

Topic 9 – Describing relationships

ENVX1002 Introduction to Statistical Methods

Si Yang Han

The University of Sydney

Feb 2025



THE UNIVERSITY OF
SYDNEY

About me

- Research topics: spatial modelling and mapping, precision agriculture, winter grains
- Timeline at USYD
 - ➡ BSc (Hons) in Agricultural Science
 - ➡ PhD in Digital Agriculture
 - ➡ Postdoc in Spatial Modelling
 - ➡ Associate Lecturer in Agricultural Data Science



Faba beans at Trangie

Learning Outcomes

- LO1. Demonstrate proficiency in utilizing R and Excel to effectively explore and describe data sets in the life sciences.
- LO2. Evaluate and interpret different types of data in the natural sciences by visualising probability distributions and calculating probabilities using RStudio and Excel.
- LO3. Apply parametric and non-parametric statistical inference methods to experimental data using RStudio and effectively interpret and communicate the results in the context of the data.
- LO4. Apply both linear and non-linear models to describe relationships between variables using RStudio and Excel, demonstrating creativity in developing models that effectively represent complex data patterns.
- LO5. Articulate statistical and modelling results clearly and convincingly in both written reports and oral presentations, working effectively as an individual and collaboratively in a team, showcasing the ability to convey complex information to varied audiences.

Module overview

- **Week 9. Describing Relationships**
 - ➡ Correlation (calculation, interpretation)
 - ➡ Regression (model structure, model fitting)
 - ➡ What/when/why/how
- Week 10. Simple Linear Regression
 - ➡ Can we use the model?(assumptions, hypothesis testing)
 - ➡ How good is the model?(interpretation, model fit)
- Week 11. Multiple Linear Regression
 - ➡ Multiple Linear Regression (MLR) modelling
 - ➡ Assumptions, interpretation and the principle of parsimony
- Week 12. Nonlinear Regression
 - ➡ Common nonlinear functions
 - ➡ Transformations

Module overview

- **Week 9. Describing Relationships**

- ➡ Correlation (calculation, interpretation)
- ➡ Regression (model structure, model fitting)
- ➡ What/when/why/how

Example - Galton's Data

► Code

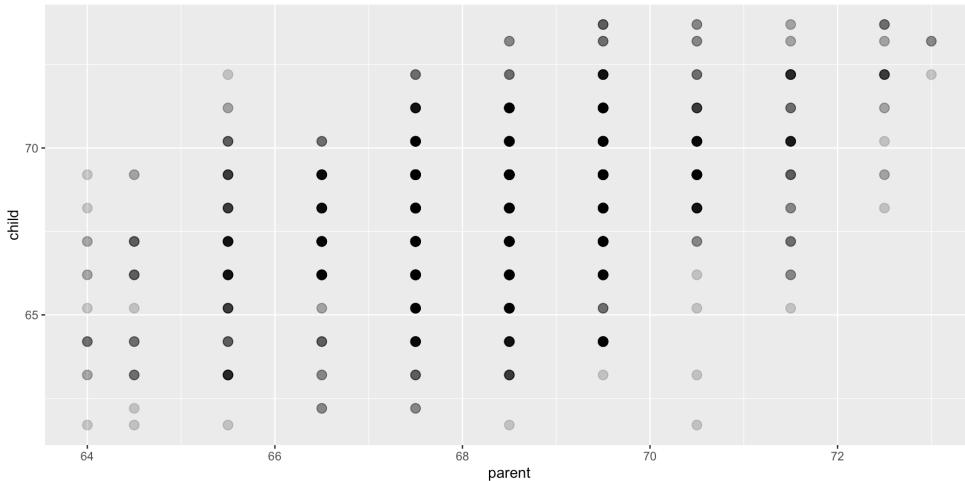
```
parent child
1    70.5  61.7
2    68.5  61.7
3    65.5  61.7
4    64.5  61.7
5    64.0  61.7
6    67.5  62.2
```

- 928 children of 205 pairs of parents
- Average height of both parents and child measured in inches
- Size classes were binned (hence data looks discrete)

Example - Galton's Data

We can visually inspect the relationship between the two variables using a scatterplot:

► Code



... but this is not a very good way to assess the strength of the relationship between the two variables.

Is the relationship:

- Linear?
- Positive or negative?
- Weak, moderate or strong?

Correlation

Correlation

The **correlation coefficient** is a number between -1 and 1 that describes the relationship between two **continuous** variables.

- **Direction:**

- Positive – both variables increase together
- Negative – one variable increases as the other decreases

- **Strength:**

- -1 → perfect negative relationship
- 0 → no relationship
- 1 → perfect positive relationship
- Common terms: *weak* (~0.1–0.3), *moderate* (~0.4–0.6) or *strong* (~0.7–1.0) – useful but subjective!

Correlation

Pearson's Correlation (r) Formula:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Covariance divided by the product of the standard deviations.

Karl Pearson developed the **correlation coefficient** in 1800s (based on the work by Francis Galton)



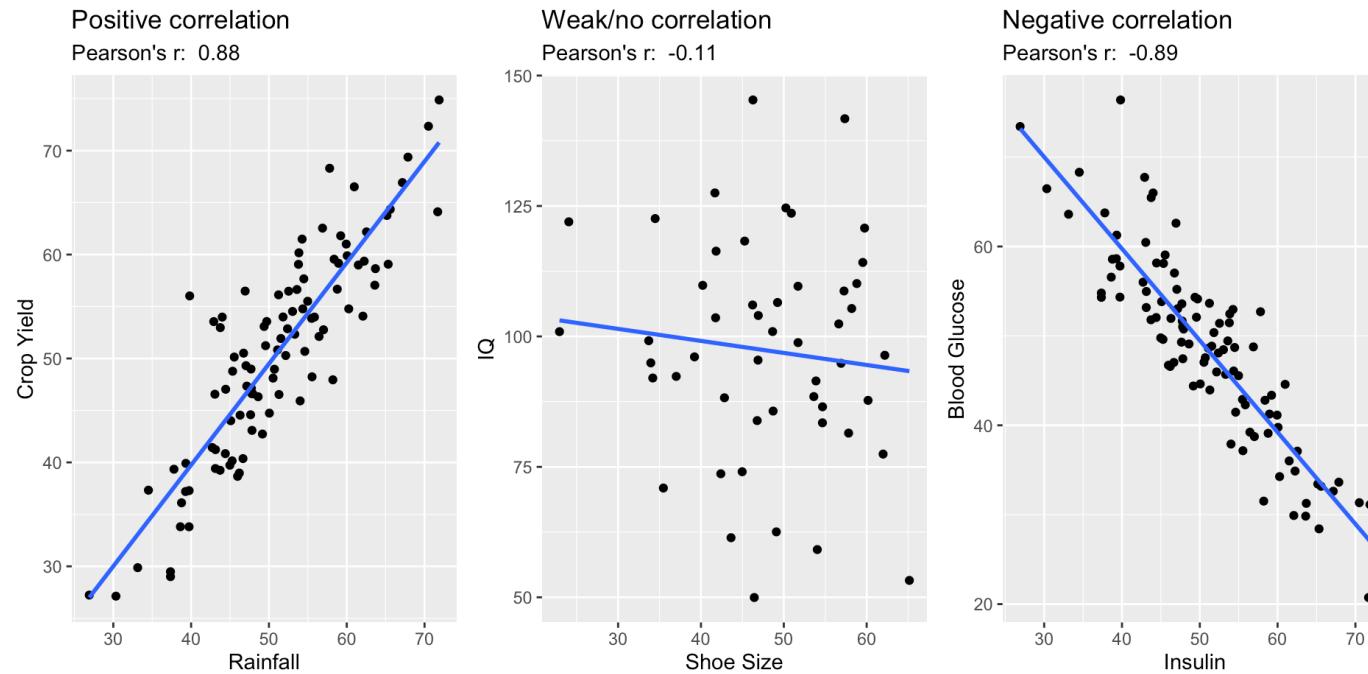
Pearson's r is:

- The most commonly used correlation coefficient
- For normally distributed data only (parametric)
- For linear relationships only (i.e. a straight line)

Luckily we have Excel and R

- Excel: `=CORREL()` formula, or use the Analysis Toolpak
- R: `cor()` function
 - ➡ By default this calculates Pearson's correlation coefficient
 - ➡ Can also calculate other types of correlation (`cor(x, y, method = "spearman")`)

► Code



(Note: not real data)

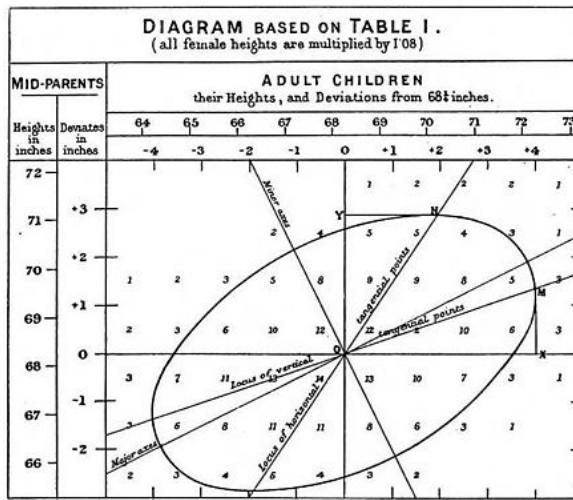
A **weak** or **nonexistent** relationship is one where the correlation coefficient is close to 0.

A **moderate** relationship falls closer to 0.5.

A **strong** relationship is one where the correlation coefficient is close to -1 or 1.

