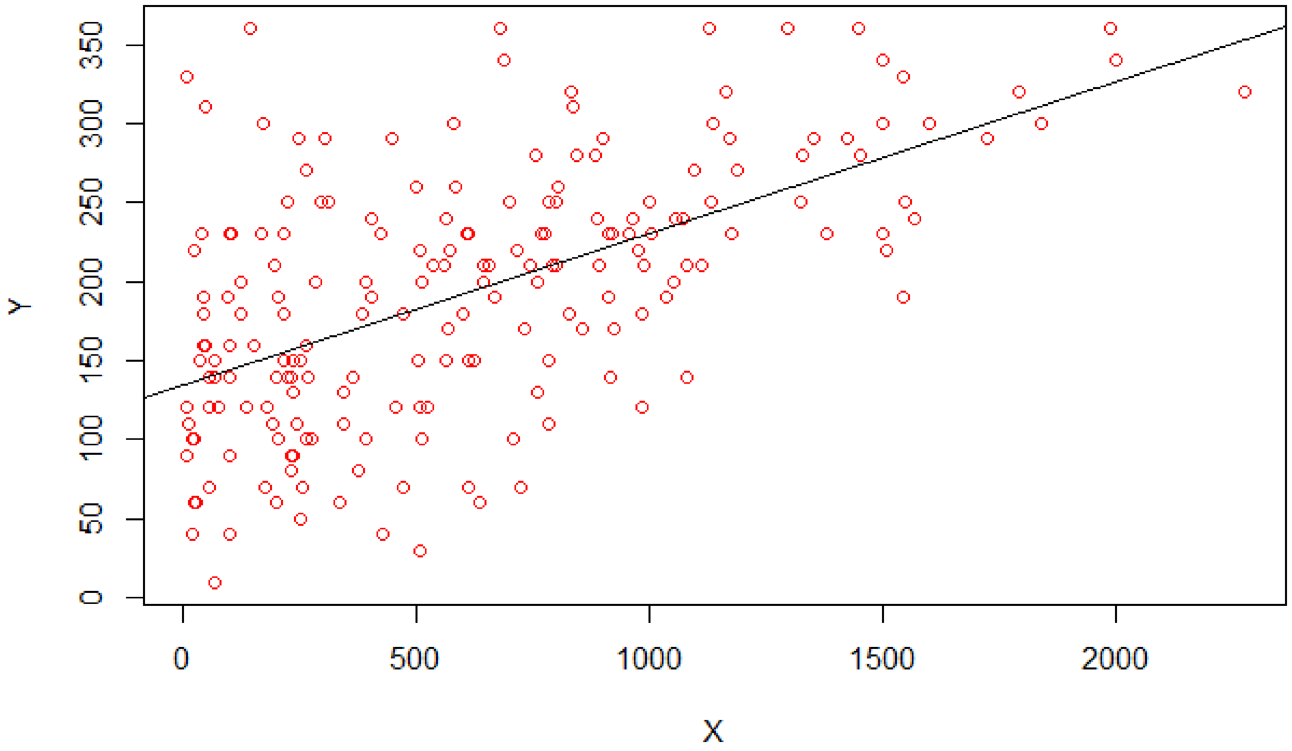
**Null hypothesis**:

There is no association between sales of record albums and the amount spent in advertisements.

**Alternative hypothesis**:

Sales of record albums and the amount spent in advertisements are associated.

