Correlation and Regression

27th February 2025

Session Plan

- Correlation analysis
- Linear regression – how to find the line of best fit
- Interpreting regression output
- Assumptions and limitations of regression

Tutorial Session

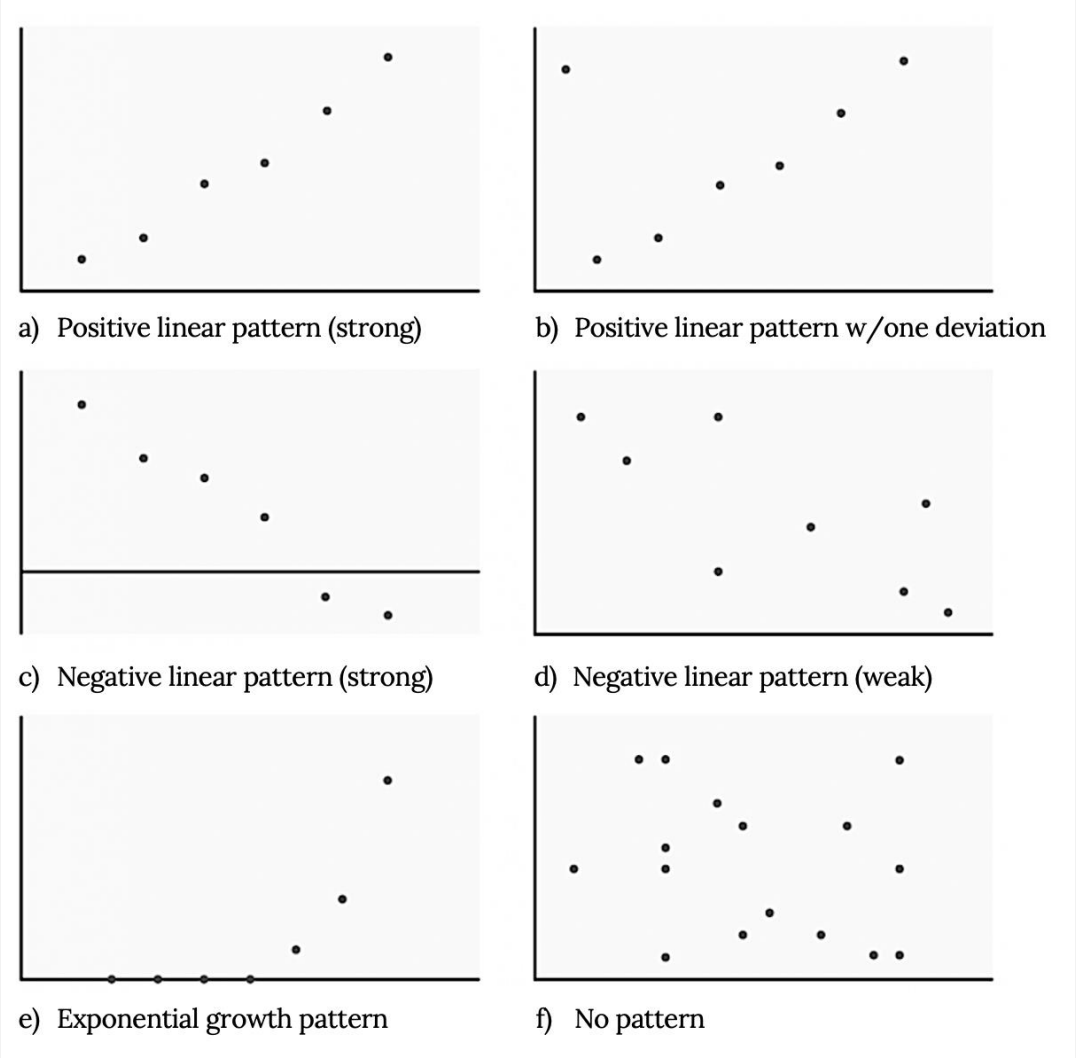
- Discussion session 1 for the formative assessment

Reminders 1: scatter plots

- Once you have plotted your data, you can start to consider patterns and relationships.

- You can describe the
  - Trend (+/-)
  - Shape
  - Strength

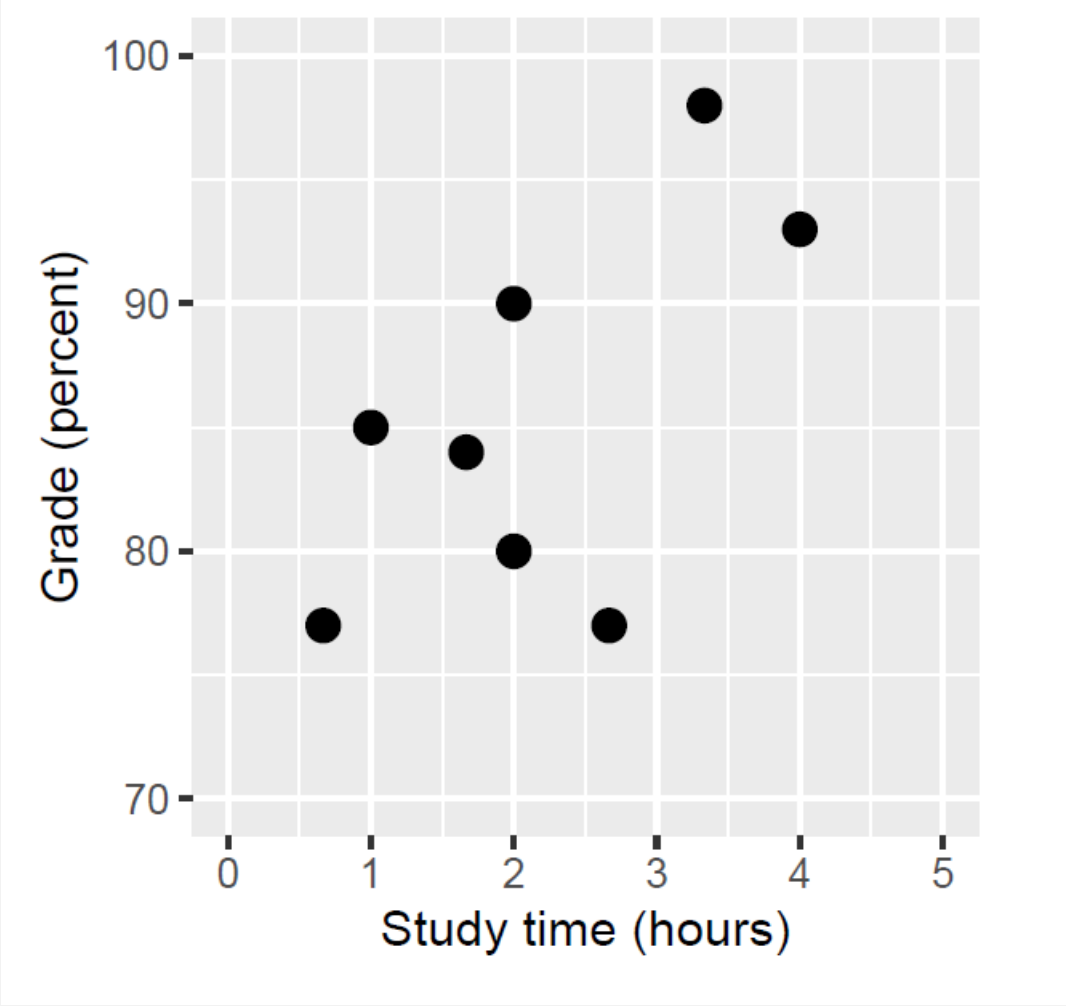
  of relationships between the two data series



Reminders 2: Some ideas

- Our general goal is to find a model that minimises error (s.t. simplicity, generalisability etc)
- You've just done a task of model development by drawing those lines.
- A reminder of 2 terms:
  - Dependent variable (**y**) – the outcome variable that our model seeks to explain
  - Independent variable (**x**) – the variable we want to use to explain the dependent variable

We saw in week 2 how scatter plots can show a relationship between variables



Describe this chart: What is the

- Trend (+/-)
- Strength
- Shape
  - Linear: follows a straight line
  - Non-linear: follows a curved pattern

of the fit?

Is there a statistically significant relationship between grade and study time?

What grade should I expect if I study for 3 hours?

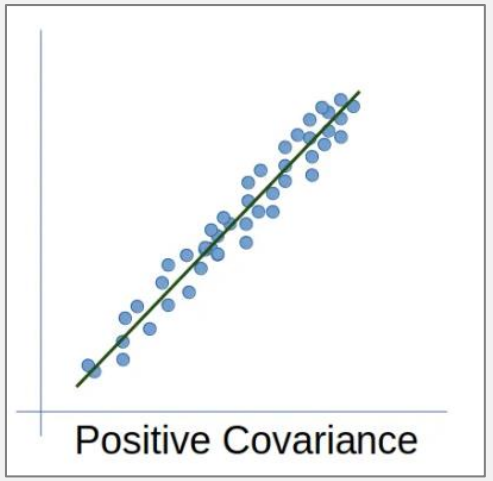
Remember, data = model + error

Covariance

- Another way to describe the **linear** relationship between 2 variables is through the **covariance** (refer to week 2!)

$$covariance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

- The covariance tells us whether there is a relation between the deviations of two different linear variables across observations
- If **covariance is positive**, we expect both sets of variables to increase together
- If it is **negative**, one variable will decrease while the other increases.
- The further the number is from zero, the deviations from their respective means are similar



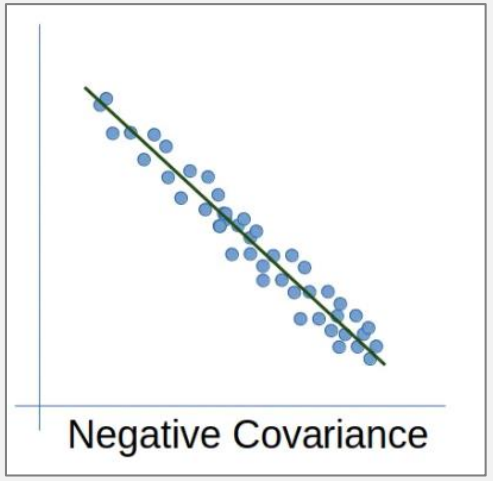
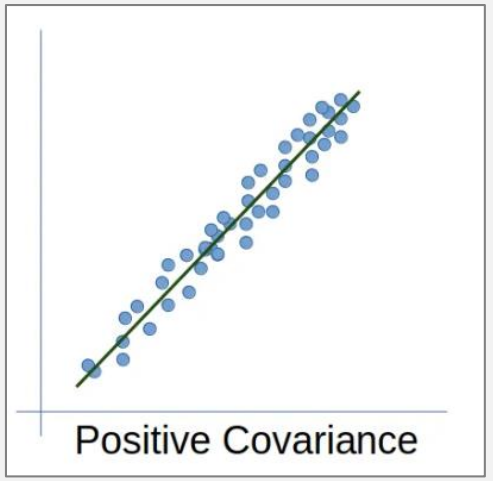


Image source

Covariance – step by step

1. Calculate the mean of each variable ($\bar{x}$ and $\bar{y}$)
2. **For each observation**, calculate the deviation from the mean ($x_{dev} = x_i - \bar{x}$)
3. Multiply these deviations to get the cross product for each observation
4. Sum all crossproducts and divide by (N-1)

$$covariance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$



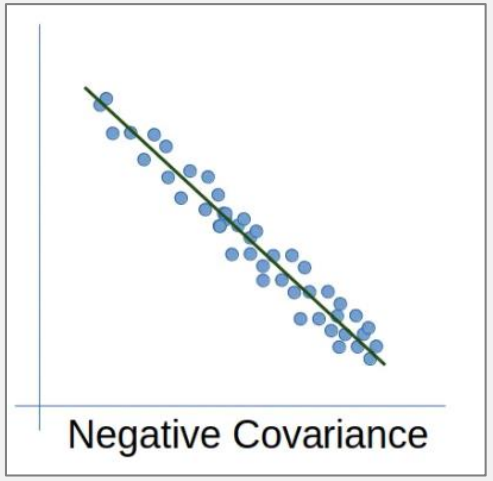


Image source

Calculate the Covariance

$$covariance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

x	y	x_dev	y_dev	crossproduct
3	5			
5	4			
8	7			
10	10			
12	17			

Covariance = 17.05

- The covariance is simply the mean of the cross-products, this means that it varies with the overall level of variance in the data.