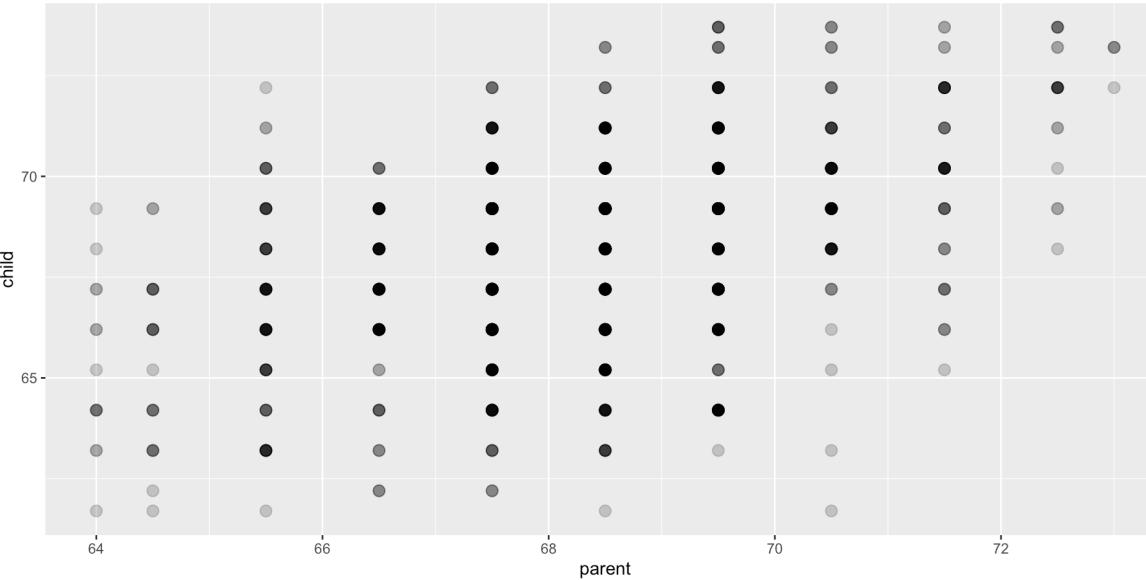
The exact values are subjective and vary between fields of study.

Example - Galton's Data



Galton's Results

► Code



► Code

```
[1] 0.4587624
```

Is the relationship:

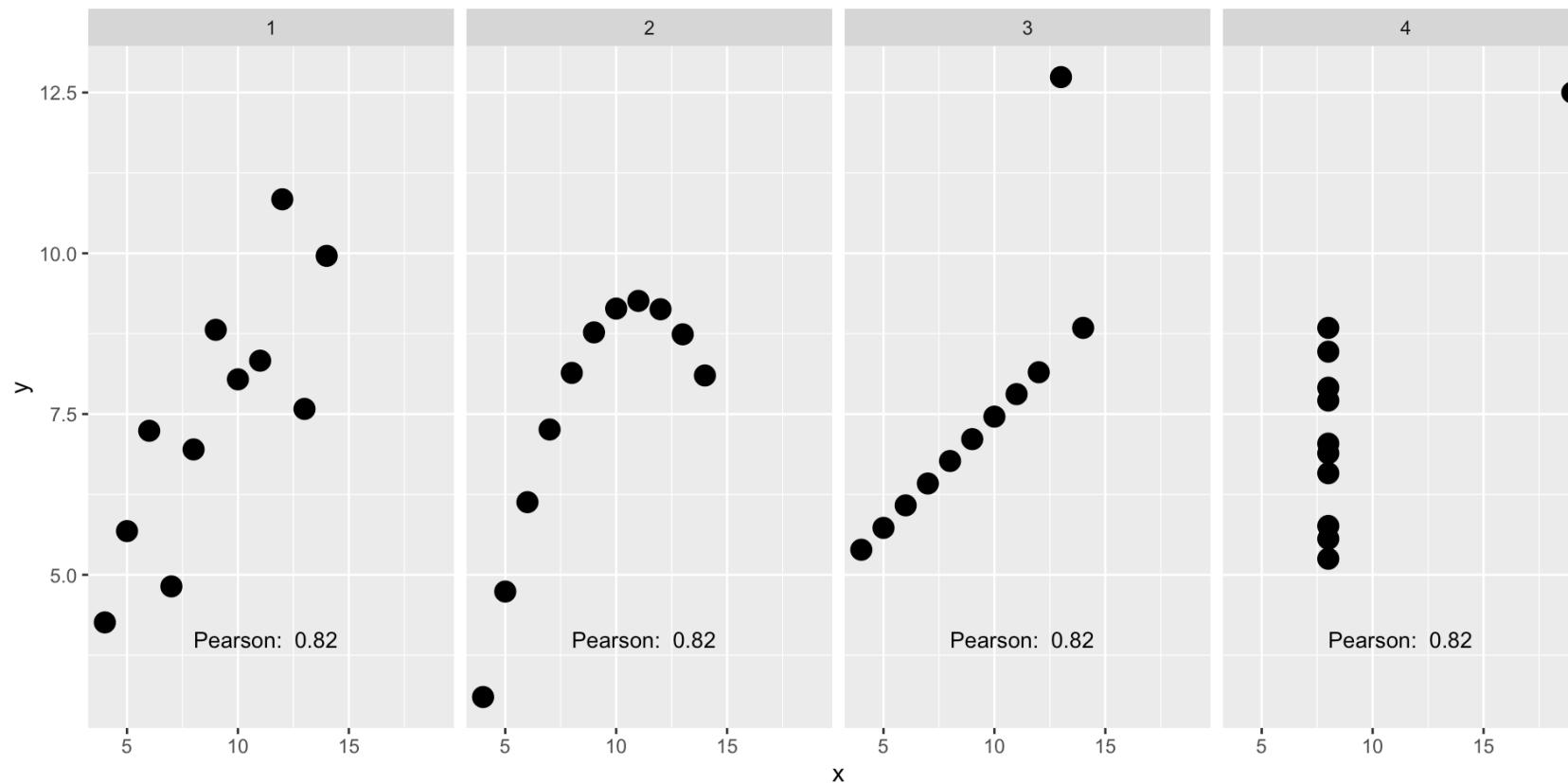
- Linear? **Seems to be**
- Positive or negative? **Positive**
- Weak, moderate or strong? **Moderate**

In words: "There is a moderate positive linear relationship between the height of parents and the height of their children."

Anscombe's Quartet

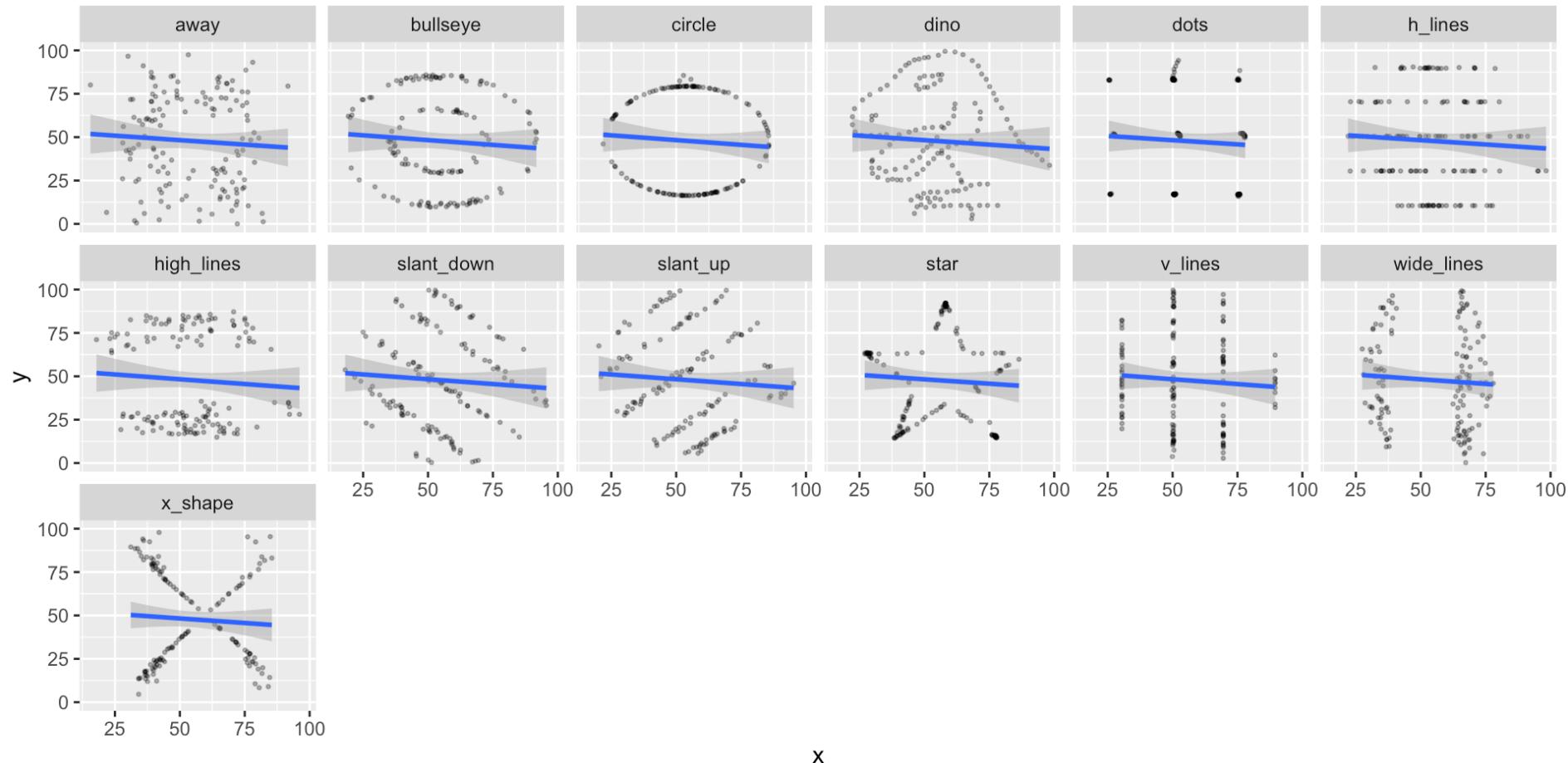
A set of four datasets that are nearly identical in simple descriptive statistics but have very different distributions.

► Code



Datasaurus Dozen

► Code



All of these data have a correlation coefficient close to zero!