# Example 1 – Album sales data

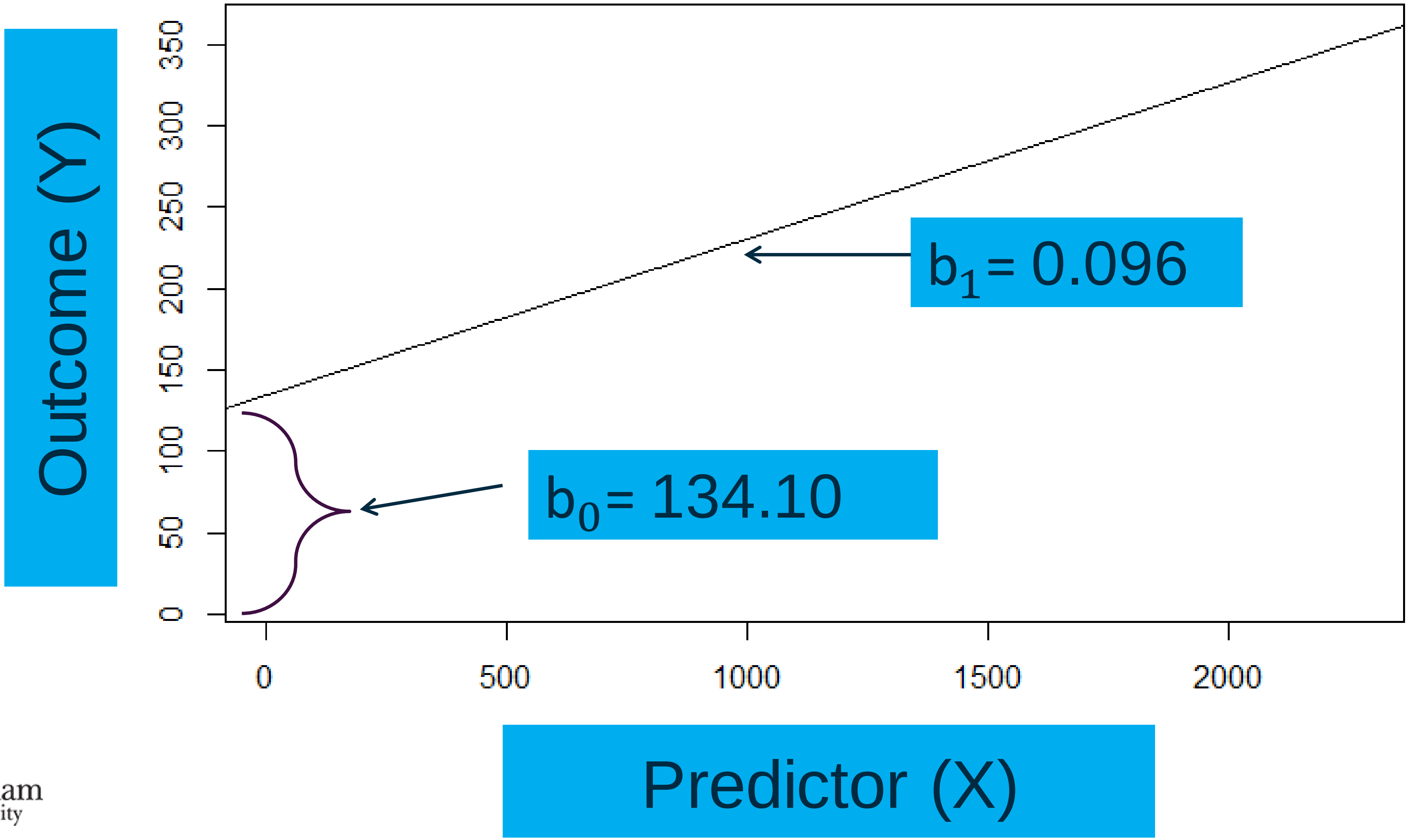


Least squares method gives the ‘line of best fit’

Record sales of albums (per thousands) is our dependent variable and the amount spent in advertisements (per thousand pounds) is the predictor

| | Estimate | Standard error | t value | Pr(>\|t\|) | 95% CI | |
|---|---|---|---|---|---|---|
| | | | | | LB | UB |
| Intercept ($\mathbf{b_0}$) | 134.10 | 7.54 | 17.80 | 0.00*** | 119.28 | 149.00 |
| adverts($\mathbf{b_1}$) | 0.096 | 0.00962 | 9.98 | 0.00*** | 0.08 | 0.12 |