- To address this, we usually use the *correlation coefficient* (aka *Pearson's correlation*)

- This coefficient is computed by scaling the covariance by the standard deviations of the two variables.

$$r = \frac{covariance}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

- This removes the dependence on the scale of the variables, making correlation coefficient values comparable across different pairs of variables. As it's a ratio of 2 things, the correlation coefficient is dimensionless.

Calculate the correlation coefficient

$$r = \frac{covariance}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

	x	y	x_dev	y_dev	crossproduct
	3	5	-4.6	-3.6	16.56
	5	4	-2.6	-4.6	11.96
	8	7	0.4	-1.6	-0.64
	10	10	2.4	1.4	3.36
	12	17	4.4	8.4	36.96
Mean	7.6	8.6		Covariance = 17.05	

Correlation coefficient r = 0.89

Correlation

- The correlation *coefficient* is helpful because it will always vary between -1 and +1
- There's a strong link between the correlation coefficient and the shape of the linear regression

- A correlation of ±1 indicates a perfectly linear relationship
- A correlation of zero suggests no linear relationship

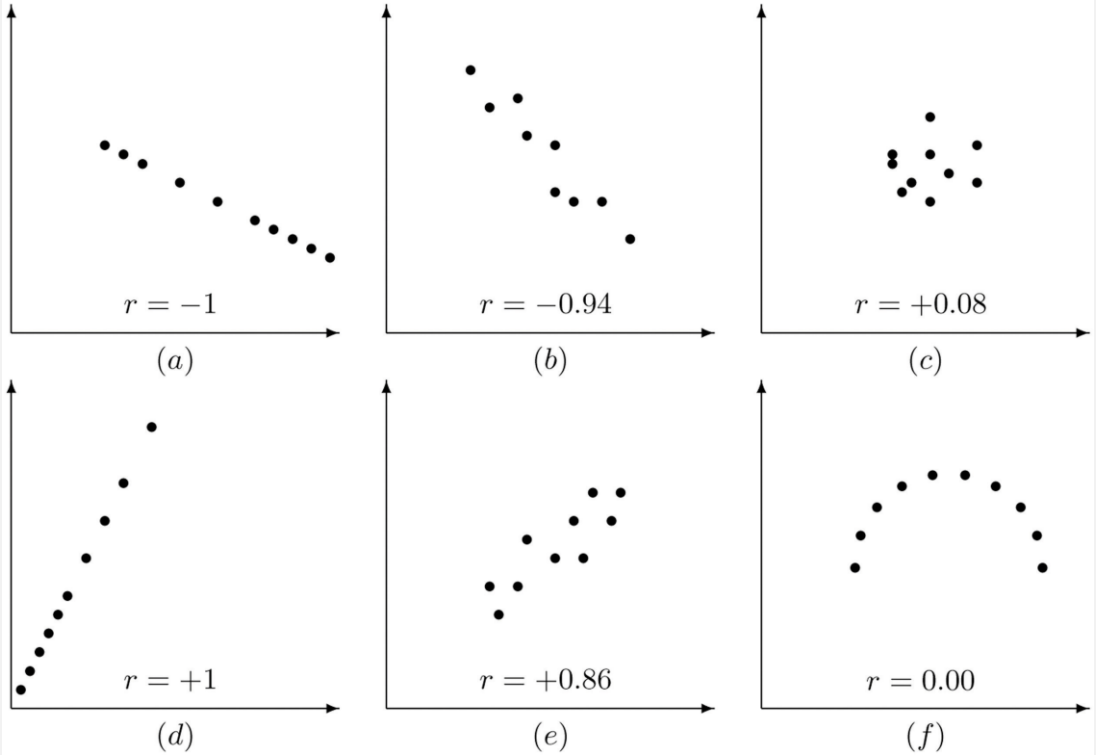


Image source

Correlation

Relationships between variables are fundamental to statistical modelling. They help us:

- Predict future outcomes based on current data
- Identify drivers of performance
- Quantify effects of interventions or strategies
- Build models for estimation and inference
- Test hypotheses about causal mechanisms.

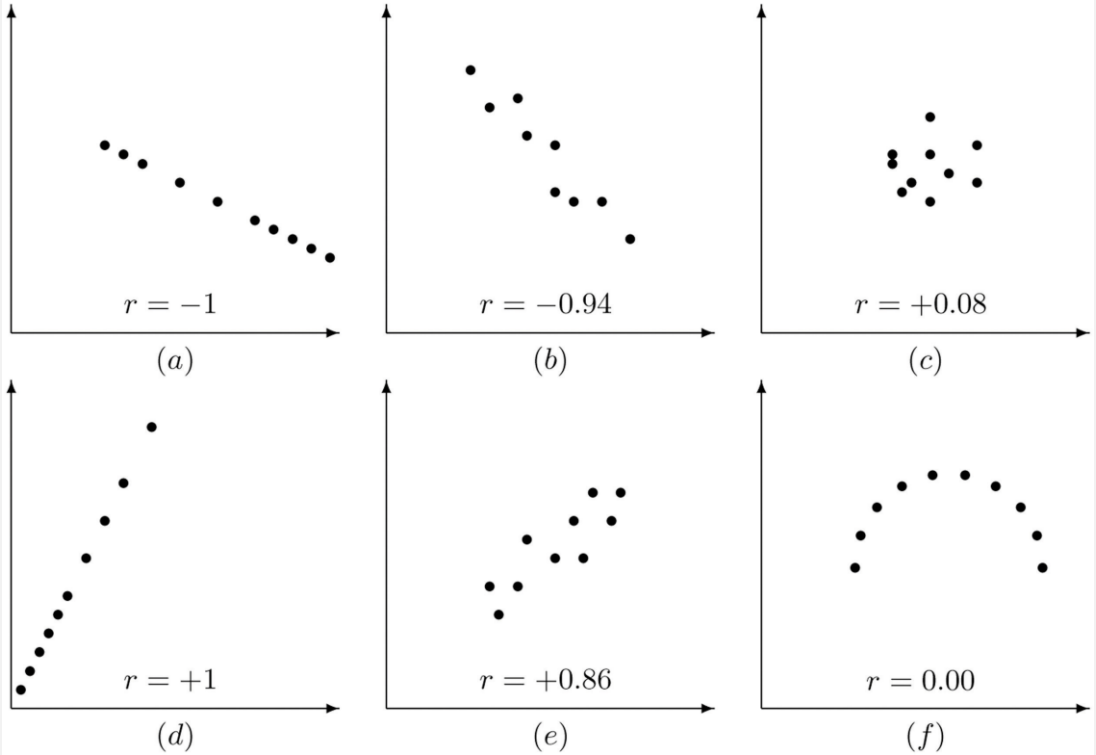


Image source

Hypothesis Testing for Correlations

How do we determine if an observed correlation is **statistically significant?**

- **Null Hypothesis** (H_0): $\rho = 0$ (no correlation in population)
- **Alternative Hypothesis** (H_A): $\rho \neq 0$ (correlation exists)
- **Test statistic:**

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- Follows a t-distribution
- Reject H_0 if p-value < α (typically 0.05)

Example: “We found a strong positive correlation between study time and grades (r = 0.89, p < 0.001), suggesting that increased study time is significantly associated with higher academic performance.”

Correlation in R

Calculating a correlation coefficient and its statistical test is simple in R:

```r
{r echo=TRUE, eval=FALSE}
# Calculate correlation in R
cor(df$studyTime, df$grades)
cor.test(df$studyTime, df$grades)

# Interpreting the output
# - cor.test() gives both the correlation coefficient
# - and performs a statistical test of H₀: r = 0
# - p-value < 0.05 suggests statistically significant correlation
```

```
[1] 0.9217702


        Pearson's product-moment correlation

data:  df$studyTime and df$grades
t = 10.086, df = 18, p-value = 7.826e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8094095 0.9690238
sample estimates:
      cor
0.9217702
```

When is correlation relevant?

Pearson's correlation coefficient (*r*) is useful when:

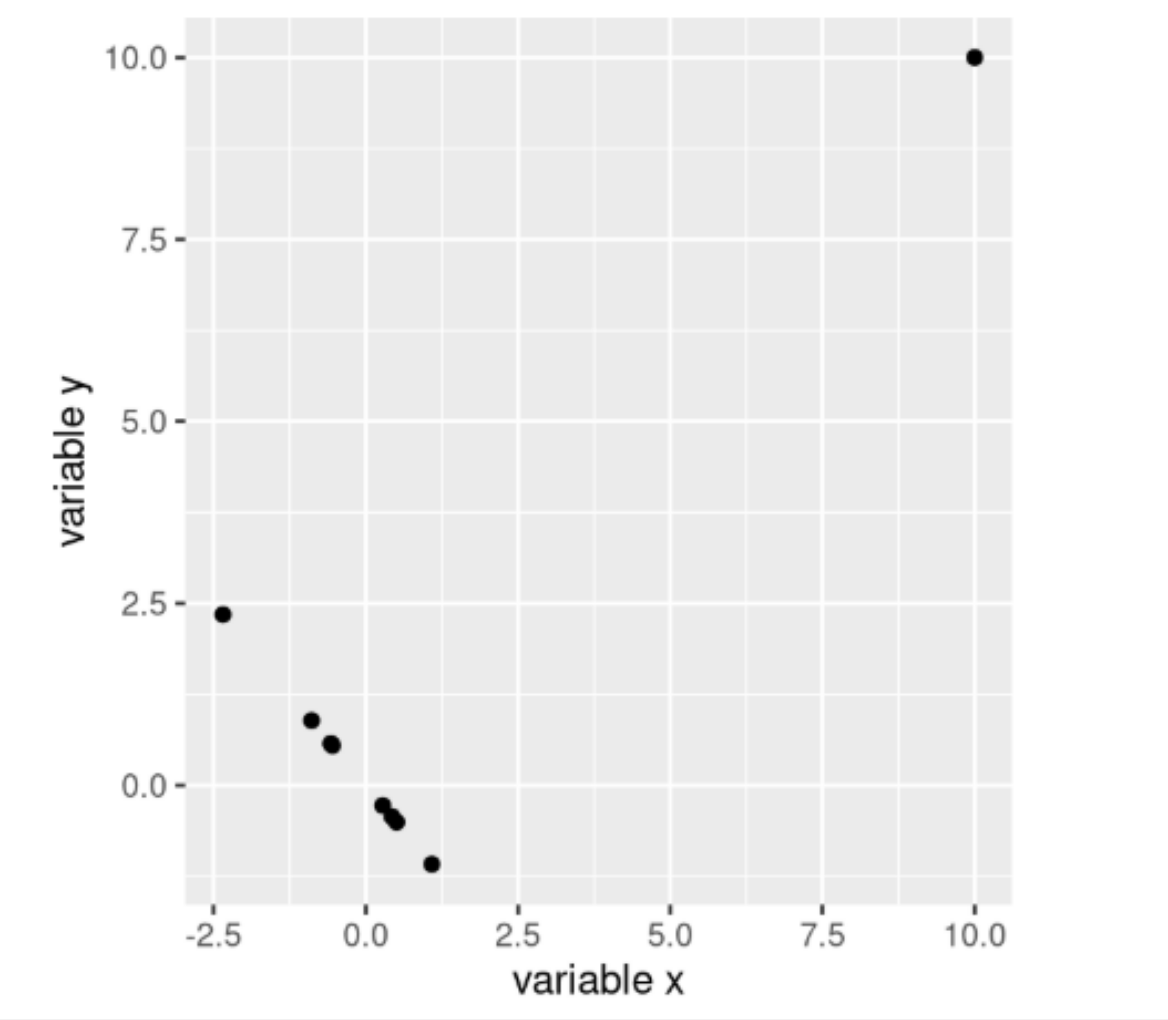
- …we want to know the relationship between two variables
- … where both data sets have an interval scale

Because Pearson's r is based on the idea of linearity, it only makes sense to use it for data that is measured on *at least* an interval scale.