Monotonic vs linear relationships

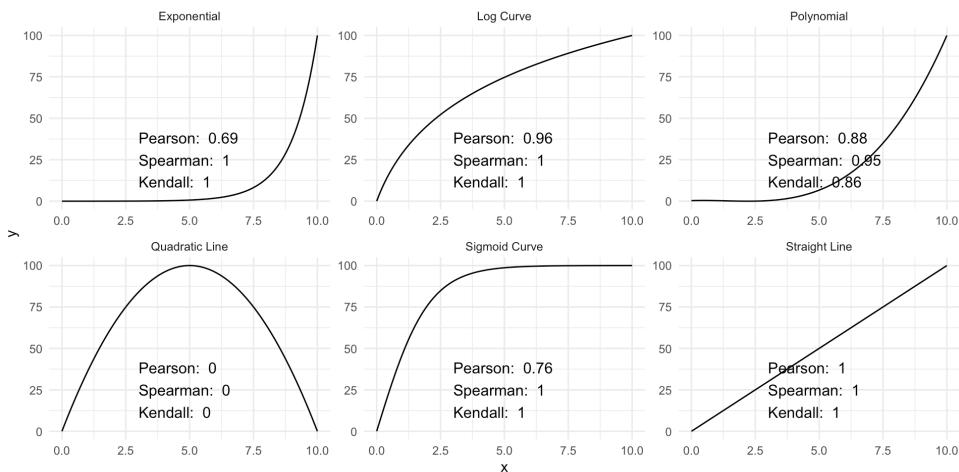
- **Monotonic**: a relationship that is consistently increasing or decreasing
- **Linear**: a relationship that is increasing or decreasing at a *constant rate* i.e. a straight line

Pearson's correlation is only for linear relationships.

Spearman's rank and **Kendall's tau** are correlation coefficients for all monotonic relationships.

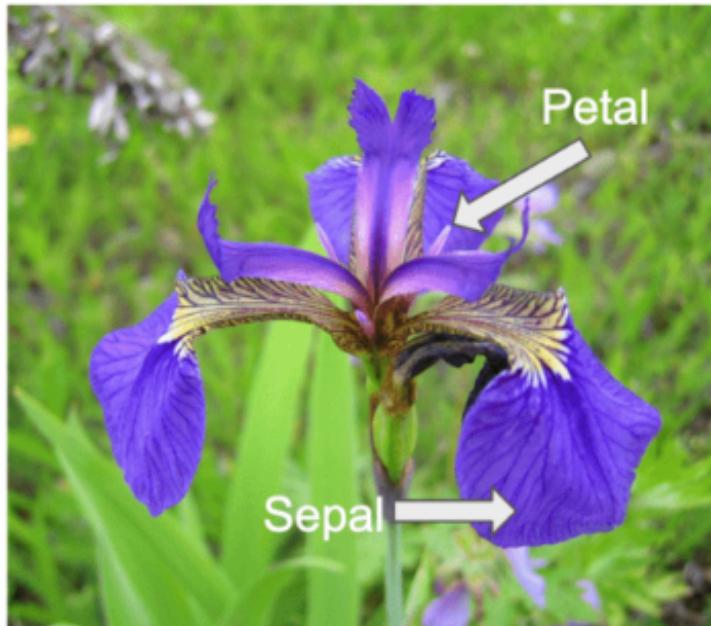
- Works with non-parametric data
- More 'conservative' i.e. values can be smaller in magnitude, but more robust against outliers

► Code

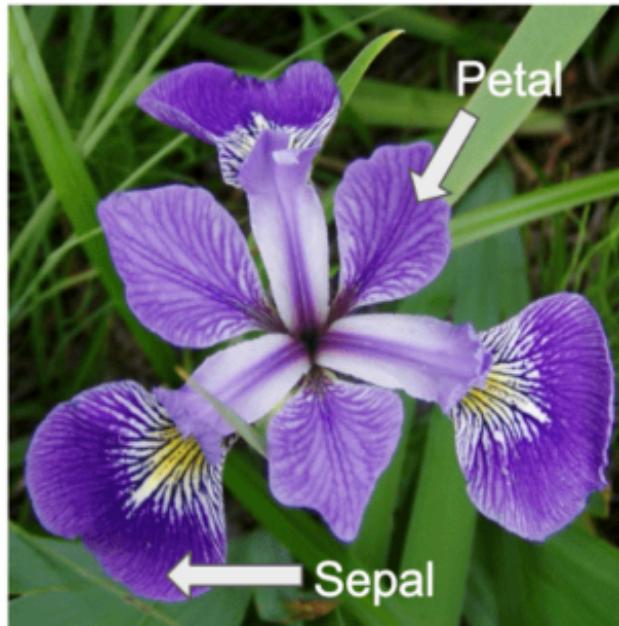


Example - iris Dataset

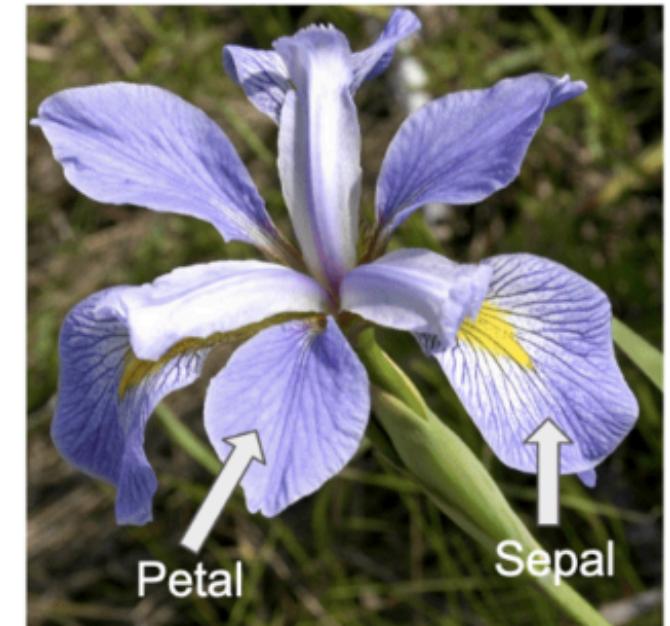
Iris setosa



Iris versicolor



Iris virginica



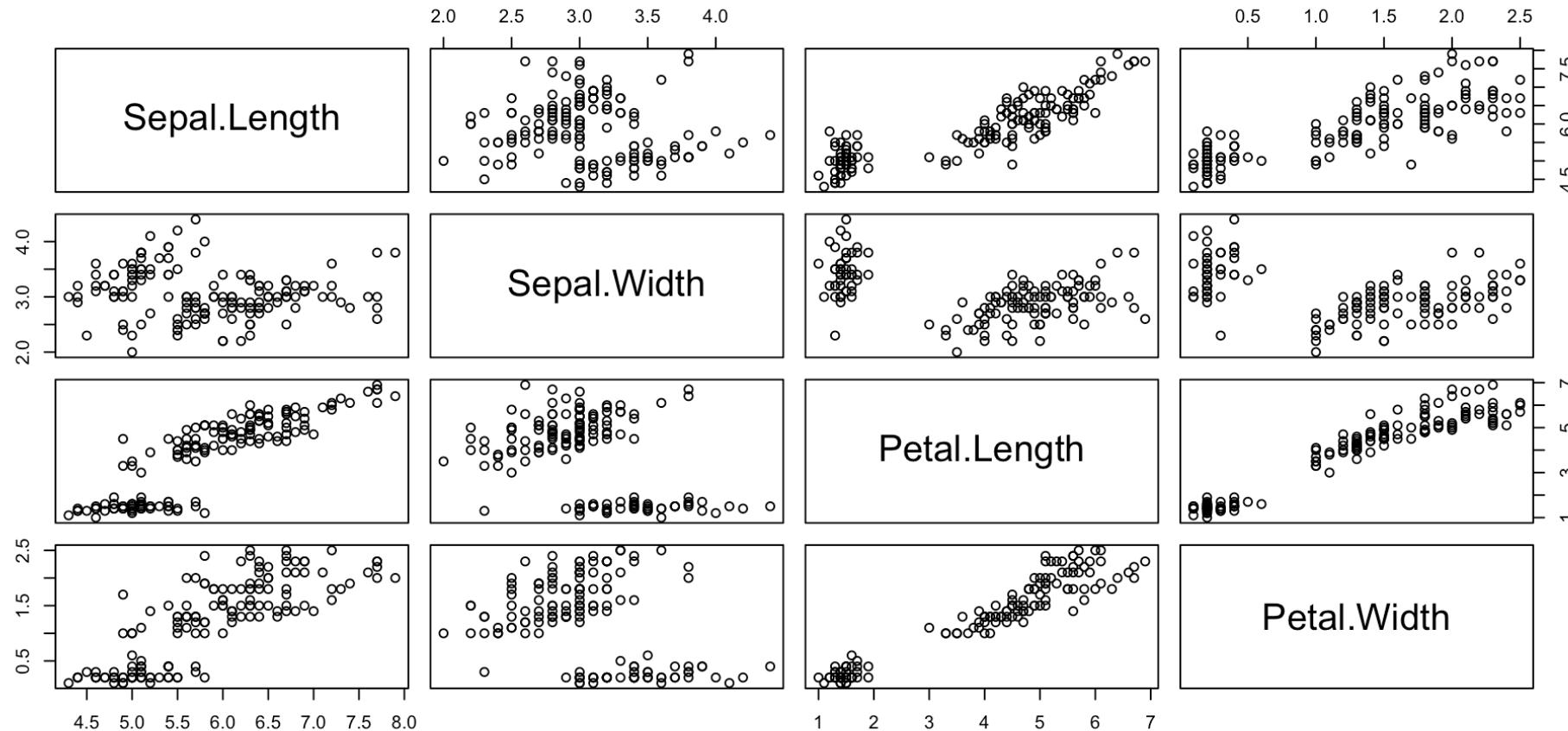
► Code

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Example - `iris` Dataset

Which correlation coefficient do we use?