$\widehat{sales}_i$ =134.10 + 0.096 x $advertisement\ budget_i$



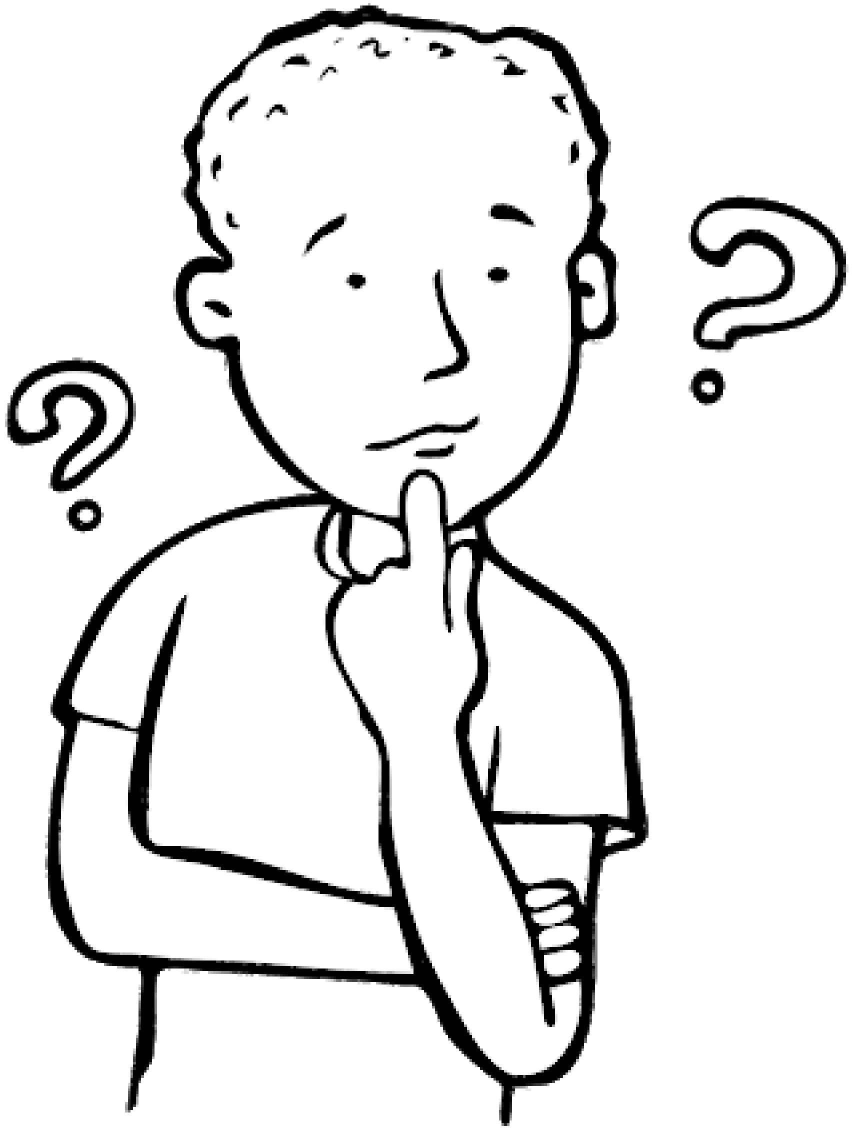
$b_1 = 0.096$

$b_0 = 134.10$

Record sales of albums (per thousands) is our dependent variable and the amount spent in advertisements (thousand pounds) is the predictor

| | Estimate | Standard error | t value | Pr(>\|t\|) | 95% CI | |
|---|---|---|---|---|---|---|
| | | | | | LB | UB |
| Intercept ($\mathbf{b_0}$) | 134.10 | 7.54 | 17.80 | 0.00*** | 119.28 | 149.00 |
| adverts($\mathbf{b_1}$) | 0.096 | 0.00962 | 9.98 | 0.00*** | 0.08 | 0.12 |

t-statistic = (b - 0)/SE(b) i.e. t=(0.096-0)/0.00962 =9.98



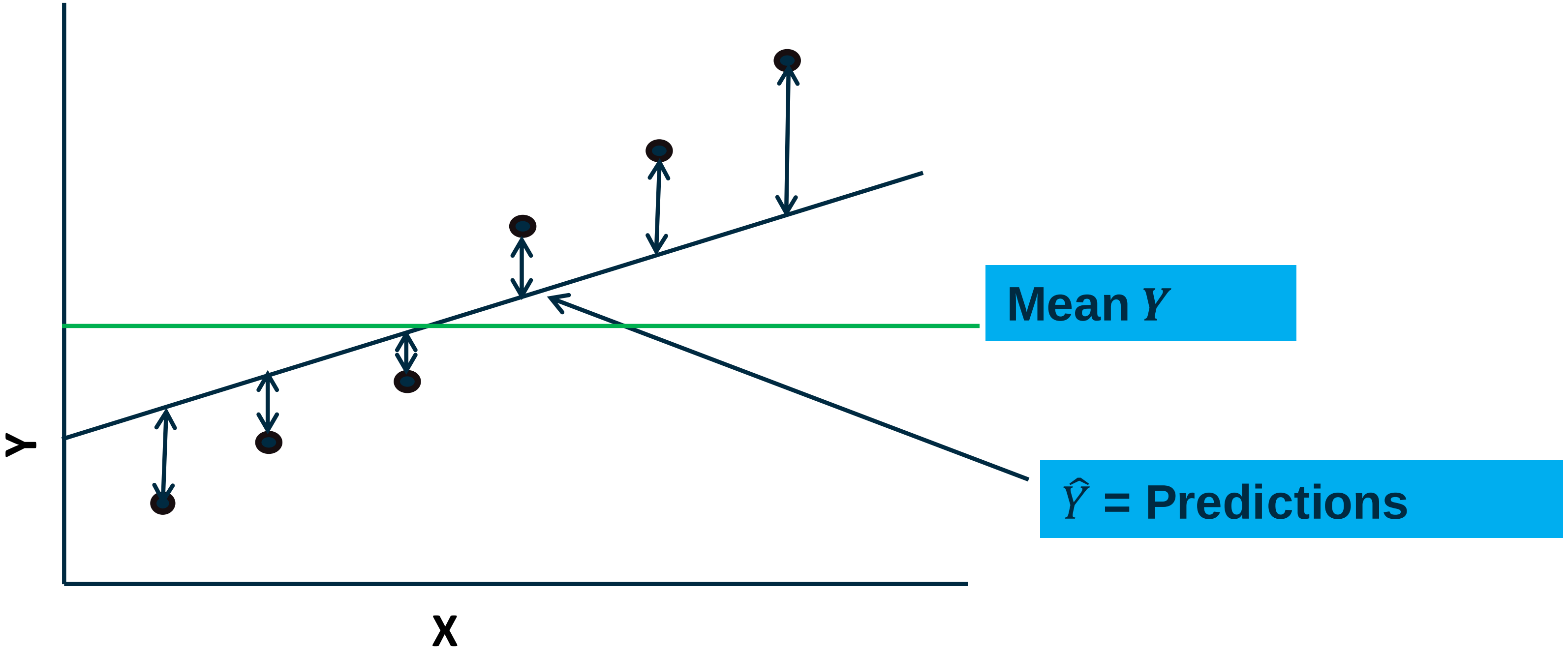
How do I tell if my model is useful?

# Sum of Squares

- The regression line minimises the distance between observed & predicted values.

- Error Sum of Squares (SSE) summarises the distance between the data and the model.