Pearson's *r* is good at measuring the strength of linear associations BUT,

- It can be quite misleading in the presence of a non-linear relationships
- Extrapolation beyond the limits of observed data can be dangerous
- It's not so helpful where there are outliers …

Robust Correlations

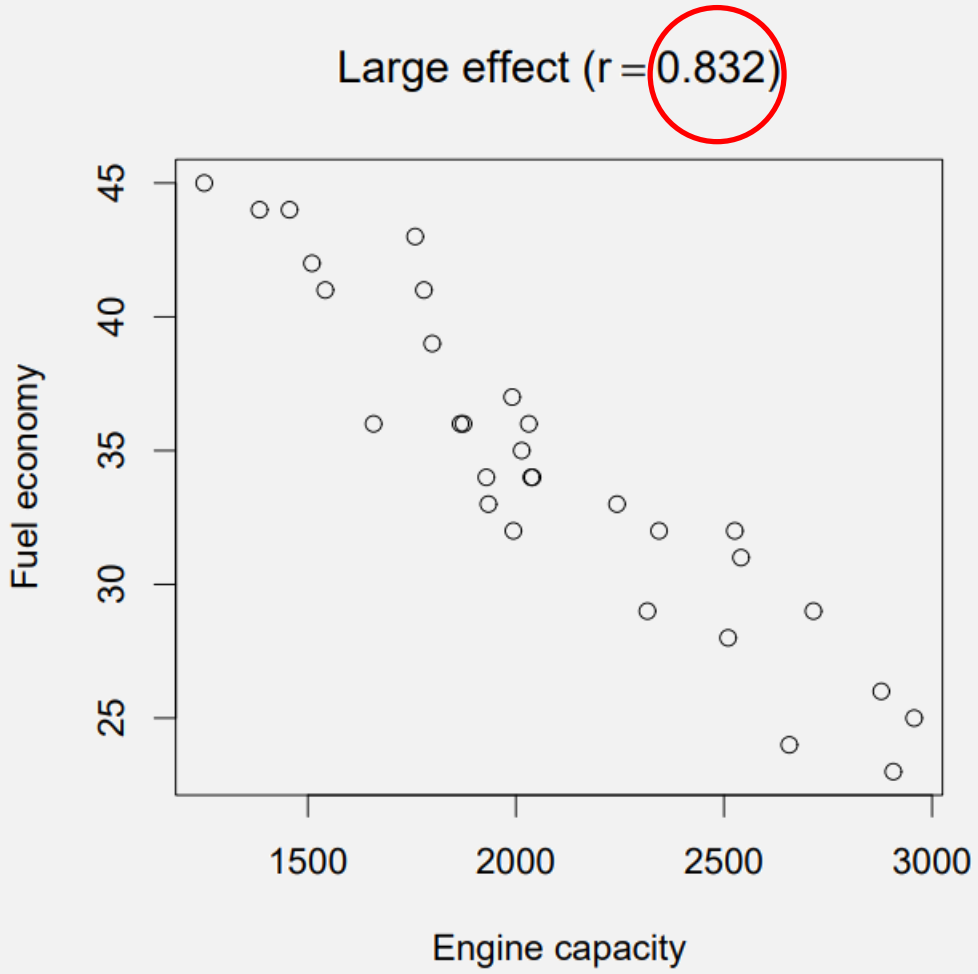


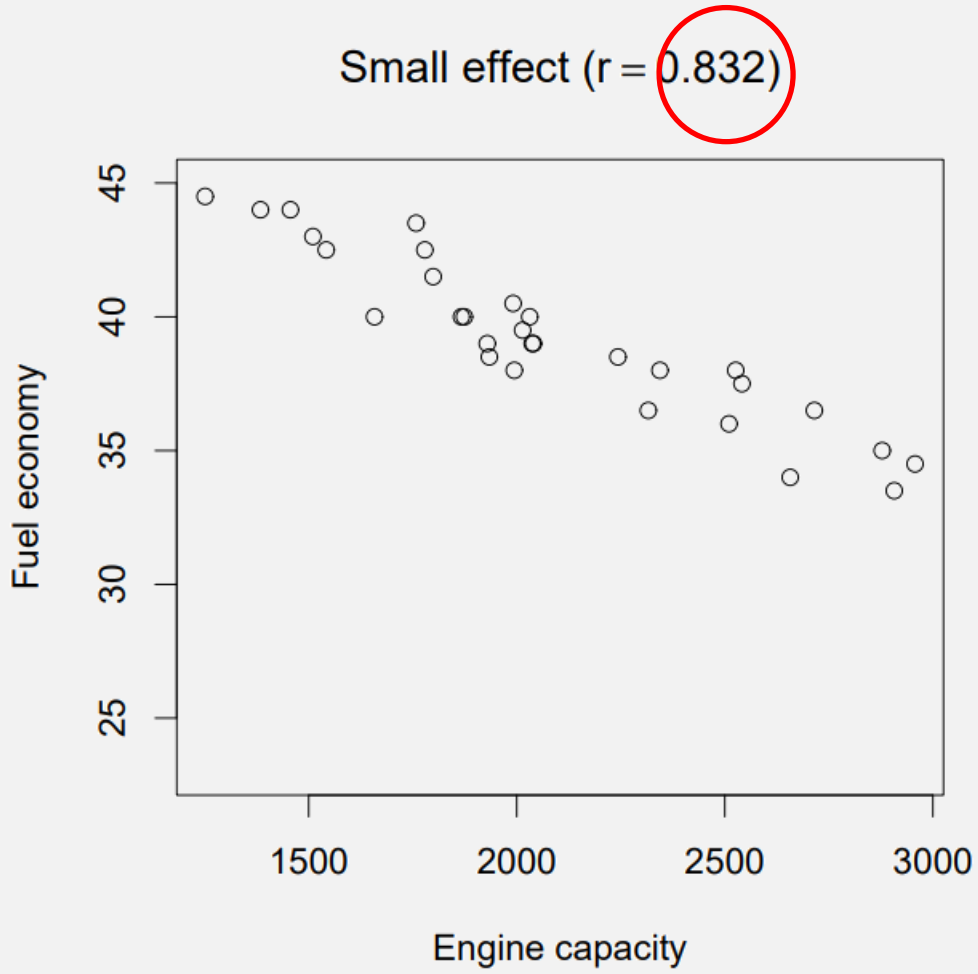
A correlation coefficient can be badly affected by outlying observations

- This chart has a data outlier
- With this outlier r=+0.83
- If we remove the outlier, r=-1.00

The Pearson coefficient is not a **robust estimator** because it is not resistant – small changes can affect its value

A robust statistic is resistant to errors in the results, produced by deviations from assumptions





Even if a relationship is genuine, a strong correlation doesn't necessarily imply that a change in one variable will produce a large change in the other one.

Source

Rank coefficients

One way to address outliers is to compute the correlation on the ranks of the data after ordering (ranking) them, rather than on the data themselves.

e.g. Spearman's rank correlation

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

ρ = 0.67

English (mark)	Maths (mark)	Rank (English)	Rank (maths)	d	d^2
80	77	1	1	0	0
76	67	2	3	1	1
75	70	3	2	1	1
71	60	4	7	3	9
64	56	5	9	4	16
62	65	6	5	1	1
61	63	7	6	1	1
58	59	8	8	0	0
56	66	9	4	5	25
45	40	10	10	0	0

NB: Correlation is not causation

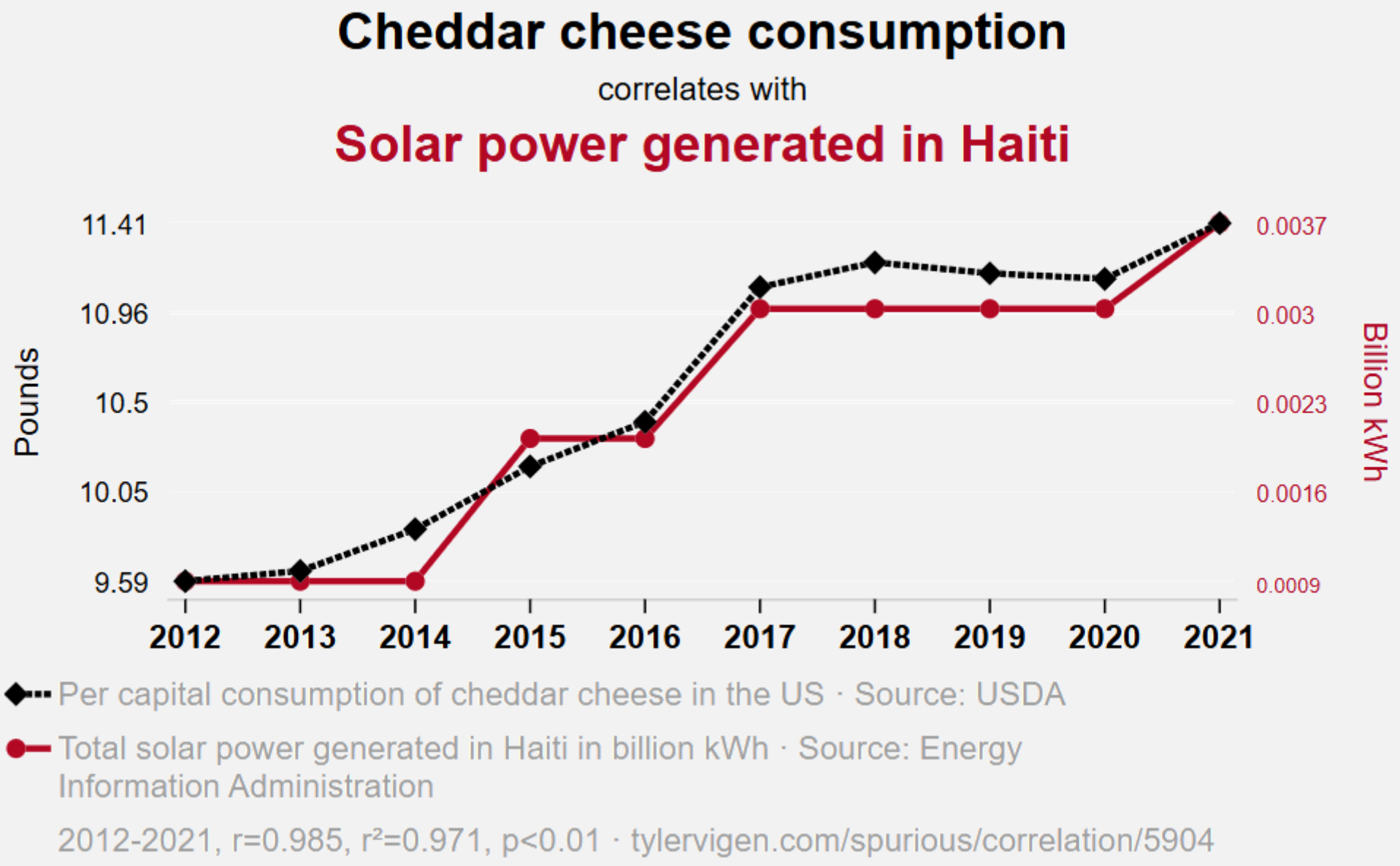
A strong (or even statistically significant) **correlation between variables does not *necessarily* mean one causes the other!**

Possible explanations for correlation:

1. X causes Y
2. Y causes X
3. Both X and Y are caused by Z (confounding)
4. Pure coincidence (especially with small samples or many comparisons)
5. Complex network of causal relationships.