► Code



Example - `iris` Dataset

► Code

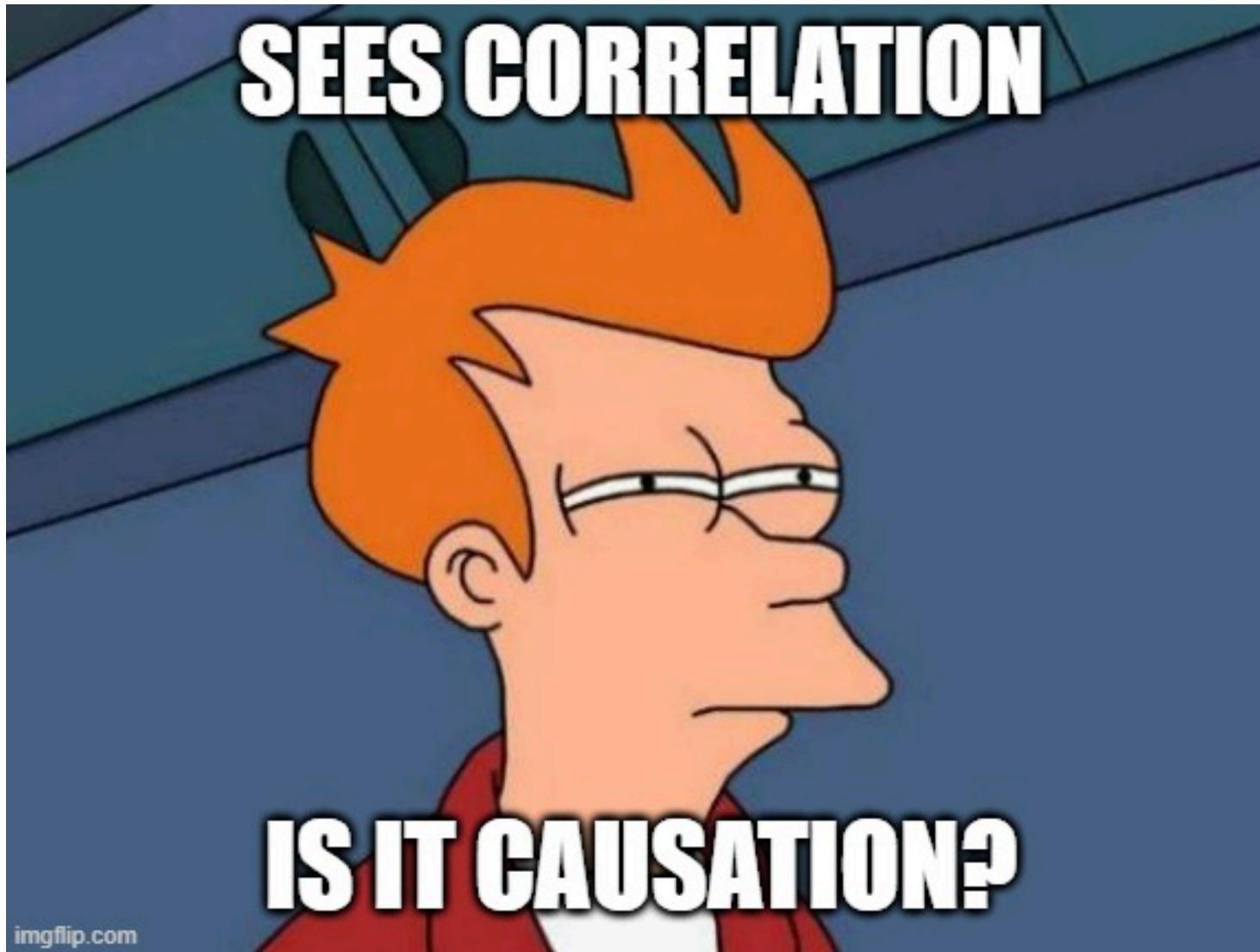
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.00	-0.12	0.87	0.82
Sepal.Width	-0.12	1.00	-0.43	-0.37
Petal.Length	0.87	-0.43	1.00	0.96
Petal.Width	0.82	-0.37	0.96	1.00

Correlation analysis **identifies** variables that have a strong linear relationship **quickly and easily**.

- We want to *predict* the `Petal.Width` of an iris flower
- Which variables would be *good predictors*?
 - ➡ Either `Petal.Length` and `Sepal Length` would be good predictors
- Which variables are likely to present issues for *model fitting*?
 - ➡ e.g. `Petal.Length` and `Sepal Length` are strongly correlated - including both would cause multicollinearity – (L11 MLR)

Correlation ≠ causation

Spurious correlations: a relationship between two variables does not imply that one causes the other.



What comes after correlation?

- We have data with two or more numerical variables
- We conduct correlation analysis to describe *possible* linear relationships
 - ➡ Fast, easy, interpretable, and widely used

But, we can't infer causation: - Is there *reason()* to expect a relationship between the two variables? - Do you have a hypothesis* about the relationship between the two variables?

If we have a hypothesis about the relationship between two variables, we can use **regression analysis** to test it.

Regression modelling

- Regression is a *statistical* method to fit a model to data
- In linear regression this is a straight line that best fits the data – i.e. line of best fit

Why regression?

Describe the relationship between two variables

What is the relationship between a response variable Y and a predictor variable x ?

Common Terms: Y = response, independent variable, target, outcome etc. x = predictor, dependent variable, feature, input etc.

Explain the relationship between two variables

How much variation in Y can be explained by a relationship with x ?

Predict the value of a response variable

What is the value of Y for a given value of x ?

Often we can easily measure/obtain x but not Y , so we need to *predict* Y from x .

A gateway to the world of modelling

Many types of regression models exist:

- **Simple linear regression (one predictor i.e. x)**
- Multiple linear regression (more than one predictor)
- Non-linear regression, using functions such as polynomials, exponentials, logarithms, etc.

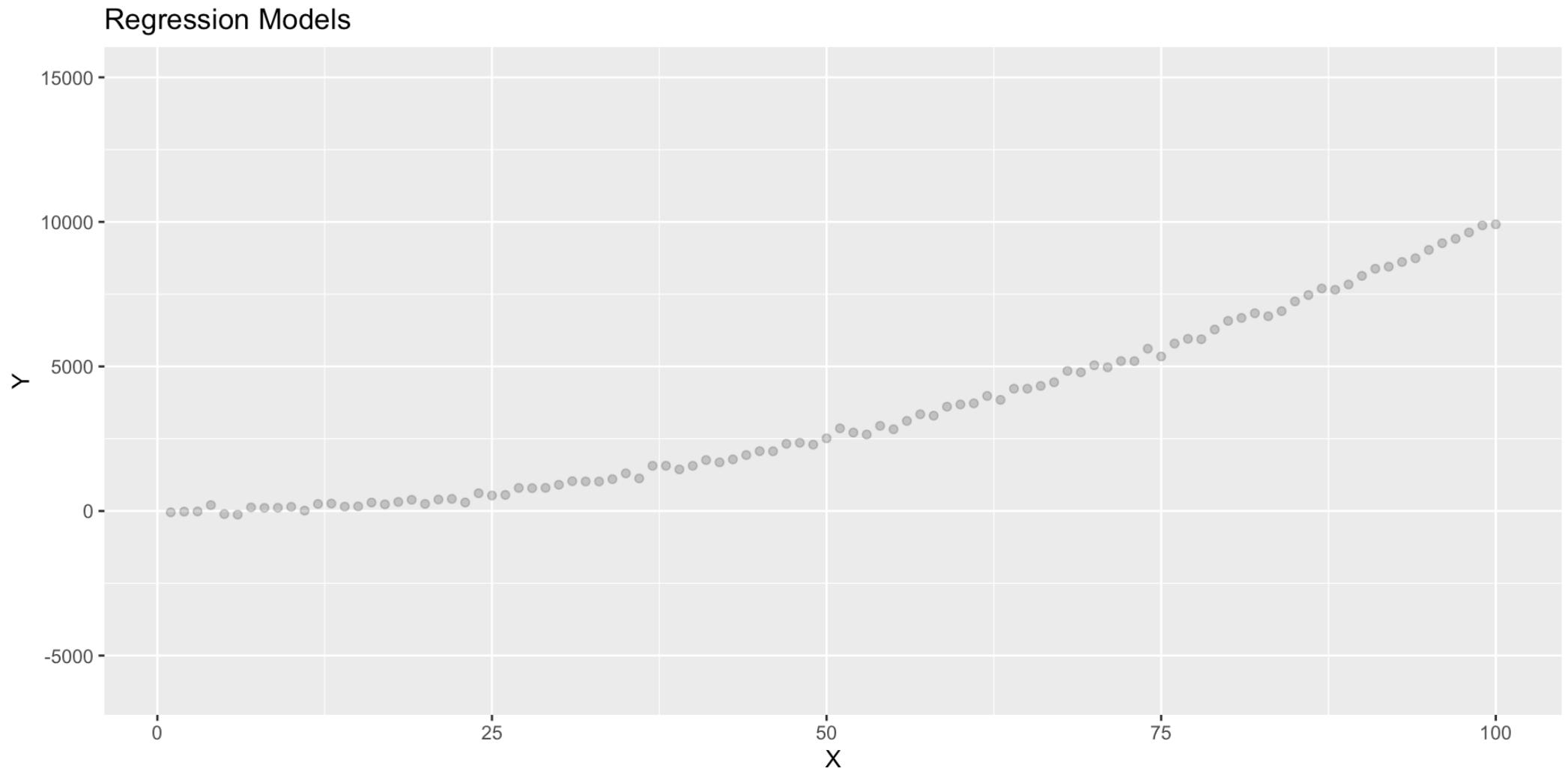
Asking ChatGPT for help with the next slide:

Using R code, can you generate some data that is useful to demonstrate simple linear regression, multiple linear regression, polynomial, exponential and logarithmic regressions in ggplot2?

Sure! Here's an example code that generates a sample dataset and visualizes it using ggplot2 library in R.