# Sum of Squares



- **Regression Sum of Squares (SSR)** is measured by squaring the differences (deviations) between the predicted *Y* value and the mean, for each data point, and adding up.
- **Error Sum of Squares (SSE):** measures the variation that remained unexplained.

# Sum of Squares

Error Sum of Squares

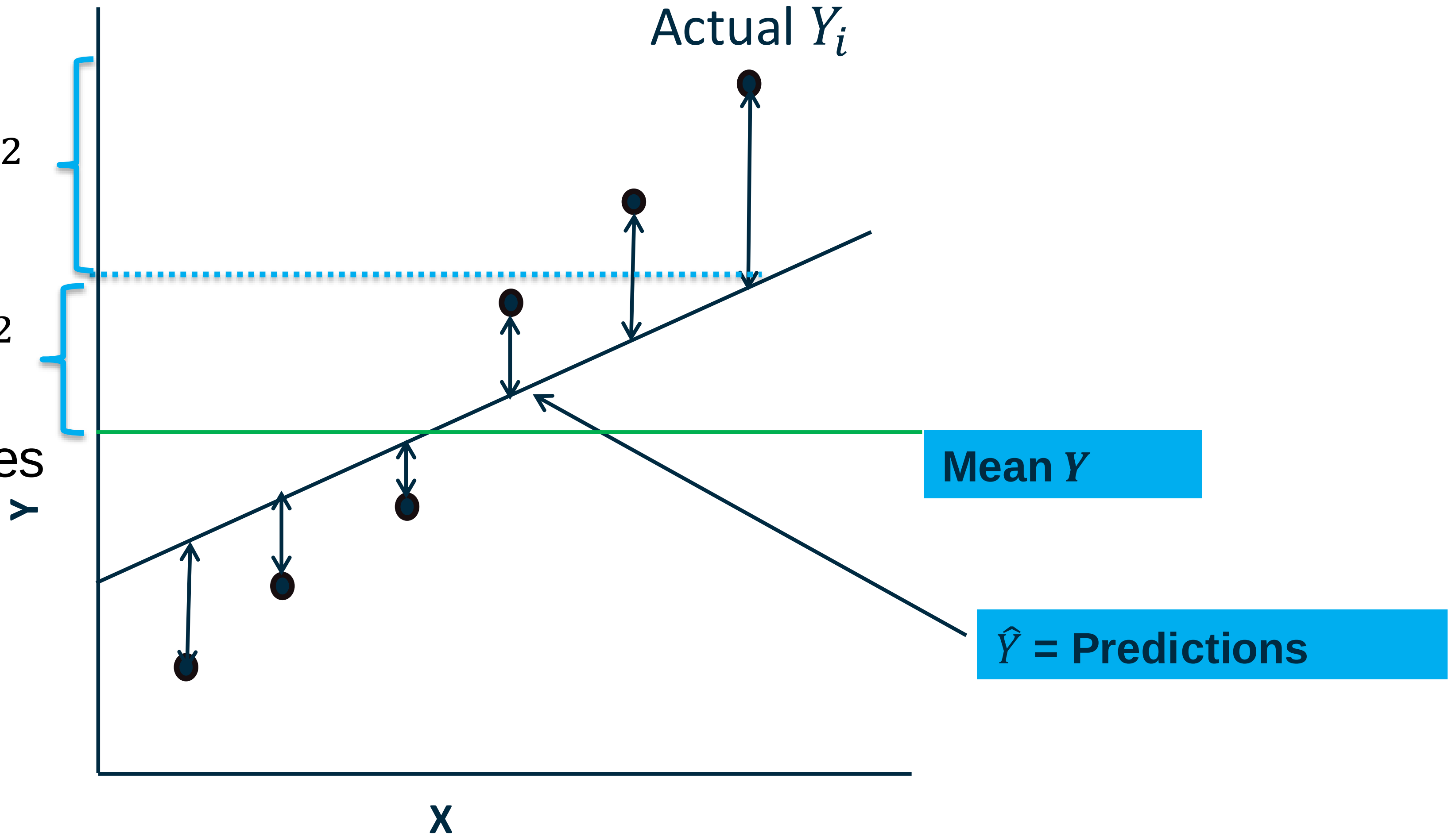
SSE=$\sum(Y_i - Y_{fitted})^2$

SSR=$\sum(Y_{fitted} - Y_{mean})^2$

Regression Sum of Squares

Total Sum of Squares

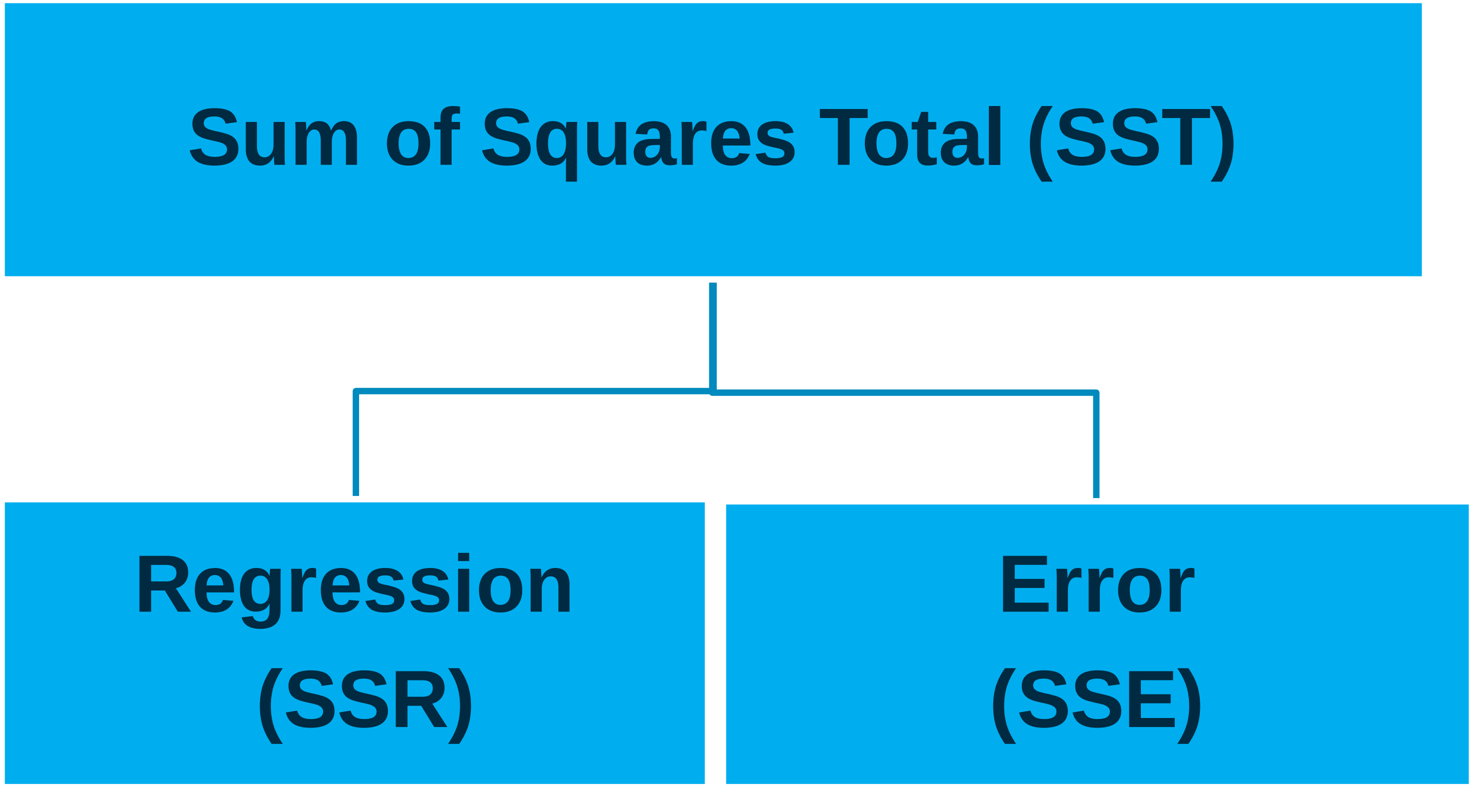
SST=$\sum(Y_i - Y_{mean})^2$



# Sum of Squares

**Total Sum of Squares SST = SSR + SSE**

- **SSR = Regression Sum of Squares**



Sum of Squares Total (SST)

Regression (SSR)

Error (SSE)

# Coefficient of Determination - $R^2$

$R^2$ represents the amount of variance in the outcome explained by the predictors (independent variables) in the regression model. In other words, $R^2$ tells how well the regression model fits the data.

**$R^2$**

$$\frac{SSR}{SST} = 1- \frac{SSE}{SST}$$

$R^2$ range between 0 and 1

Note: In the single predictor variable case, $R^2 = r^2$

|  | Beta estimate | Standard error | t value | Pr(>\|t\|) | 95% CI |  |
|---|---|---|---|---|---|---|
|  |  |  |  |  | LB | UB |
| Intercept ($\mathbf{b_0}$) | 134.10 | 7.54 | 17.80 | 0.00*** | 119.28 | 149.00 |
| adverts($\mathbf{b_1}$) | 0.096 | 0.00962 | 9.98 | 0.00*** | 0.08 | 0.12 |

$R^2$= 0.3346. Are you convinced of the results?

# Regression with categorical predictors

The values of a categorical variable indicate group membership of observations rather than a quantitative measurement

Examples:

- **Binary** – e.g. smoker; “Yes”, “No”
- **Ordered** – e.g. obesity; “underweight”, “normal”, “overweight”, “obese”
- **Unordered** – e.g. region; “Bristol”, “London” and “Stoke”

# Regression with categorical predictors

We can also create categorical variables from continuous variables by grouping values or using cut-offs.