Recommended reading: *The Book of Why* by Judea Pearl & Dana Mackenzie

NB: Correlation is not causation



NB: Correlation is not causation

To establish causation, we typically need:

- Controlled experiments
- Causal modelling techniques
- Understanding of underlying mechanisms

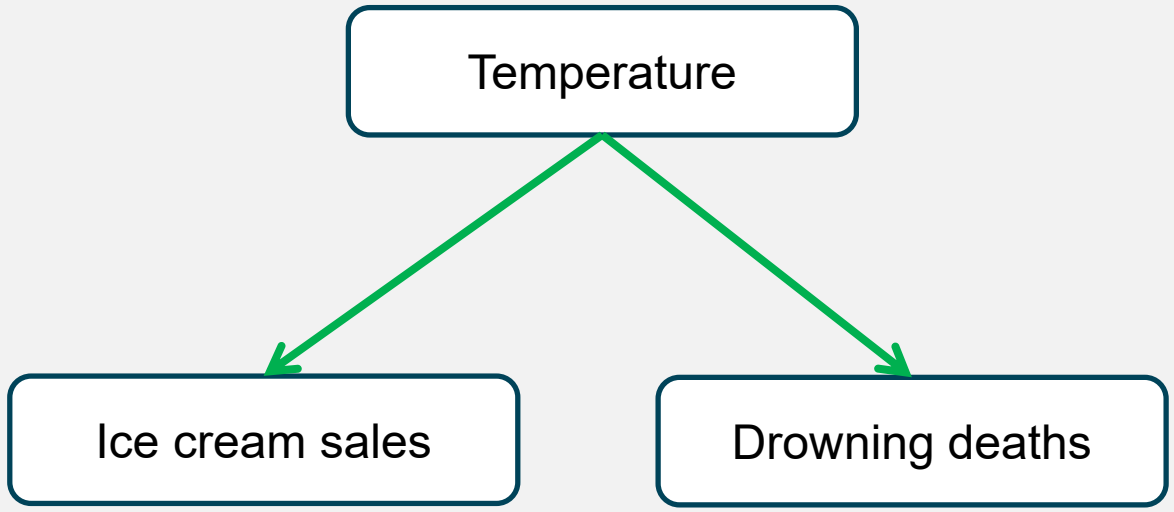
When we say that one thing *causes* another, we mean that manipulating the value of *x* should also change the value of *y.* This is the foundation of experimental science, where we control one variable and observe changes in another.

Thinking about causal relationships

Example: Ice cream sales and drowning deaths are correlated positively. Does ice cream cause drowning?

No – both are affected by a third variable: temperature/season. During summer months, both ice cream consumption and swimming activities increase, leading to more drowning.

Causal graphs

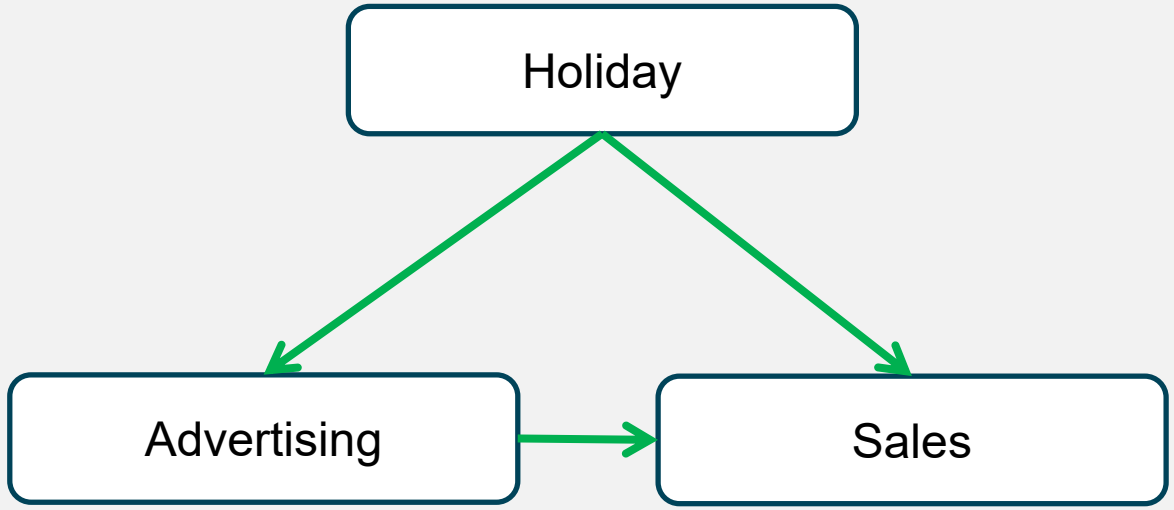


Thinking about causal relationships

Example: A positive correlation is observed between **advertising spend** and **sales**.

- Do we know that increased ad spending causes increased sales?
- What other relationships could produce this correlation?

Possibility: Both increase during holiday seasons

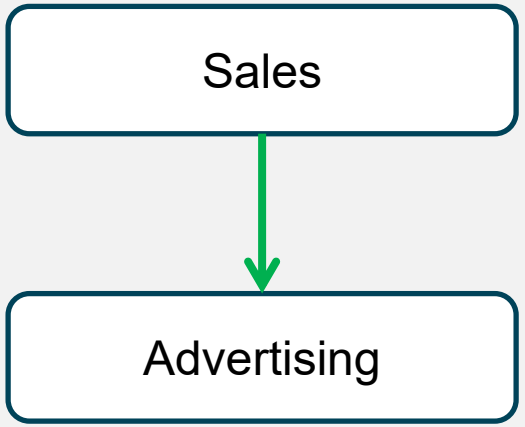


Thinking about causal relationships

Example: A positive correlation is observed between **advertising spend** and **sales**.

- Do we know that increased ad spending causes increased sales?
- What other relationships could produce this correlation?

Possibility: A company's sales increase, therefore they have more available to spend on advertising

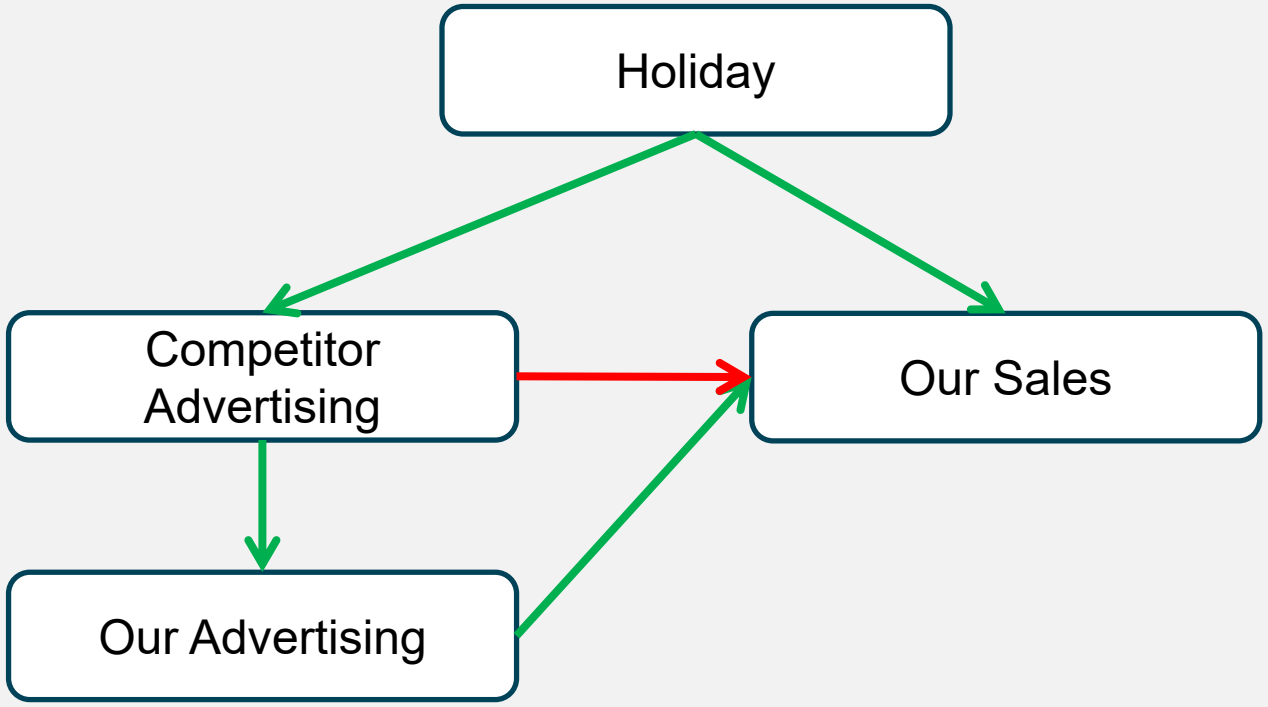


Thinking about causal relationships

Example: A causal relationship has been shown between ad spend and sales for a company.

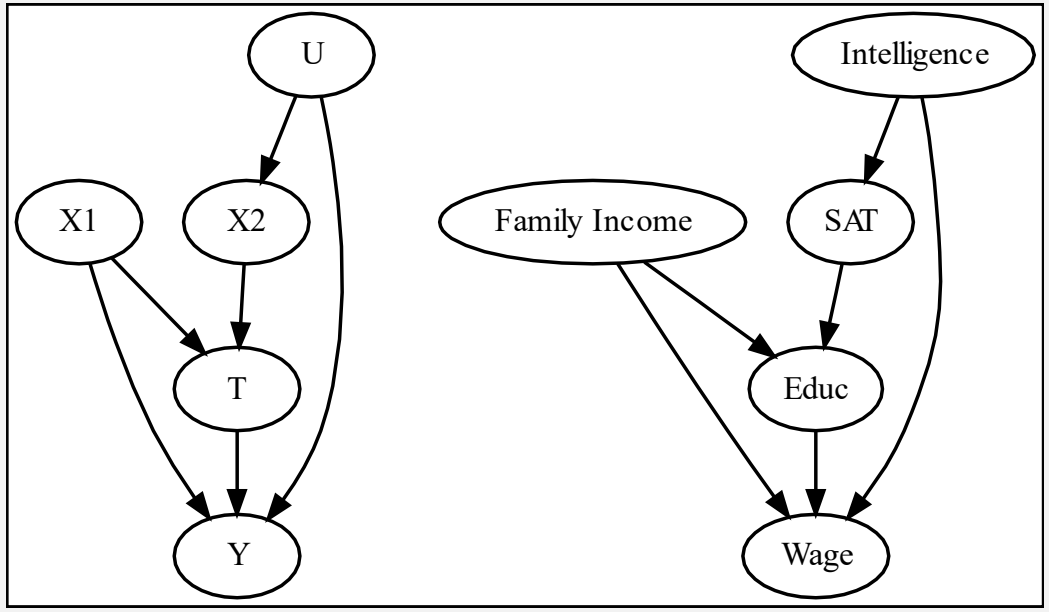
One year, this correlation has disappeared – why might this be?