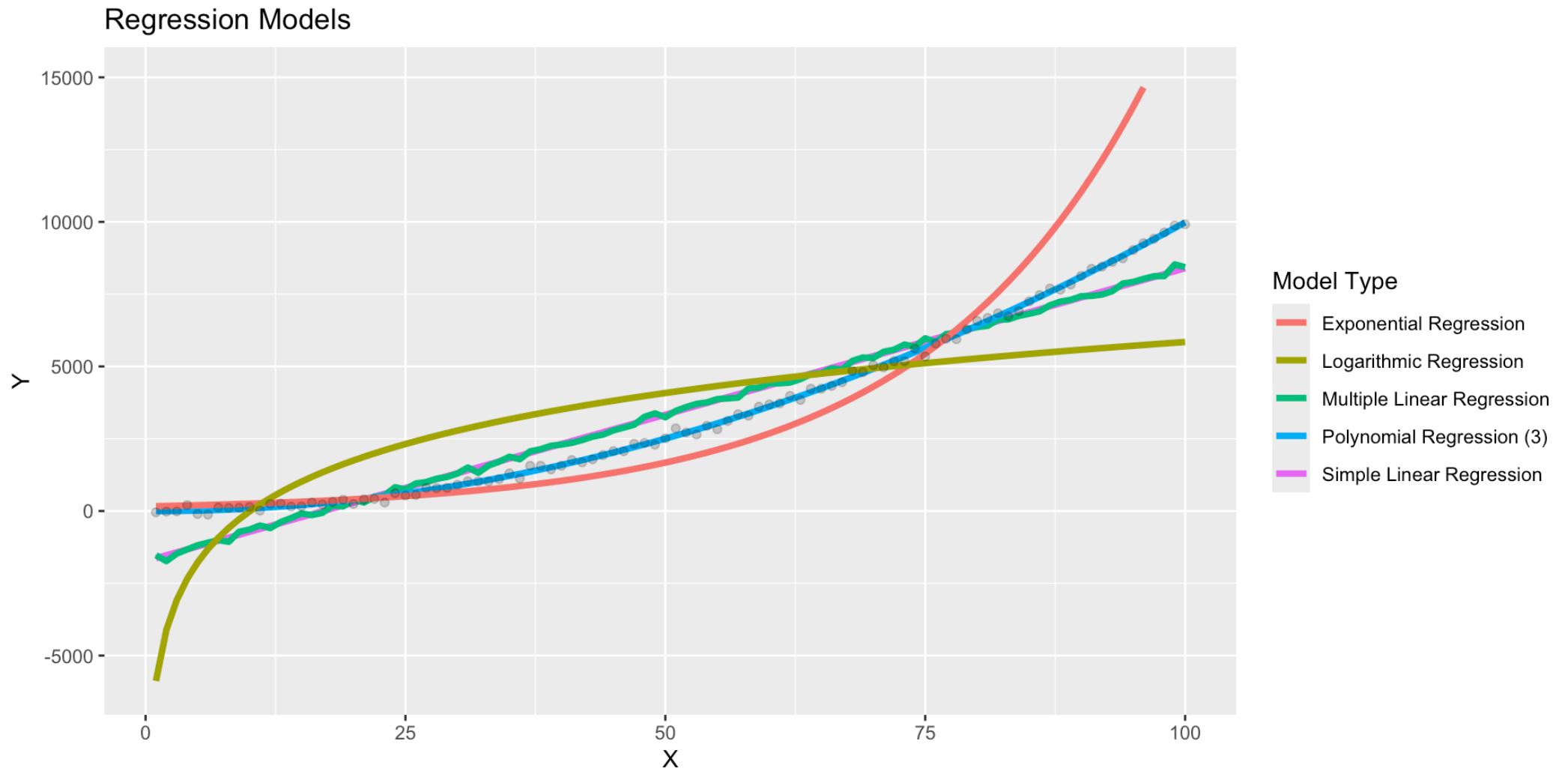
Visualising regression models

► Code

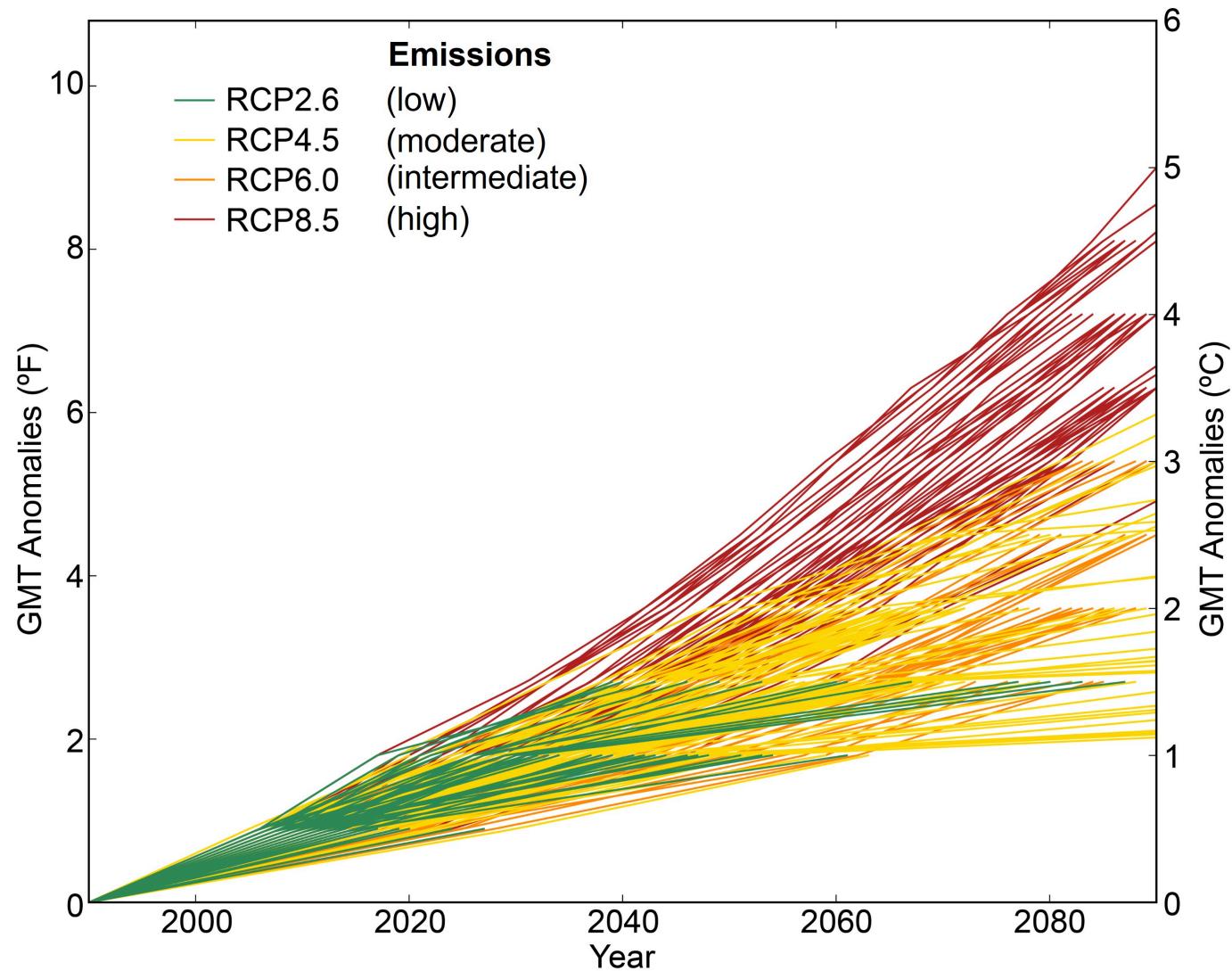


Visualising regression models

► Code

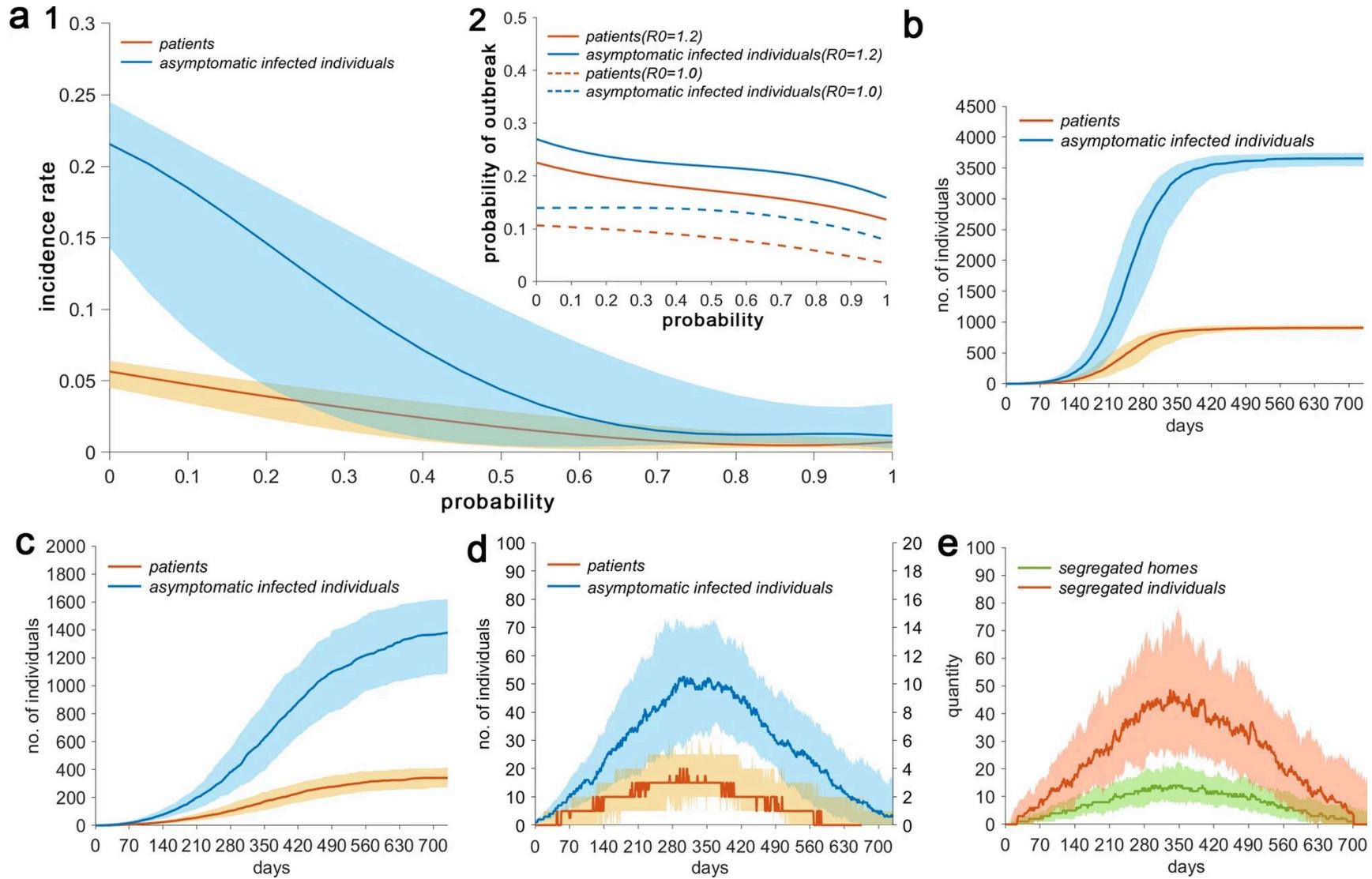


Example: climate change modelling



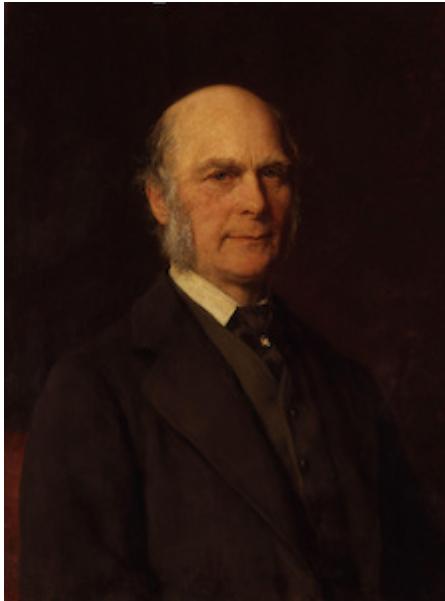
Source: <https://science2017.globalchange.gov/chapter/4/>

Example: COVID-19 transmission modelling



Source: <https://www.nature.com/articles/s41598-021-84893-4/figures/1>

How does regression work?



Adrien-Marie Legendre, Carl Friedrich Gauss, Francis Galton

Note

Many other people contributed to the development of regression analysis, but these three are the most well-known.

How does regression work?