For example:

- Classify Body Mass Index (BMI) into clinically relevant groups
  - Underweight: $\text{BMI} < 20$
  - Normal: $20 \leq \text{BMI} < 25$
  - Overweight: $25 \leq \text{BMI} < 30$
  - Obese: $\text{BMI} \geq 30$

# Regression with categorical predictors: Indicator variables

A binary variable is a special type of categorical variable called an indicator variable taking the values of 0 and 1

Including indicator variables in a regression will model a difference in means between groups

Indicator variables can also be used to represent categorical variables which have more than two categories

# Regression with categorical predictors: Indicator variables

When a categorical variable has more than two categories effects are estimated by introducing a series of indicator variables

First we choose a reference group (usually the lowest coded value of the variable) to which the other groups will be compared

If a variable has k levels then k-1 indicators are included

The regression coefficient for each indicator variable is the difference in the mean outcome for that group compared to the reference group

# Assumptions

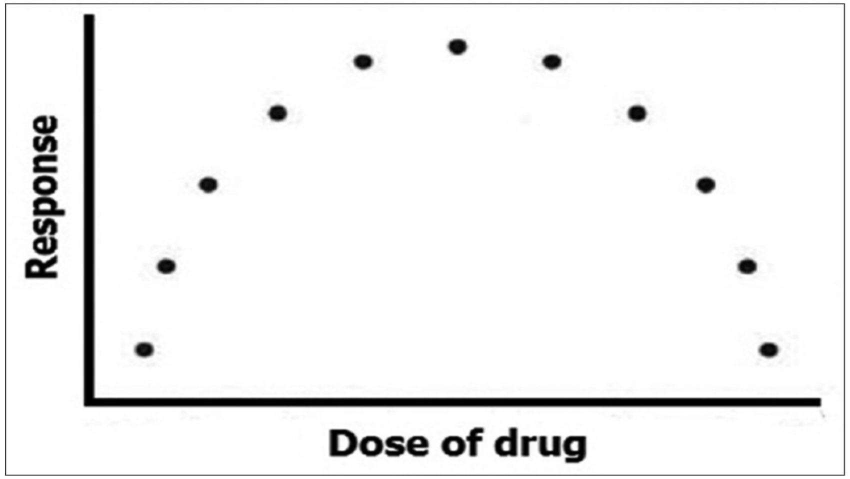
There are four basic assumptions for every linear model with a continuous outcome. There are different views on the importance of the assumptions; our view is that at this level all the assumptions should be considered as equally important.

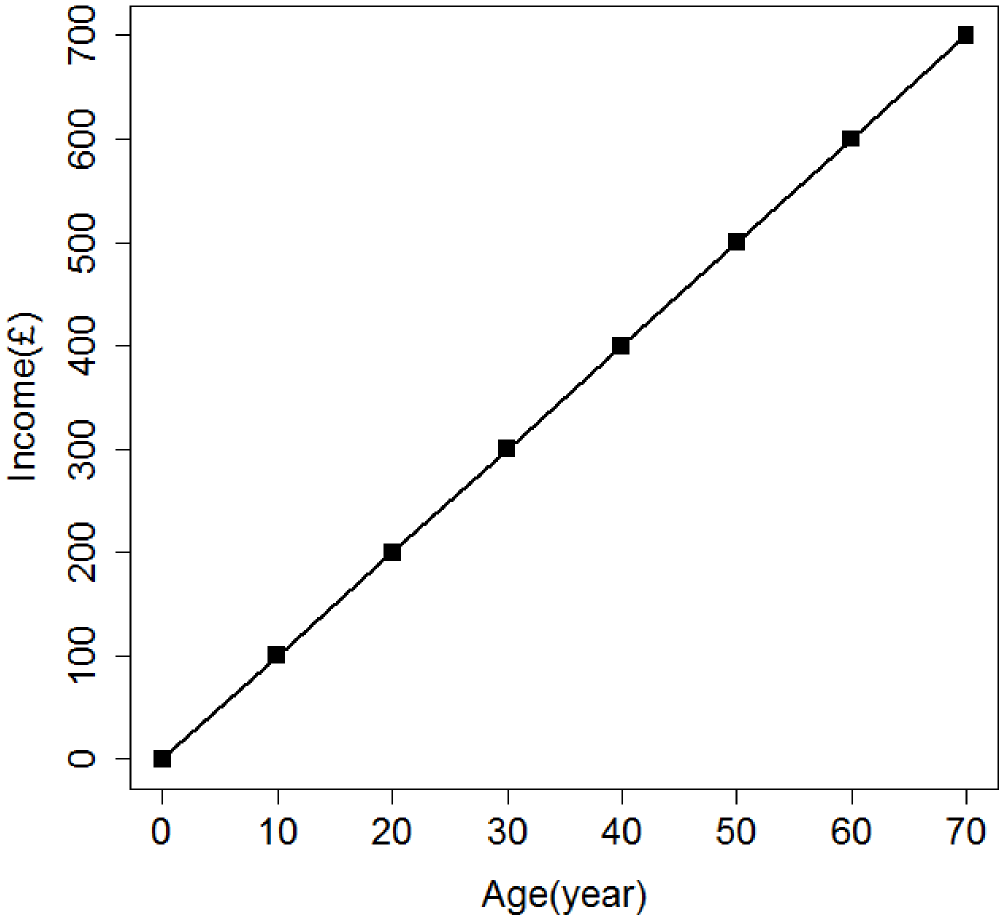
**Outcome variable:** MUST be continuous