Competitor's ads increase during the holidays, leading our company to increase ads, just to avoid losing sales.

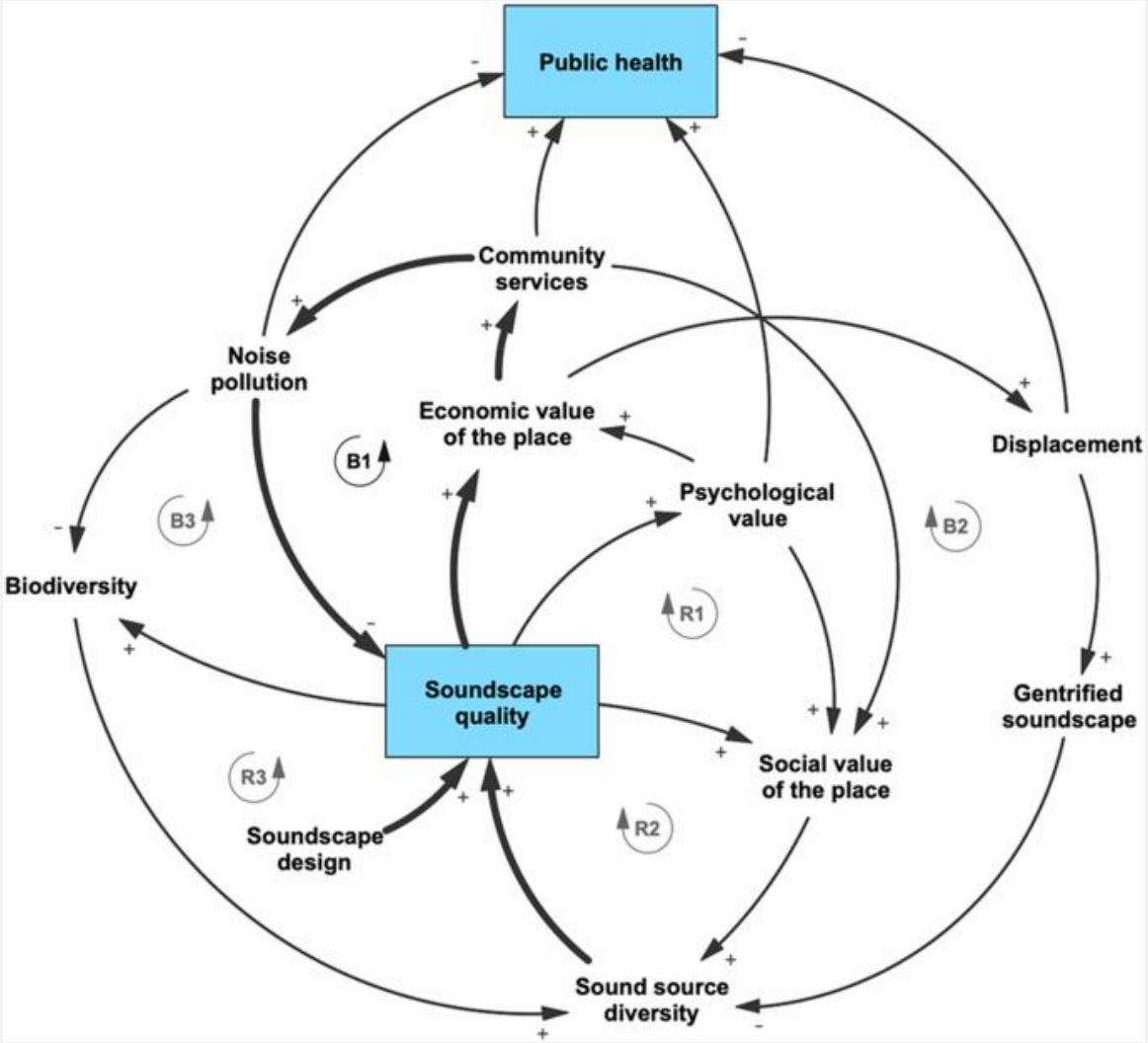


Thinking about causal relationships

Causal graphs can become very complex – even if you are not explicitly creating one, you should keep this in mind when performing statistical analysis



https://matheusfacure.github.io/python-causality-handbook/04-Graphical-Causal-Models.html



Aletta, F., Zhou, K., Mitchell, A. *et al.* Exploring the relationships between soundscape quality and public health using a systems thinking approach. *npj Acoust.* **1**, 3 (2025). https://doi.org/10.1038/s44384-025-00003-y

Break

Mid-module Feedback

Menti: 5723 0408

https://www.menti.com/aliw8jqrtbv4

Linear regression

Data = model + error

- We can use the general linear model to describe the relation between two variables and to decide whether that relationship is statistically significant

- The specific version of the GLM that we use for this is referred to as linear regression.

- A model with one explanatory variable is a **simple linear regression**; a model with two or more explanatory variables is **a multiple linear regression**.

The term *regression* was coined by Francis Galton (C19),

He observed that the heights children of extremely tall or short parents generally fell closer to the average than did their parents.

His description of 'regression to the mean' was originally only applied in a biological sense but was later extended to a more general statistical context.

- The simplest manifestation of the linear regression model is:

$$y = x * \beta_x + \beta_0 + \varepsilon$$

- β_x tells us how much we would expect y (the independent variable) to change given a one-unit change in x
- β_0 is the intercept on the y axis – the value of y when x is zero
- ε is the error (**residuals**) left over once the model has been fit to the data

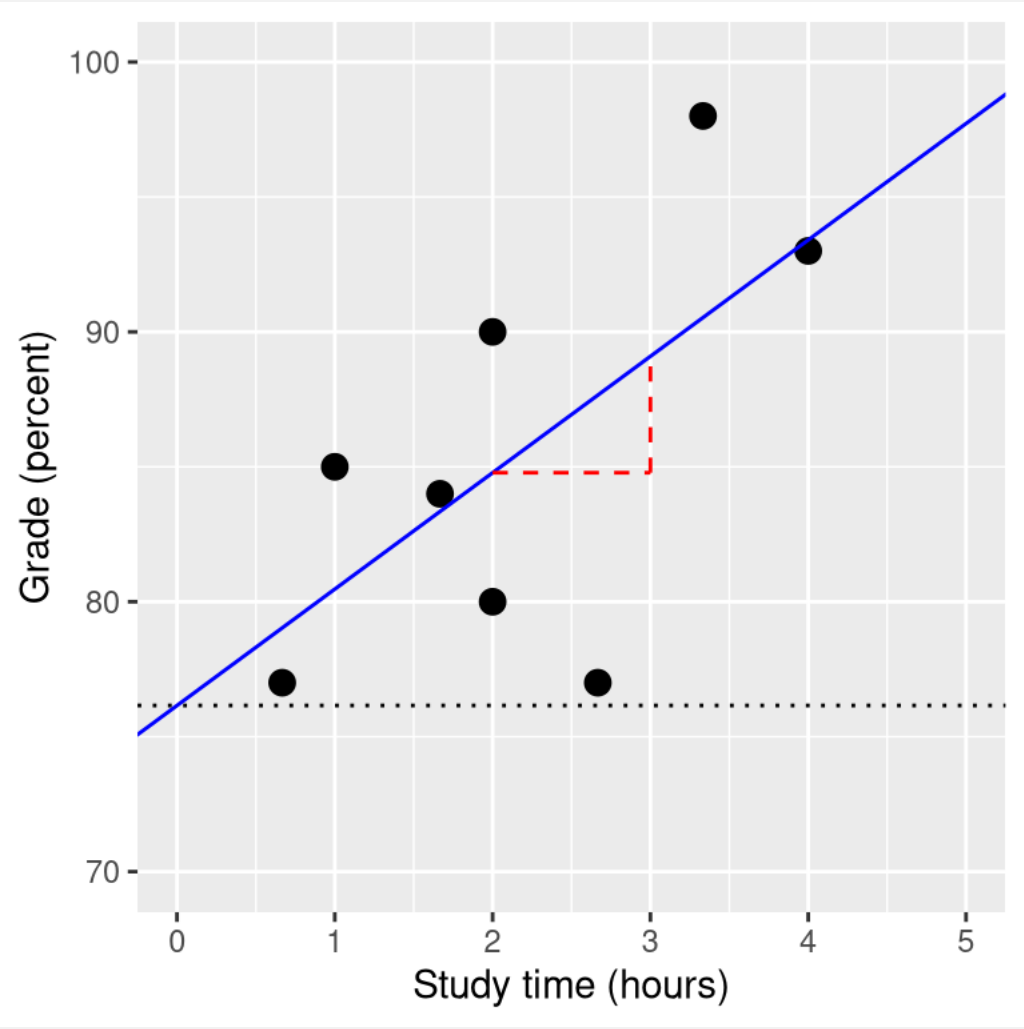
- If we are trying to predict y after β has been estimated, we can drop the error

$$\hat{y} = x * \hat{\beta}_x + \hat{\beta}_0$$

The least squares regression line

- Simple linear regression aims to find a linear relationship to describe the correlation between an independent and possibly dependent variable.
- The blue line represents the “line of best fit” through the data points.
- The regression line can be used to predict or estimate missing values, this is known as interpolation.

- We can determine the regression line through the method of **ordinary least squares (OLS)**.
- In this approach the idea is to minimise the sum of the vertical distance between all the data points and the line of best fit



Ordinary Least Squares

The equation of the LSR line should be familiar

$$y = x * \beta_x + \beta_0$$

- Estimates line by minimizing sum of squared residuals
- Residual = Actual Y – Predicted Y
- Sum of Squared Residuals (SSR):

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Ordinary Least Squares

The equation of the LSR line should be familiar

$$y = x * \beta_x + \beta_0$$

OLS estimators are calculated as:

$$\beta_x = \frac{\mathrm{cov}(x,\ y)}{var(x)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\beta_0 = \bar{y} - \bar{x} * \beta_x$$

Ordinary Least Squares

The OLS formula for the slope can be rewritten as:

$$\beta_x = \frac{\text{cov}(x,\ y)}{var(x)} = r_{xy}\frac{s_y}{s_x}$$

where

- cov(*x, y)* is the covariance between *x* and *y*
- *var(x)* is the variance of x
- r_{xy} is the **correlation between *x* and *y***
- s_x and s_y are the standard deviations

This shows the **direct relationship** between correlation and regression:

the slope coefficient is the correlation multiplied by the ratio of standard deviations.