- **Method of least squares** first theorised by Adrien-Marie Legendre in 1805
- **Technique of least squares** first used by Carl Friedrich Gauss in 1809 (to fit a parabola to the orbit of the asteroid Ceres)
- **Model fitting** first published by Francis Galton in 1886 (predicting the height of a child from the height of the parents)

Least squares

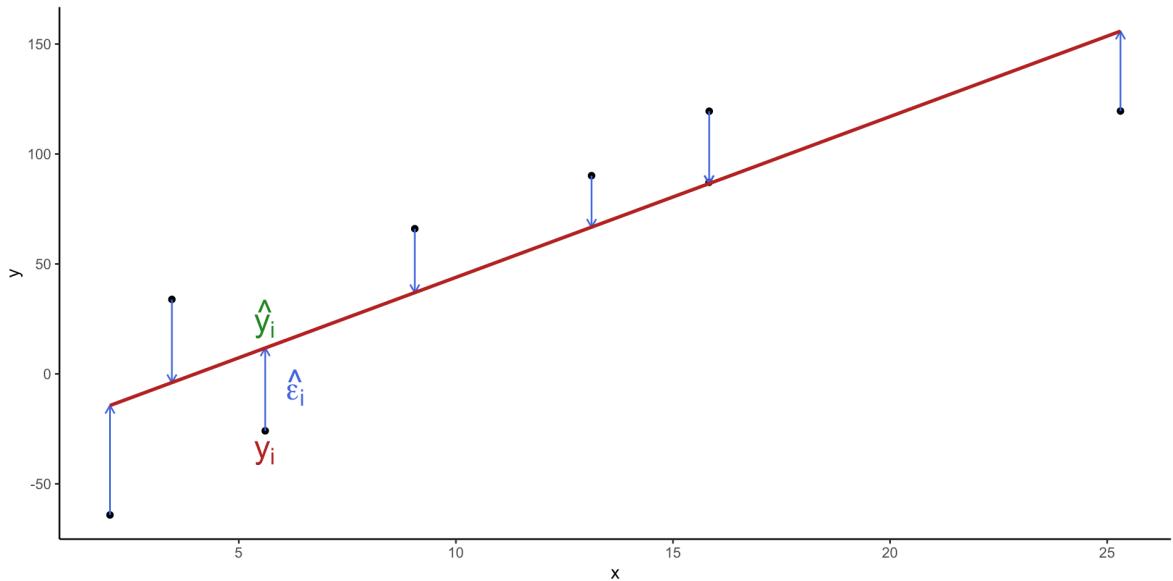
When we fit a line, there is error between the observed and predicted values.

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

The **method of least squares** fits a line of best fit by minimising the sum of the squared errors.

$$\sum_{i=1}^n (\hat{\epsilon}_i)^2$$

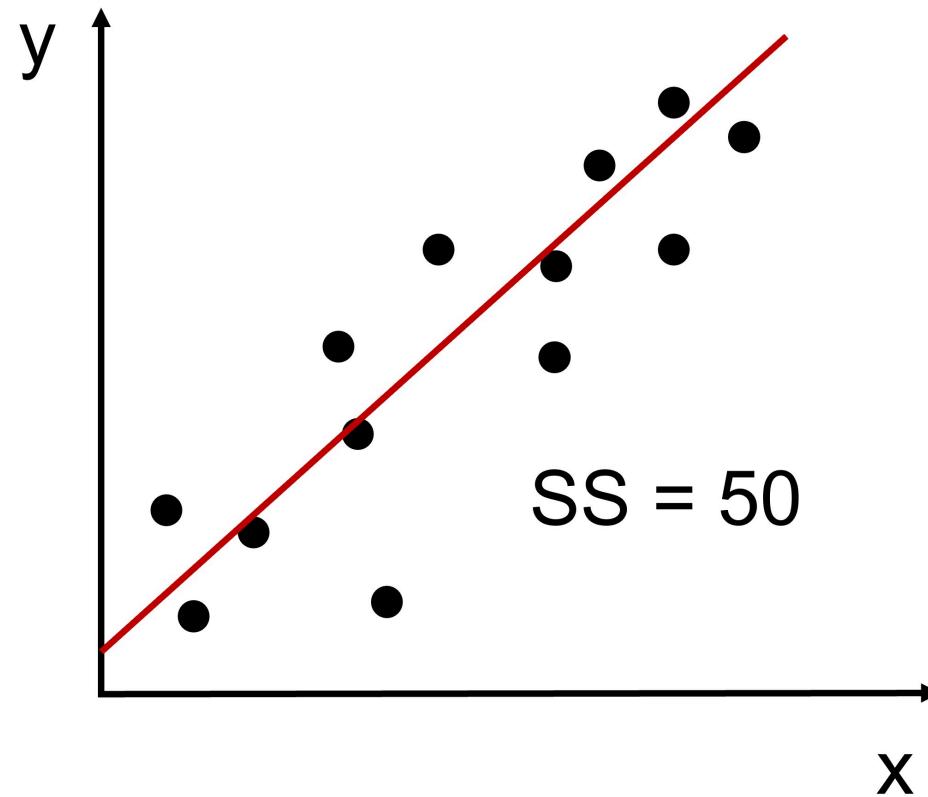
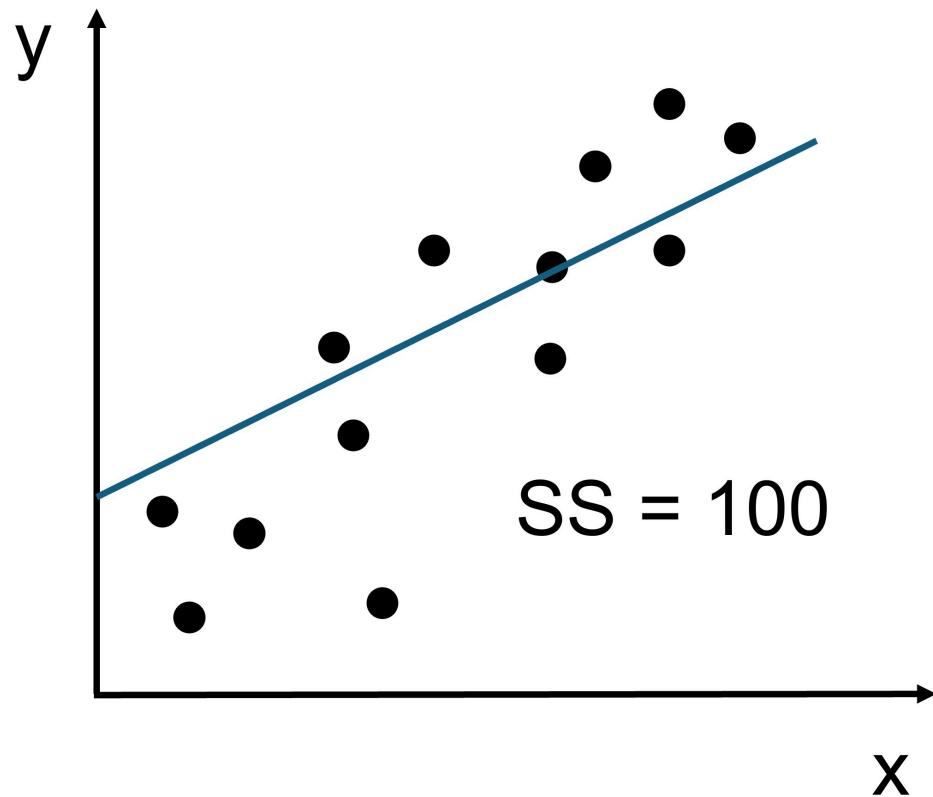
► Code