**Predictor or independent variable:**
Assume continuous for now.

# Basic Assumptions

**Linearity:**
The relationship between the outcome variable and the predictor is **linear**.





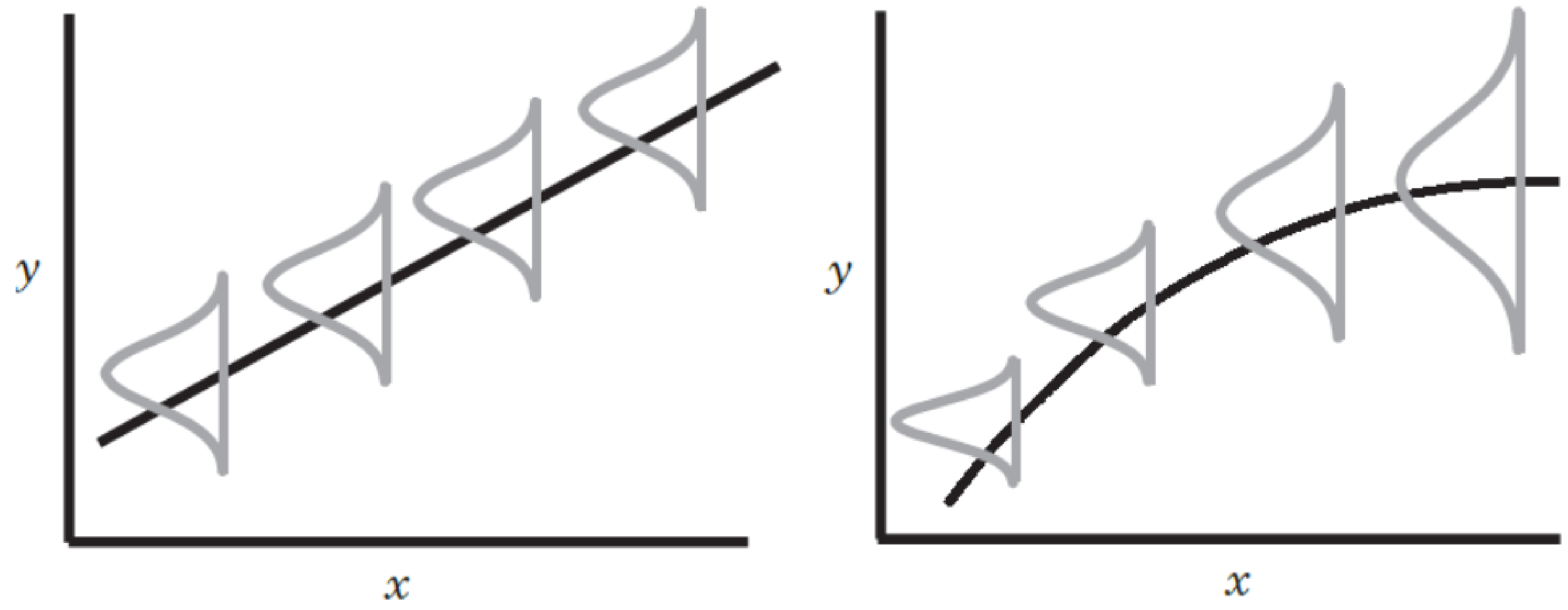
# Basic Assumptions

**Normally distributed errors:**
It is assumed that the residuals in the model are normally distributed.

**Independent errors:**
For any two observations the residual terms should be independent.

# Basic Assumptions

**Homogeneity:** The variation of the residuals at any level of the predictor is the same. Variance of residuals is constant for all cases in the model. This assumption assumes all cases have the same variance.



Which data are homogenous?

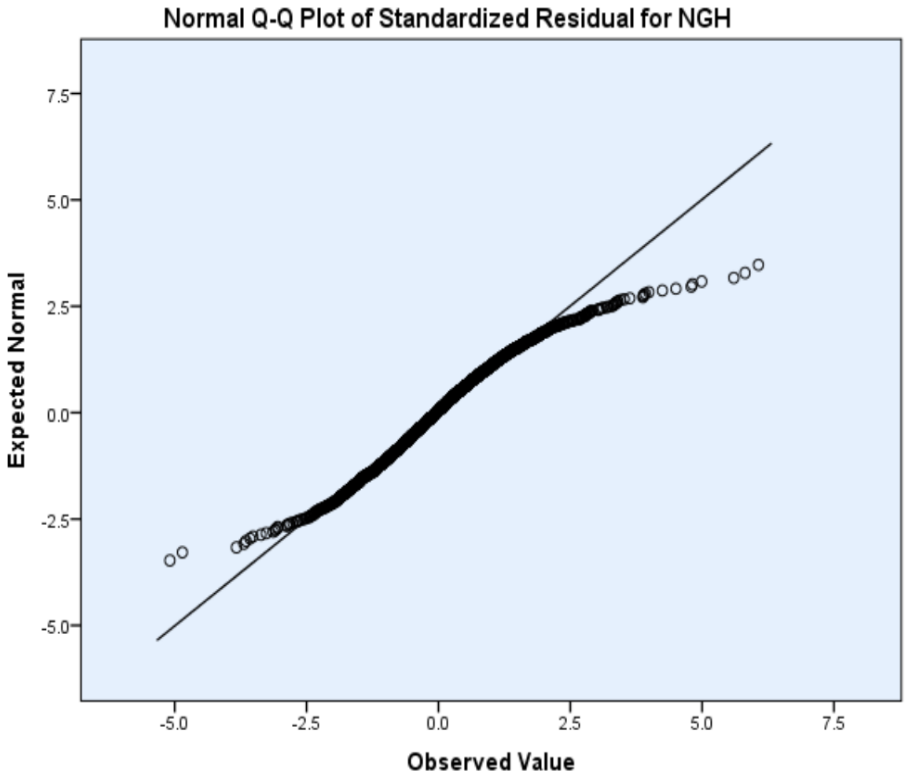
# Diagnostics for Normality of residuals:

- Use quantile-quantile plot (q-q plot) to compare the distribution of the residuals from the model with a standard Normal distribution.

- If the residuals are Normally distributed, then the expected and observed quantiles will match (roughly)

- Deviation from the diagonal, especially at the tails may be an indication for the violation of the Normality assumption.

- Formal testing can be done using either Kolmogorov-Smirnov test or Shapiro-Wilk tests. Both tests are very sensitive to outliers, especially in large samples.



# Diagnostics for Constant variance of residuals:

- We recommend to always look at the plot of the residuals against predicted values. If the assumption of constant variance is satisfied, the scatter plots should be random without any pattern. One may also look at the plot of the residuals and the actual outcome data.

- A formal test using Levene's test can also be performed to test for constancy of variance.

# Diagnostic for Independence of residuals:

- Checking independence of the error terms is not straightforward. If this assumption is violated, it would usually imply that the residuals are not independent.
- One possibility is to check trend in the residuals.

# Diagnostic for Linearity:

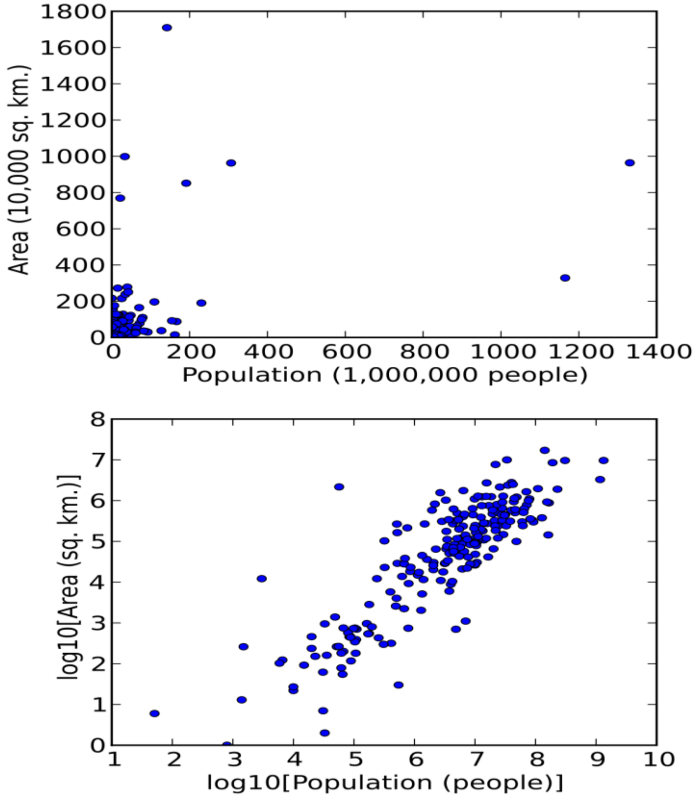
- Linearity can be explored by looking at the plot of the residuals and the suspicious continuous predictor.

Formal testing can be performed by using LIKELIHOOD RATIO TEST that examine whether higher power of a predictor is needed or not (transformation of the predictor may help here – note that it will NOT help with heteroskedastic errors).

# Fixing assumption violations

- Possible solution for the violation of Normality assumption and constant variance: transformation.

- Either logarithm or square root transformation of the outcome variable can help. However, it will mean loss of interpretation as the results would no longer be on the natural scale of the outcome variable.
- If after transformation the assumptions were still not satisfied, the model should be reported and interpreted with clear statement on the assumptions violation.



# Outliers

- It is important to understand how the model is affected by the outlying cases. Some statistics are more sensitive to outliers than others.
- Outliers can be investigated using standardised residuals (std. residuals) from the model.
- A rule of thumb is that cases with absolute standardised residuals greater than 3 may be outliers. Note that a standardised residual is the residual from the model divided by its standard deviation.

# Overview of simple regression

- Simple regression model is linear relationship between outcome and predictor.
- Slope (Gradient): Corresponds to the change in the Outcome (dependent) variable for a unit change in the Predictor
- Positive slope of the regression line means a positive association.
- Negative slope of the regression line means a negative association.
- When slope is zero, it means there is no association between the outcome and the predictor variables.
- The unit of the dependent and predictor are not required to be the same.
- Fitted regression equation can predict values of outcome (Y) for any value of predictor (X).

Thanks for your attention!

Durham University
Research Methods Centre