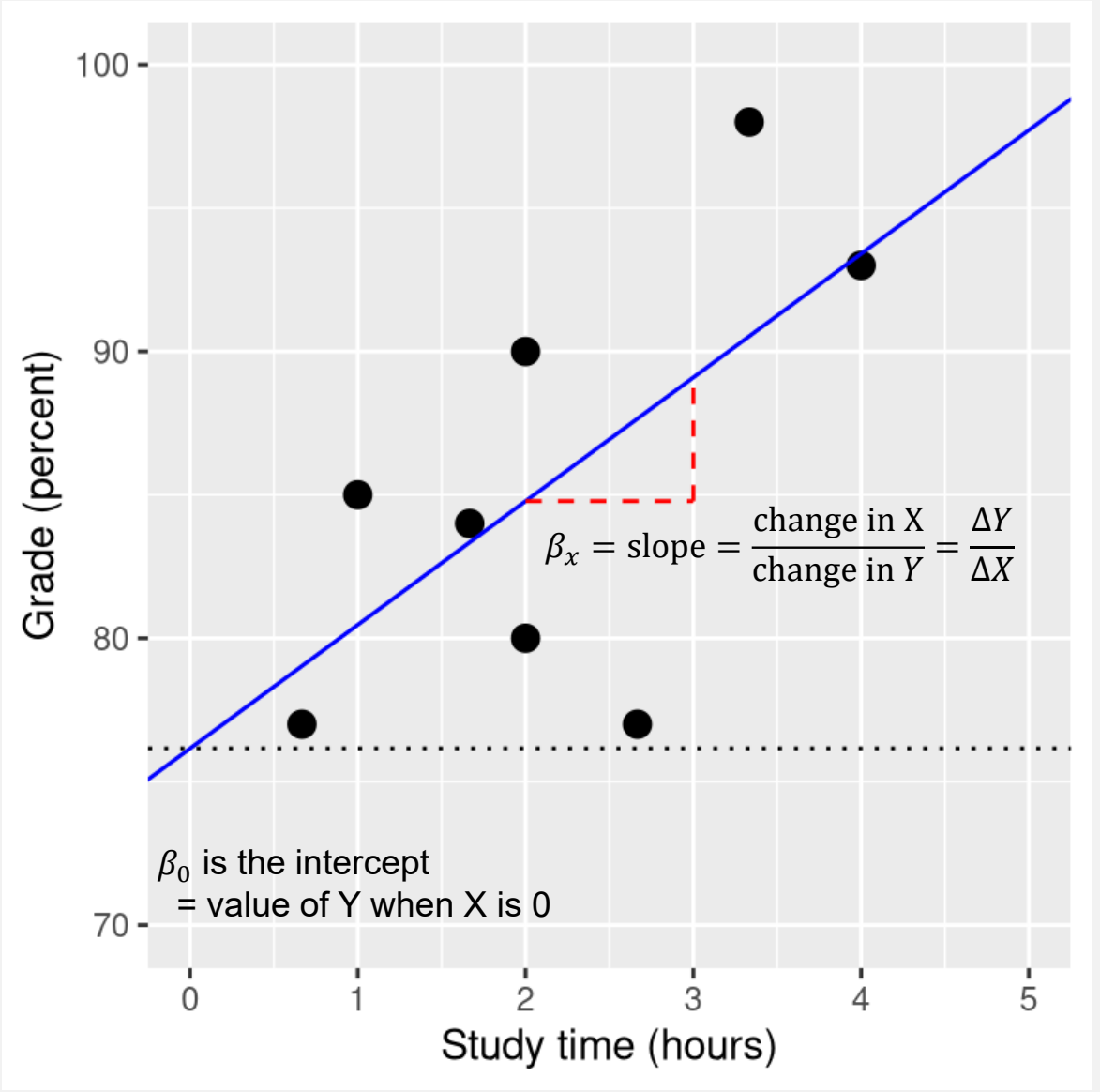
This means the magnitude of β_x depends not only on the correlation strength, but also on the spread of both variables.

Regression coefficients cannot be directly compared across different predictors – they depend on the scale of the variables.

$$y = x * \beta_x + \beta_0$$

Where:

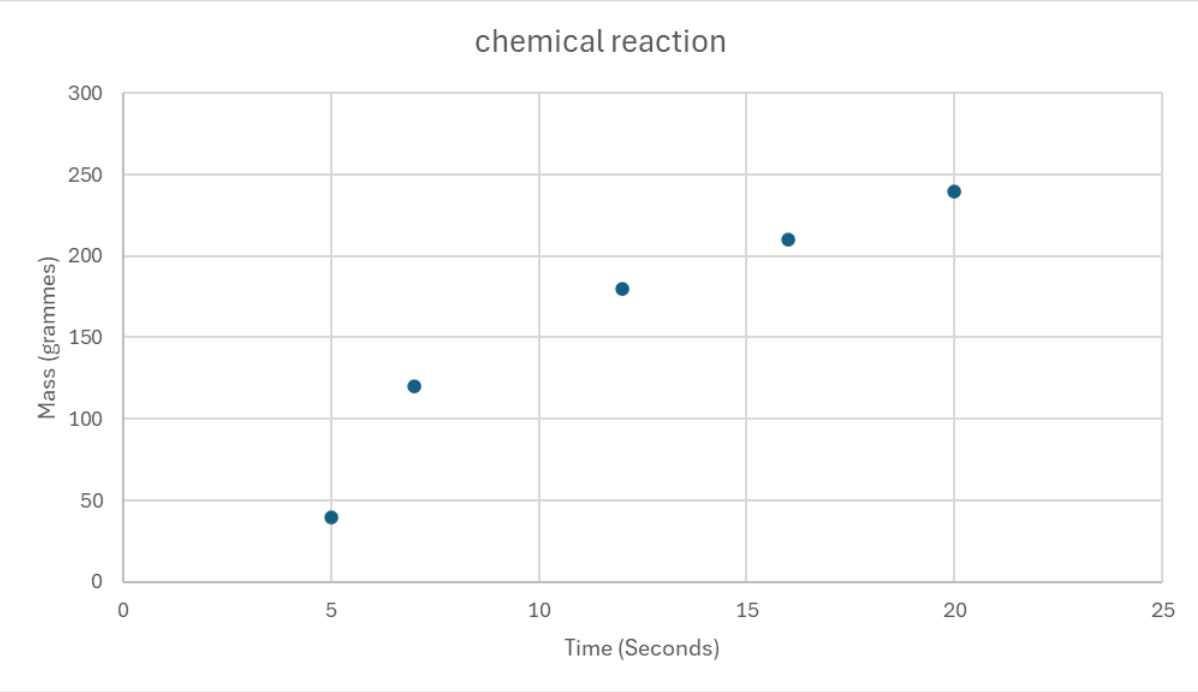
- $\beta_0 = \bar{y} - \bar{x} * \beta_x$
- $\beta_x = \frac{cov(x,y)}{var(x)}$



Example

Consider the example below where the mass, y (grams), of a chemical is related to the time, x (seconds), for which the chemical reaction has been taking place according to the table

Time (x) seconds	5	7	12	16	20
Mass (y) grammes	40	120	180	210	240



Find the equation of the regression line

Linear regression

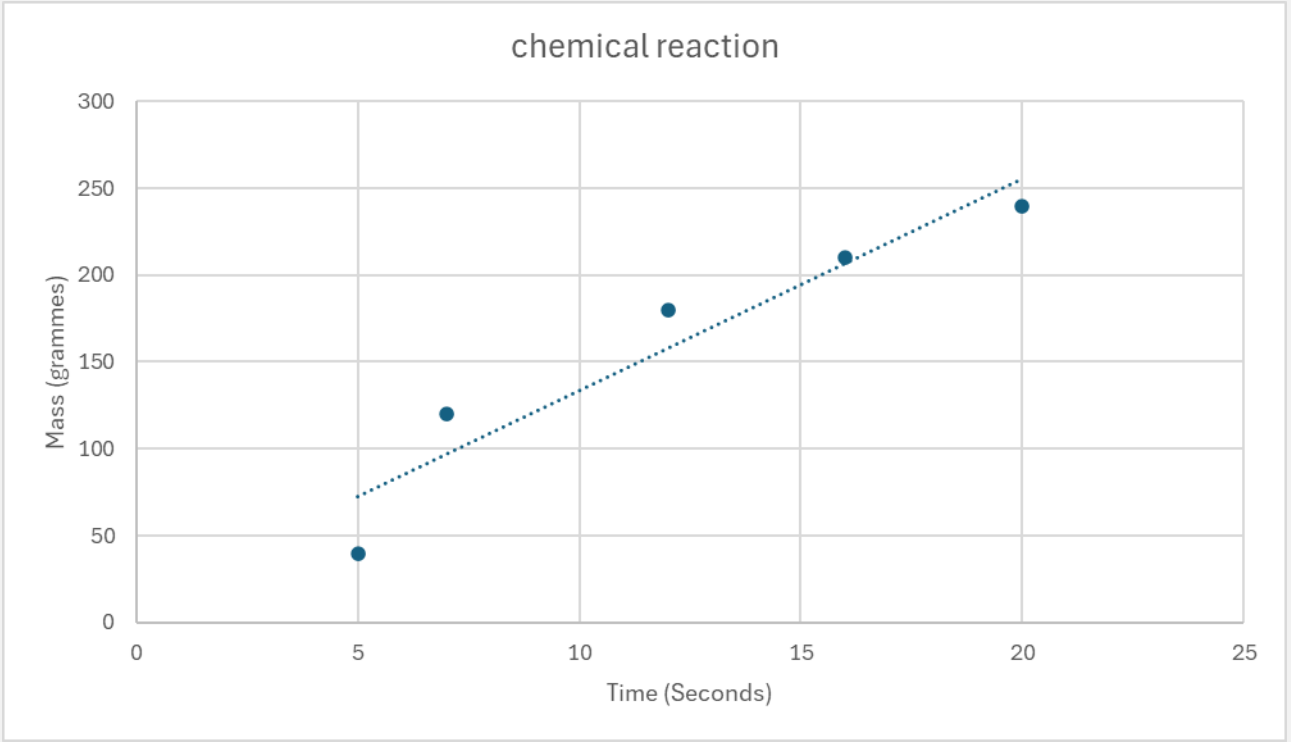
$$y = x * \beta_x + \beta_0$$

Where

$$\beta_0 = \bar{y} - \bar{x} * \beta_x$$

$$\beta_x = S_{xy} / S_{xx}$$

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



Linear regression

	x	y	x_dev	y_dev	S_{xy}	S_{xx}
	5	40	-7	-118	826	49
	7	120	-5	-38	190	25
	12	180	0	22	0	0
	16	210	4	52	208	16
	20	240	8	82	656	64
Mean	12	158				
Sum	60	790			1880	154

$\beta_x = S_{xy} / S_{xx} = 1880 / 154 = 12.208$ (3dp)

$B_0 = \bar{y} - \bar{x} * \beta_x = 158 - (12*12.208) = 11.506$

$\therefore \hat{y} = 11.506 + 12.208x$

Linear Regression in R

We use the lm() function in R.

The linear model formula in R takes the form:

$$y \sim x$$

This can be extended to include multiple predictors:

$$y \sim x1 + x2 + x3$$

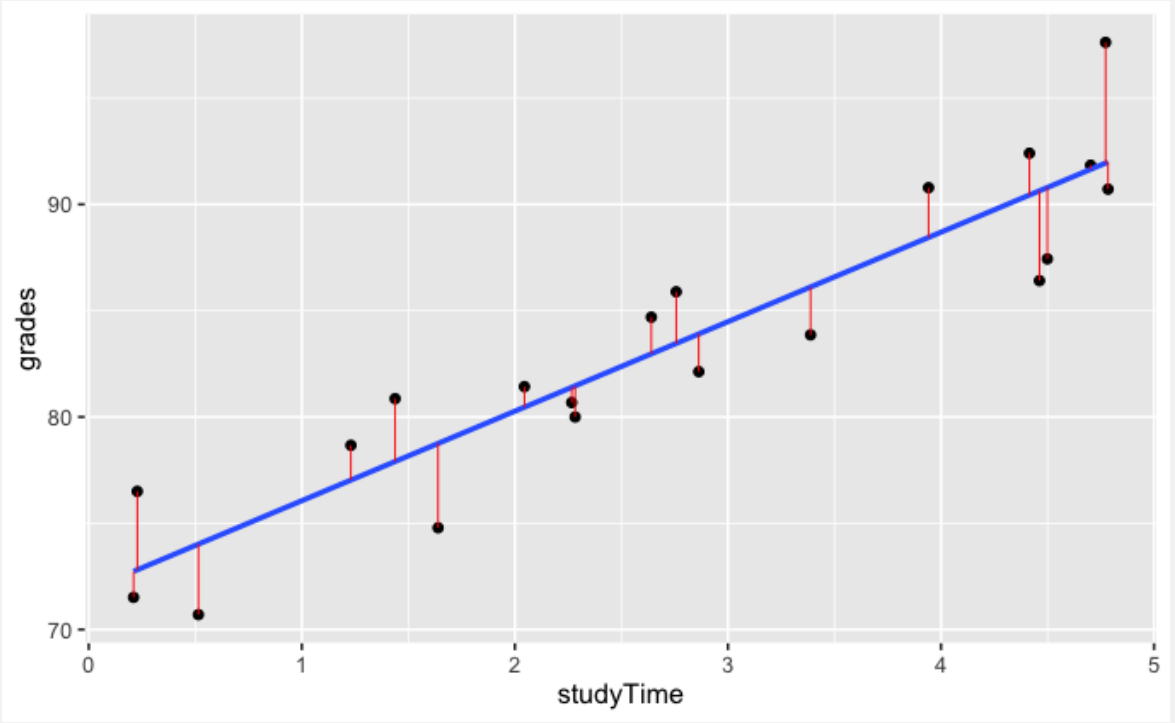
```r
{r echo=TRUE, eval=FALSE}
# Fitting a linear regression model in R
model <- lm(grades ~ studyTime, data = df)
summary(model)

# The summary() function provides:
# - Estimated coefficients (intercept and slope)
# - Standard errors for each coefficient
# - t-values and p-values for significance tests
# - R-squared and adjusted R-squared values
# - F-statistic and its p-value
# - Residual standard error
```

```
Call:
lm(formula = grades ~ studyTime, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2284 -1.8960 -0.2605  2.0599  5.6778

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.8488     1.3149   54.64  < 2e-16 ***
studyTime     4.2111     0.4175   10.09 7.83e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.852 on 18 degrees of freedom
Multiple R-squared:  0.8497,    Adjusted R-squared:  0.8413
F-statistic: 101.7 on 1 and 18 DF,  p-value: 7.826e-09
```

Checking our assumptions

There are broadly 4 assumptions of a linear regression:

1. There is a linear relationship between the predictors (x) and the outcome (y)

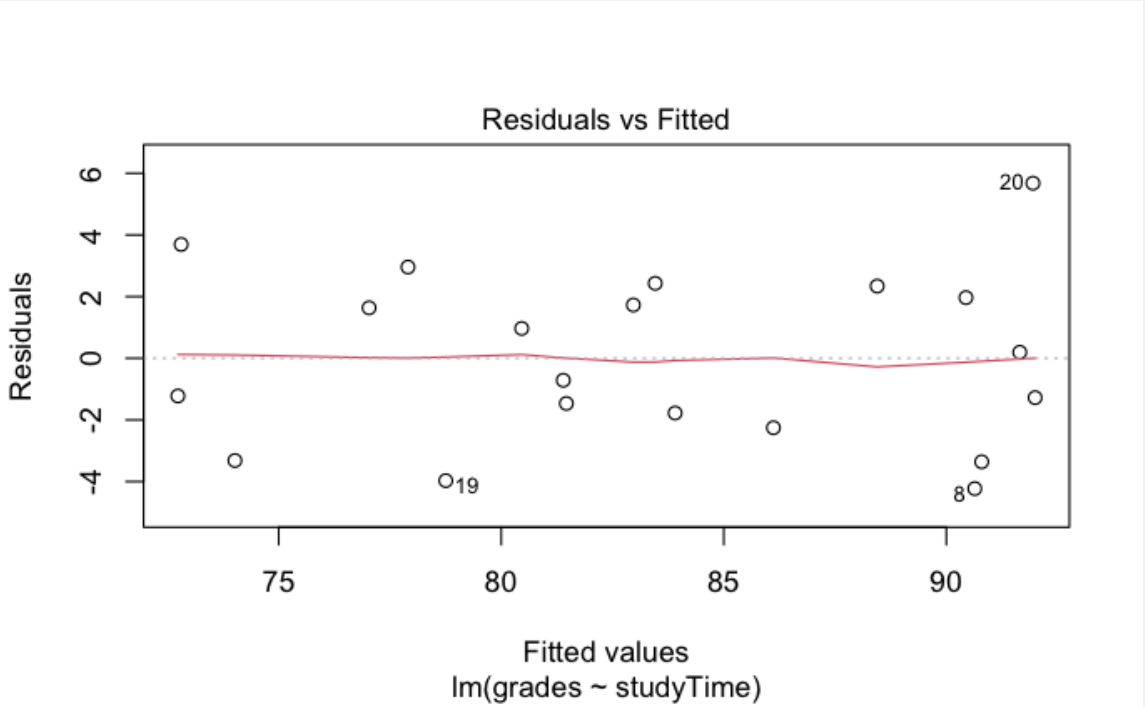


Checking our assumptions

There are broadly 4 assumptions of a linear regression:

1. There is a linear relationship between the predictors (x) and the outcome (y)
2. Residual Errors have a mean value of zero - Normalcy

```r
{r}
plot(model)
```

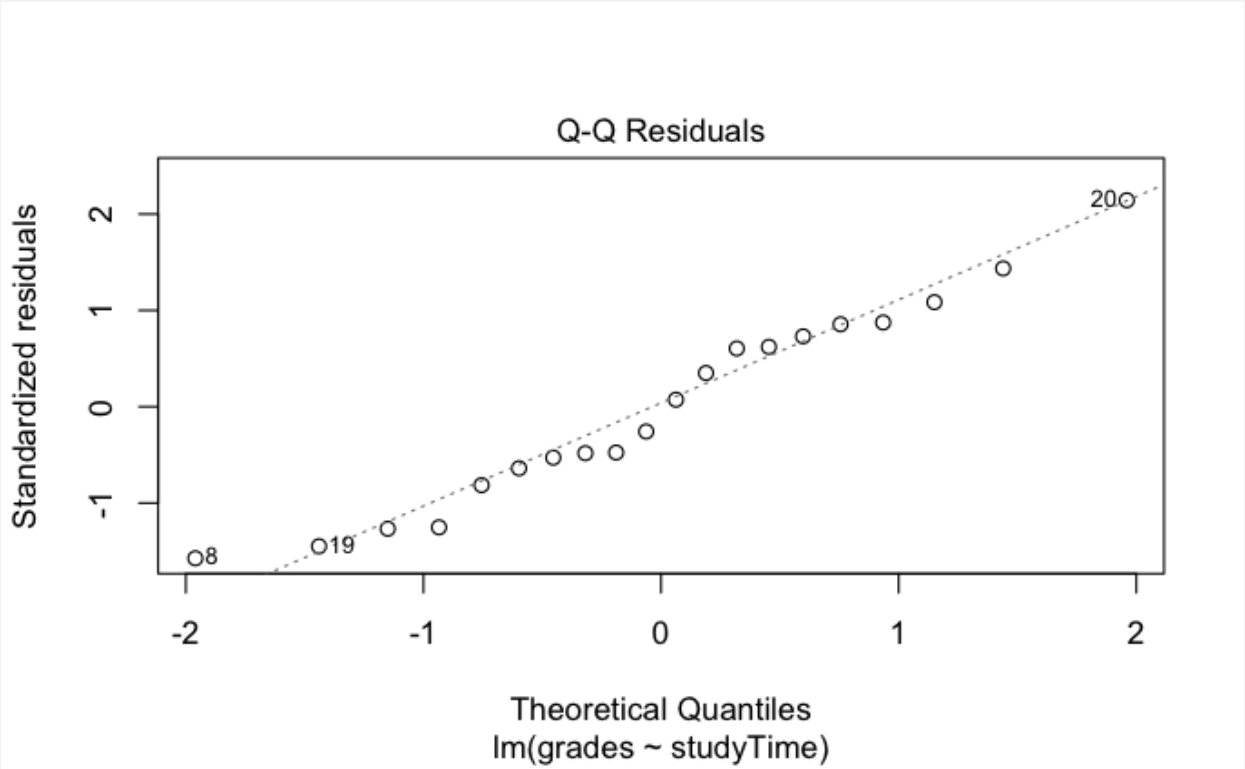


Checking our assumptions

There are broadly 4 assumptions of a linear regression:

1. There is a linear relationship between the predictors (x) and the outcome (y)
2. Residual Errors have a mean value of zero - Normalcy
3. Residual Errors have constant variance - Homoscedasticity

```r
{r}
plot(model)
```

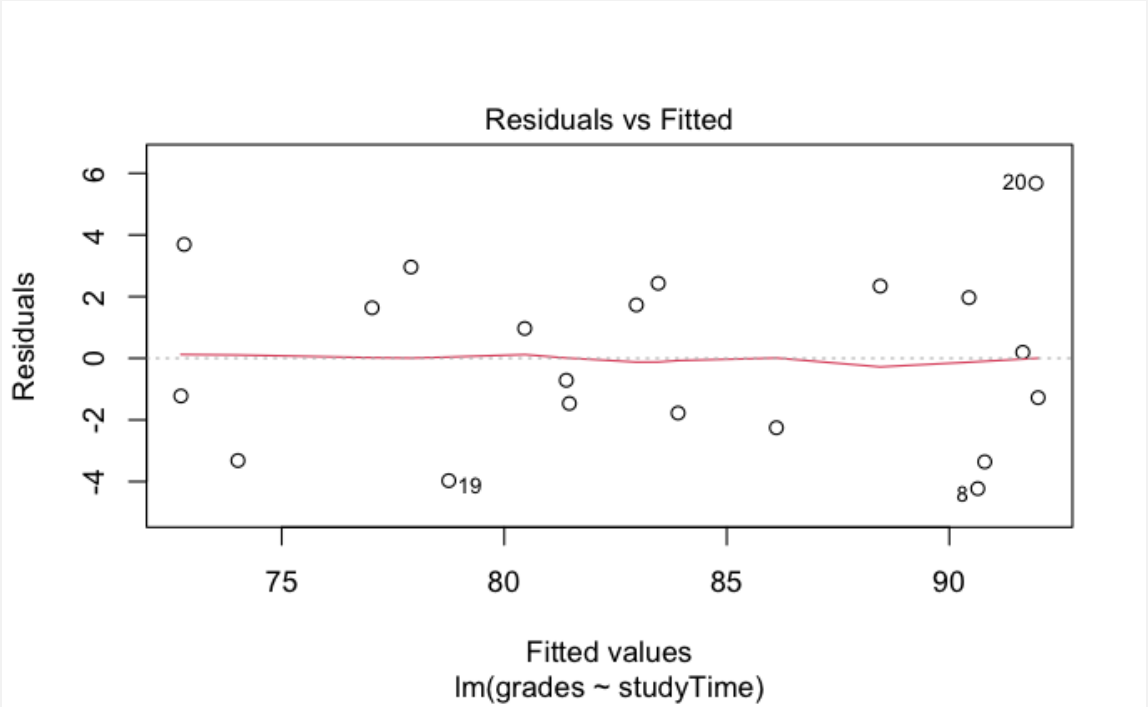


Checking our assumptions

There are broadly 4 assumptions of a linear regression:

1. There is a linear relationship between the predictors (x) and the outcome (y)
2. Residual Errors have a mean value of zero - Normalcy
3. Residual Errors have constant variance - Homoscedasticity
4. Residual Errors are independent from each other and predictors (x) - Independence

```r
{r}
plot(model)
```



Limitations of regression

As with correlation calculations, regression has limitations:

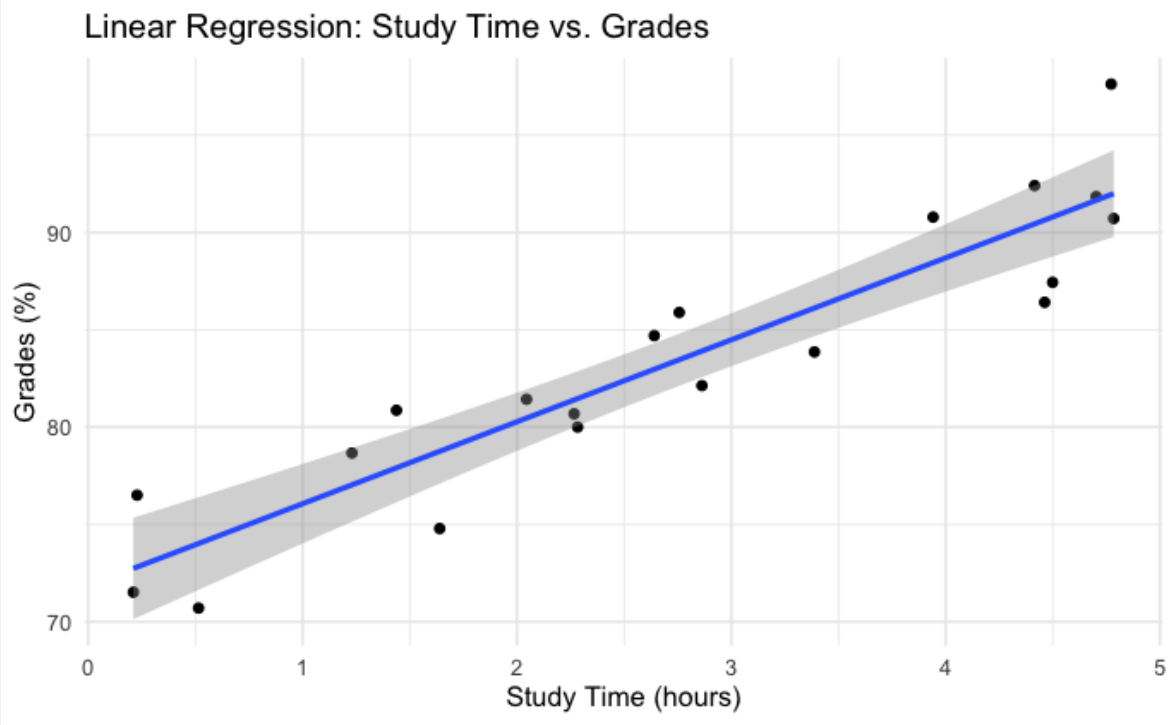
- Assumptions on the relationships between the data may not hold
- Other (known) variables may be having an influence on the dependent variables (multiple regression, next week)
- Other unknown variables may be at play
- Regression struggles with non-linear trends, dynamic and complex environments
- It's sensitive to outliers
- Historical data on which the regression is based may no longer be valid

There are other competing models for determining the regression line based on (for example) fit, parsimony or predictive power. Selecting and specifying the appropriate model requires judgment and expertise

Prediction intervals

- After you have created your regression model, the model can be used to generate predictions of the dependent variable, based on the independent variable.
- However, because the model regresses to a mean, the predicted value from the model is the mean response value
- Like any mean, there is variability around that mean.

- **Prediction intervals** account for the variability around the mean response.
- Prediction intervals will have a confidence level and can be two- or one-sided.



Goodness of fit

- It's useful to quantify how well the model fits the data overall
- One way to do this is to ask how much of the variability in the data is accounted for by the model. **We use a value called R^2**
- The R^2 value is also known as the ***coefficient of determination,*** and tells us the percentage of the variation in the underlying data that we have accounted for in the model

Where there's only one x variable: $R^2 = r^2$

i.e. to calculate R^2 you simply square the correlation coefficient

- The variability in the data can be broken down into two sums of squares
  - The sum of squared errors (or residuals) SS_{error}
  - The Sum of squares of the model SS_{model}
- The sum of squares of the model can be found by looking at the total variability of the data (the total sum of squares) and deducting SS_{error}

$$i.e\ SS_{model} = SS_{total} - SS_{error}$$

$$Then\ R^2 = \frac{SS_{model}}{SS_{total}} = 1 - \frac{SS_{error}}{SS_{total}}$$

A small value of R^2 tells us that even if the model fit is statistically significant, it may only explain a small amount of information in the data.

Standard errors for regression models

- When we estimate the regression coefficients β_0 and β_x these are just point estimates from our sample.
- Just like any statistic, there is uncertainty in these estimates.
- The standard error of the regression coefficient tells us how much the coefficient value might vary across different samples.
- The standard error for the slope coefficient is calculated as:

$$SE_{\beta_x} = \frac{SE_{model}}{\sqrt{\sum(x_i - \bar{x})^2}}$$

Confidence intervals for regression coefficients

We can construct confidence intervals for regression coefficients using their standard errors:

$$\widehat{\beta_x} \pm t\alpha_{/2} \times SE_{\beta_x}$$

Interpretation:

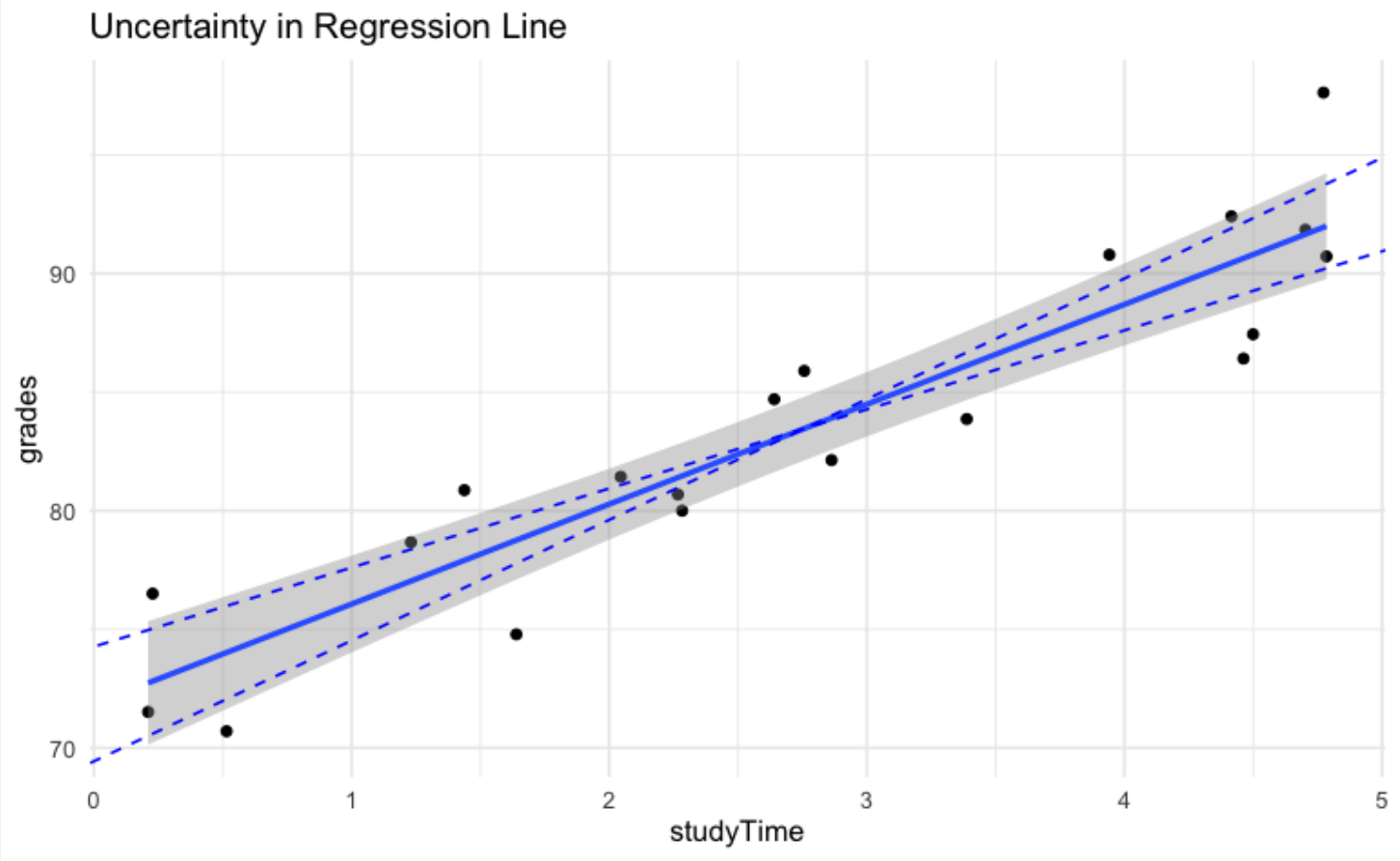
”We are 95% confident that the true effect of X on Y lies between [lower bound] and [upper bound].”

Confidence intervals for coefficients:

```r
{r echo=TRUE, eval=FALSE}
confint(model, level = 0.95)
##                   2.5 %    97.5 %
## (Intercept) 68.9376954 71.489689
## studyTime    4.5190811  5.465801
```

```
                2.5 %    97.5 %
(Intercept) 69.086337 74.611354
studyTime    3.333906  5.088231
```

Confidence intervals for regression coefficients



Confidence intervals for coefficients:

```r
{r echo=TRUE, eval=FALSE}
confint(model, level = 0.95)
##                   2.5 %    97.5 %
## (Intercept) 68.9376954 71.489689
## studyTime    4.5190811  5.465801
```

```
                2.5 %    97.5 %
(Intercept) 69.086337 74.611354
studyTime    3.333906  5.088231
```

Interpreting Regression Output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  70.2137     0.6076  115.56  < 2e-16 ***
studyTime     4.9924     0.2253   22.16  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.196 on 18 degrees of freedom
Multiple R-squared:  0.9648,    Adjusted R-squared:  0.9628
F-statistic: 491.1 on 1 and 18 DF,  p-value: < 2.2e-16
```

What this tells us:

- **Intercept (β_0) = 70.21**: The expected grade when study time is zero
- **Slope (β_1) = 4.99**: For each additional hour of study time, grades increase by 4.99 points on average
- **p-values < 2e-16**: Both coefficients are highly statistically significant
- **R^2 = 0.9648**: About 96.5% of the variance in grades is explained by study time
- **Residual standard error = 1.196**: The typical prediction error is about 1.2 grade points

Statistical significance: The p-values test the null hypothesis that the true coefficient is zero (H_0: $\beta_i = 0$).

Interpreting the slope: "Each additional hour of study time is associated with a 4.99 percentage point increase in grades, on average."

Interpreting the intercept: "A student who doesn't study at all (study time = 0) is expected to achieve a grade of 70.21%, on average."

Summary of the session

1. **Correlation**:
    1. Measures linear association between variables
    2. Range: -1 to +1
    3. Symmetrical ($r_{XY} = r_{YX}$)
    4. Does not imply causation

2. **Regression**:
    1. Models relationship between outcome and predictors
    2. Provides coefficients with interpretable meaning
    3. Allows prediction of Y from X
    4. Asymmetrical (Y ~ X ≠ X ~ Y)

Reminder: Statistical vs. Practical Significance

- Statistical significance ($p < 0.05$) means an effect is unlikely to be due to chance
- Practical significance refers to whether an effect is large enough to matter in real-world contexts
- With large samples, even tiny effects can be statistically significant
- We must consider both when interpreting results

Responsible Use:

1. Check assumptions before interpreting results
2. Be cautious about causal claims from observational data
3. Consider both statistical and practical significance
4. Validate prediction models on new data
5. Be transparent about limitations and uncertainties

Assessment Seminar

Please pull up your Reading Week Assignment

- Pair up with a partner
- Each partner present their visualisations to the other
- Guess which is the deceptive visualisation
- Provide feedback and critique on the visualisations

Mid-module Feedback

Menti: 5723 0408

https://www.menti.com/aliw8jqrtbv4