(i) Note

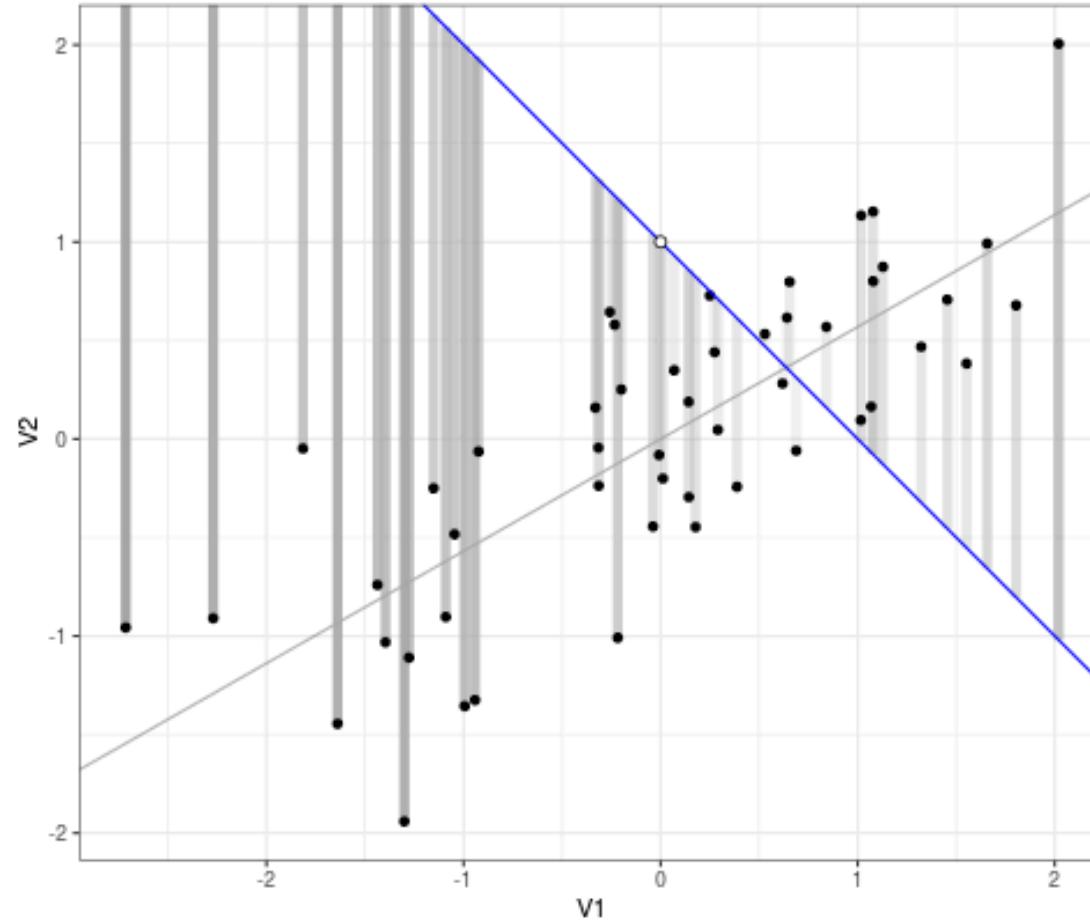
The error is squared so positive and negative errors do not cancel each other out.

Least squares



Which line fits the data better?

How does a computer fit a line of best fit?



The line is fitted again and again until the squared error stabilises. A computer can do this very quickly!

Simple linear regression

The goal is to fit the line of best fit between a numerical response and a numerical predictor.

Simple linear regression model

We want to relate the response \mathbf{Y} to a predictor \mathbf{x} for i number of observations:

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

where

$$\epsilon_i \sim N(0, \sigma^2)$$

- Y_i , the *response*, is an observed value of the dependent variable.
- β_0 , the *constant*, is the y-intercept and is **fixed**.
- β_1 is the population *slope* parameter, and like β_0 , is also **fixed**.
- ϵ_i is the error associated with predictions of y_i , and unlike β_0 or β_1 , it is *not fixed*.

(i) Note

ϵ_i is generally associated with the **residual** (*observed – predicted*). **True error** occurs during data collection (e.g. faulty instruments, selection bias, etc.) and often immeasurable.

Different wordings

Different ways to think about the response:

- Response = Prediction + Error
- Response = Signal + Noise
- Response = Model + Unexplained
- Response = Deterministic + Random
- Response = Explainable + Everything else
- $Y = f(x)$
- Dependent variable = $f(\text{Independent variable})$

Model fitting

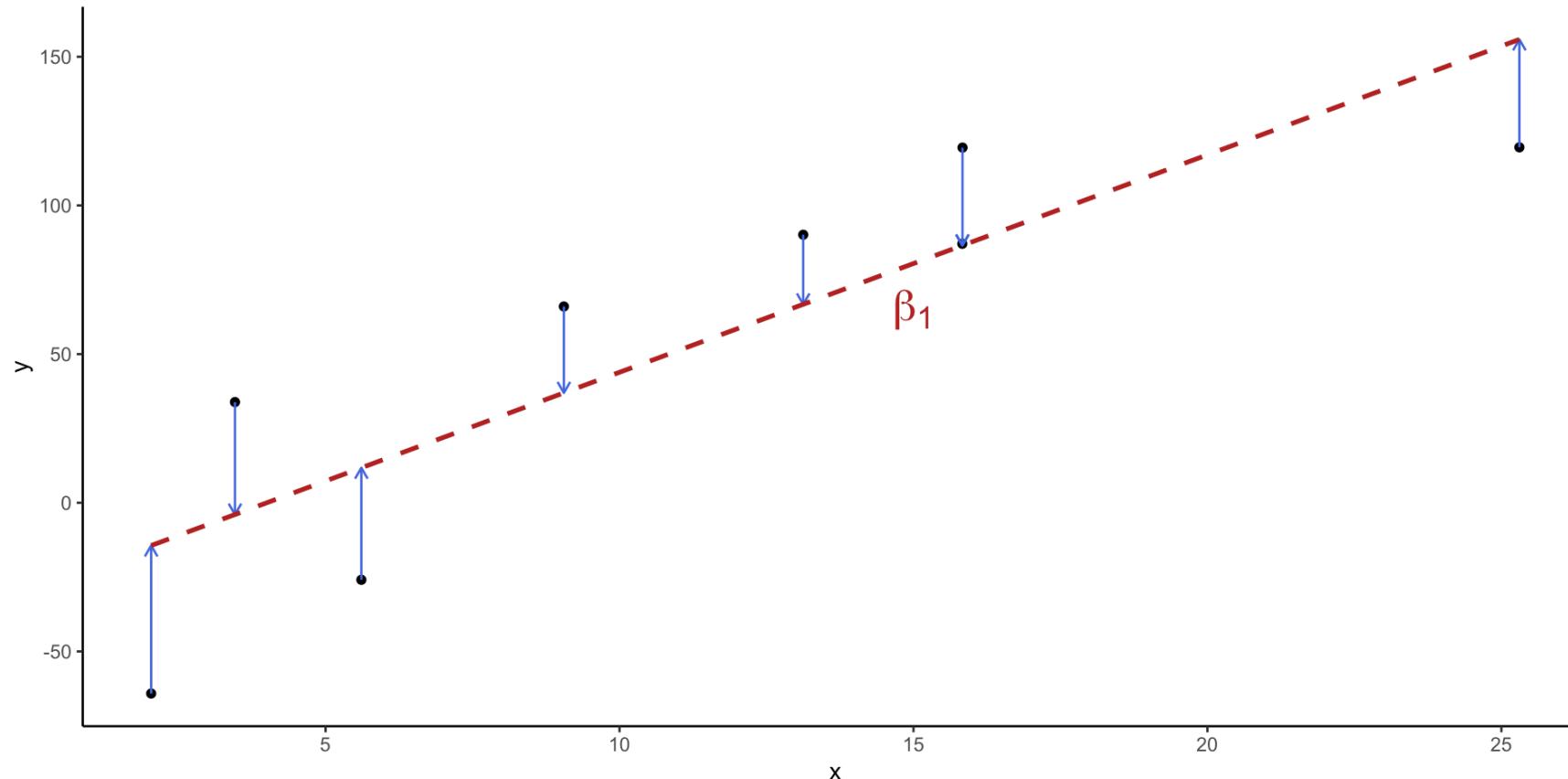
Two approaches; analytical and numerical:

- **Analytical**: equation(s) used directly to find solution
- **Numerical**: computer uses “random guesses” to find set of parameters to that minimises objective function, in this case residual sum of squares

Analytical: Slope, β_1

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{Cov(x, y)}{Var(x)} = \frac{SS_{xy}}{SS_{xx}}$$

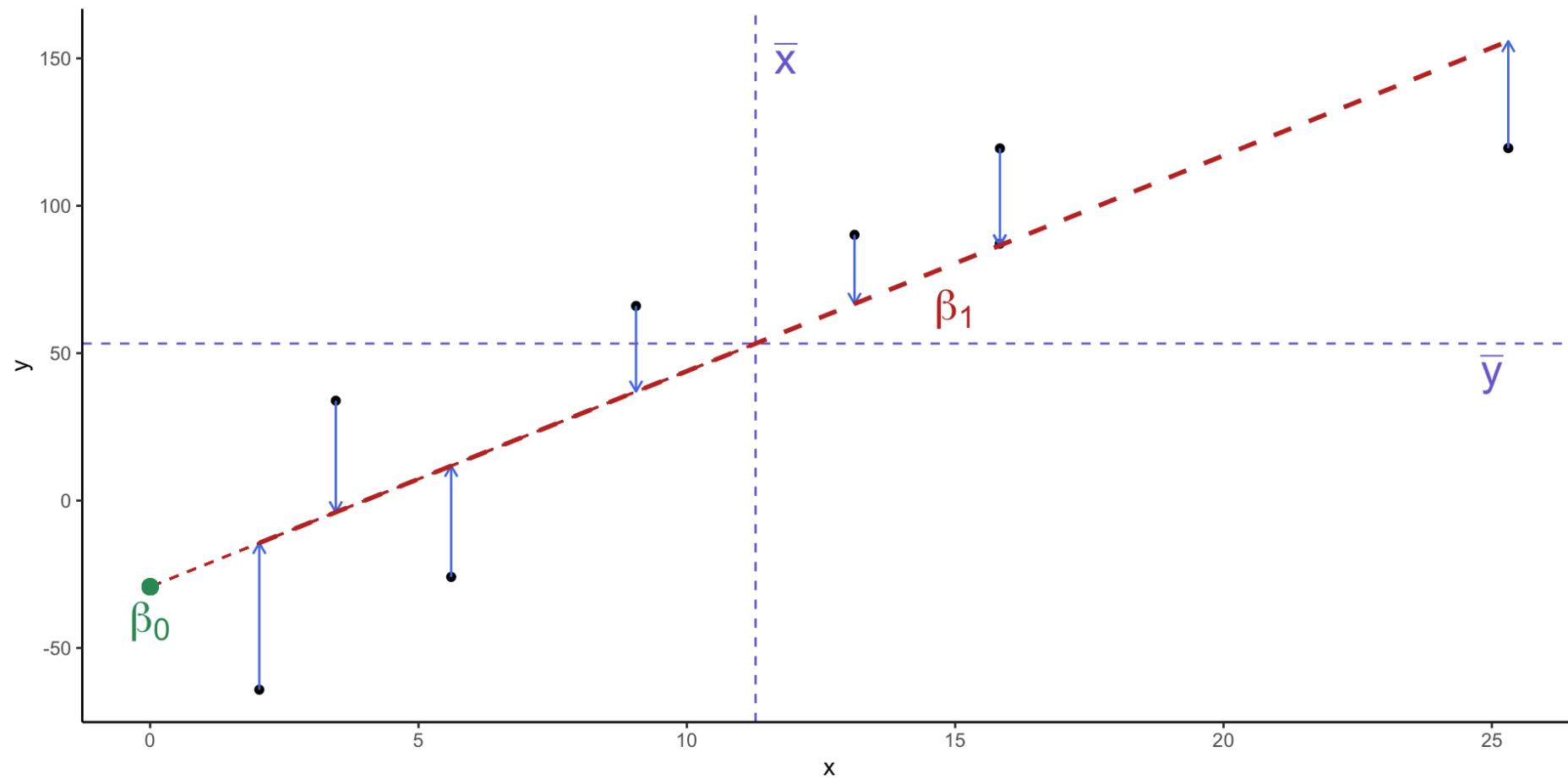
► Code



Analytical: Intercept

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

► Code



Example: back to Galton's data

What can we understand about the relationship between child height and parent height?

Let's do this in R - much easier

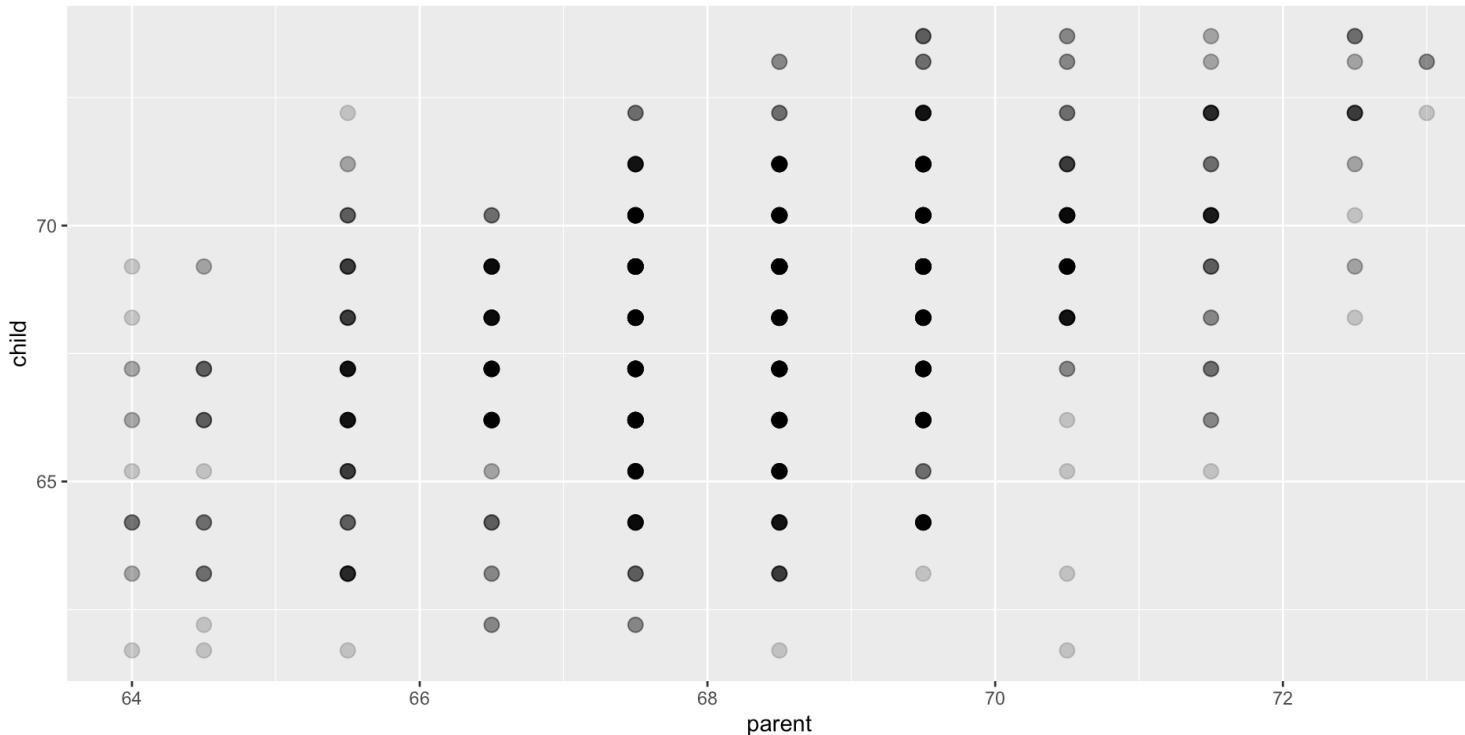
Linearity check

Recap - there appears to be a linear relationship between child height and parent height (plot). This is a moderately positive relationship (correlation).

► Code

► Code

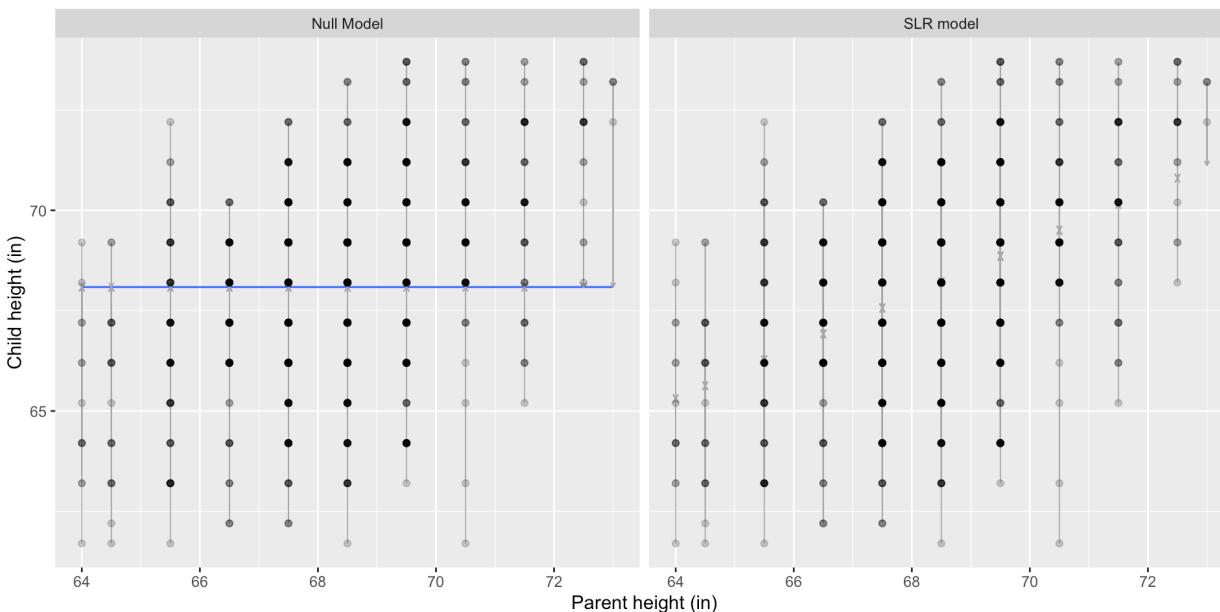
```
[1] 0.4587624
```



Hypothesis testing

- The **null hypothesis** for a linear model: $H_0 : \beta_1 = 0$
 - ➡ If there is no slope ($\beta_1 = 0$) then $y = \beta_0$ (the mean)
 - ➡ i.e. is the mean (\bar{y}) a better fit than a linear model?
- The **alternative hypothesis** for a linear model: $H_0 : \beta_1 \neq 0$
 - ➡ i.e. the estimate from the linear model (\hat{y}) fits the data better than the mean (\bar{y})

► Code



Simple linear regression

We fit the model in one simple line of code: `fit ← lm(child ~ parent, data = Galton)`

► Code

And then we can use `summary()` to get a summary of the model:

► Code

```
Call:  
lm(formula = child ~ parent, data = Galton)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-7.8050 -1.3661  0.0487  1.6339  5.9264  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 23.94153    2.81088   8.517 <2e-16 ***  
parent       0.64629    0.04114  15.711 <2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2.239 on 926 degrees of freedom  
Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096  
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

Summary

Correlation is a measure of the relationship between two numerical variables between -1 and 1.

- A correlation coefficient measures the *strength* and *direction* of the relationship.
- Pearson's correlation is for linear relationships, Spearman's/Kendall's correlation is for monotonic relationships.
- Correlation \neq causation.

Regression is an analysis that models the relationship between a dependent variable and independent variable(s).

- The most common method for regression is least squares, which minimises the sum of the squared residuals.
- Simple linear regression fits a straight line between two variables.

Next week: interpreting results and assumptions for simple linear regression.

Thanks!

This presentation is based on the [SOLES Quarto reveal.js template](#) and is licensed under a [Creative Commons Attribution 4.0 International License](#).