Topic 9 – Describing relationships

ENVX1002 Introduction to Statistical Methods

Liana Pozza
The University of Sydney

Dec 2024

THE UNIVERSITY OF SYDNEY

Quick intro

About me

- Lecturer in Agricultural Data Science
- Spatial modelling and mapping, Soil science, Precision Agriculture
- This year back to being a student again, GradCert (Higher Education)

Narrabri cotton field

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.

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

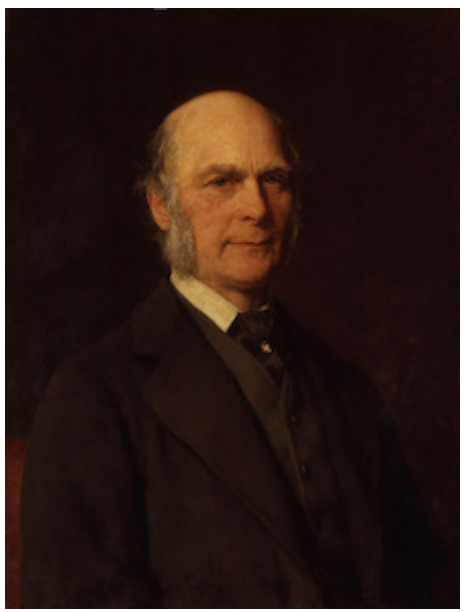
- **Week 9. Describing relationships**
  - Correlation → calculation, interpretation, things to watch out for
  - Regression → Why do we care? model structure, model fitting
- **Week 10. Linear functions**
  - Is the model worth fitting? → Assumptions, hypothesis testing
  - How good is the model? → Measures of model fit
- **Week 11. Linear functions - multiple predictors**
  - Parsimonious models
  - Introduction to Multiple Linear Regression (MLR) modelling
  - Assumptions and interpretation
- **Week 12. Nonlinear functions**
  - Common nonlinear functions
  - Transformations
  - Performing nonlinear regression

Module overview

- **Week 9. Describing relationships**
  - Correlation → calculation, interpretation, things to watch out for
  - Regression → Why do we care? model structure, model fitting

Brief history






Adrien-Marie Legendre, Carl Friedrich Gauss, Francis Galton, & Karl Pearson

Least squares, correlation and astronomy

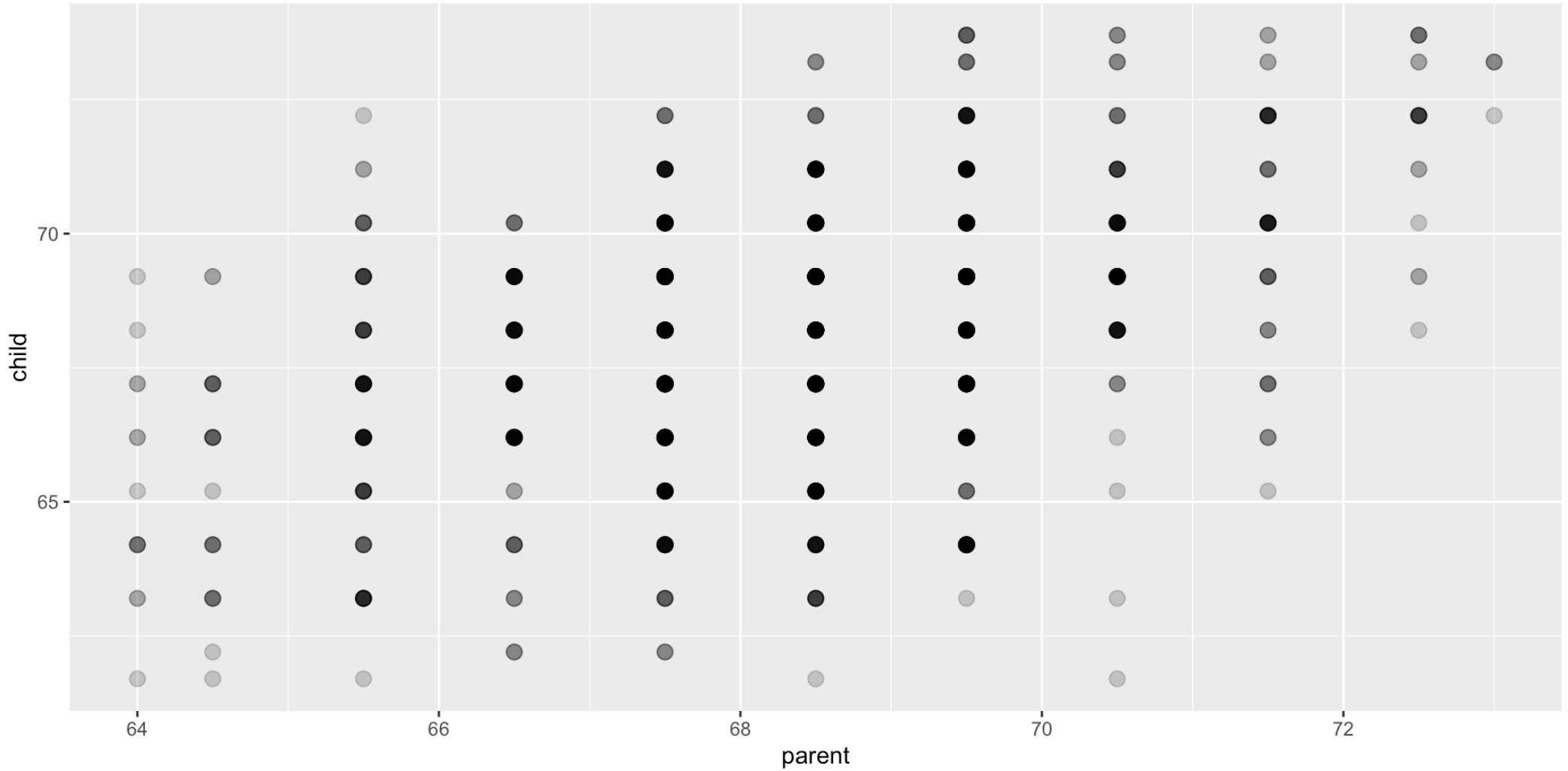
- Method of least squares first **published** paper by Adrien-Marie Legendre in 1805
- Technique of least squares used by Carl Friedrich Gauss in 1809 to fit a parabola to the orbit of the asteroid Ceres
- Model fitting first **published** by Francis Galton in 1886 to the problem of predicting the height of a child from the height of the parents
- **Karl Pearson developed the correlation coefficient in 1800s based on the work by Francis Galton**

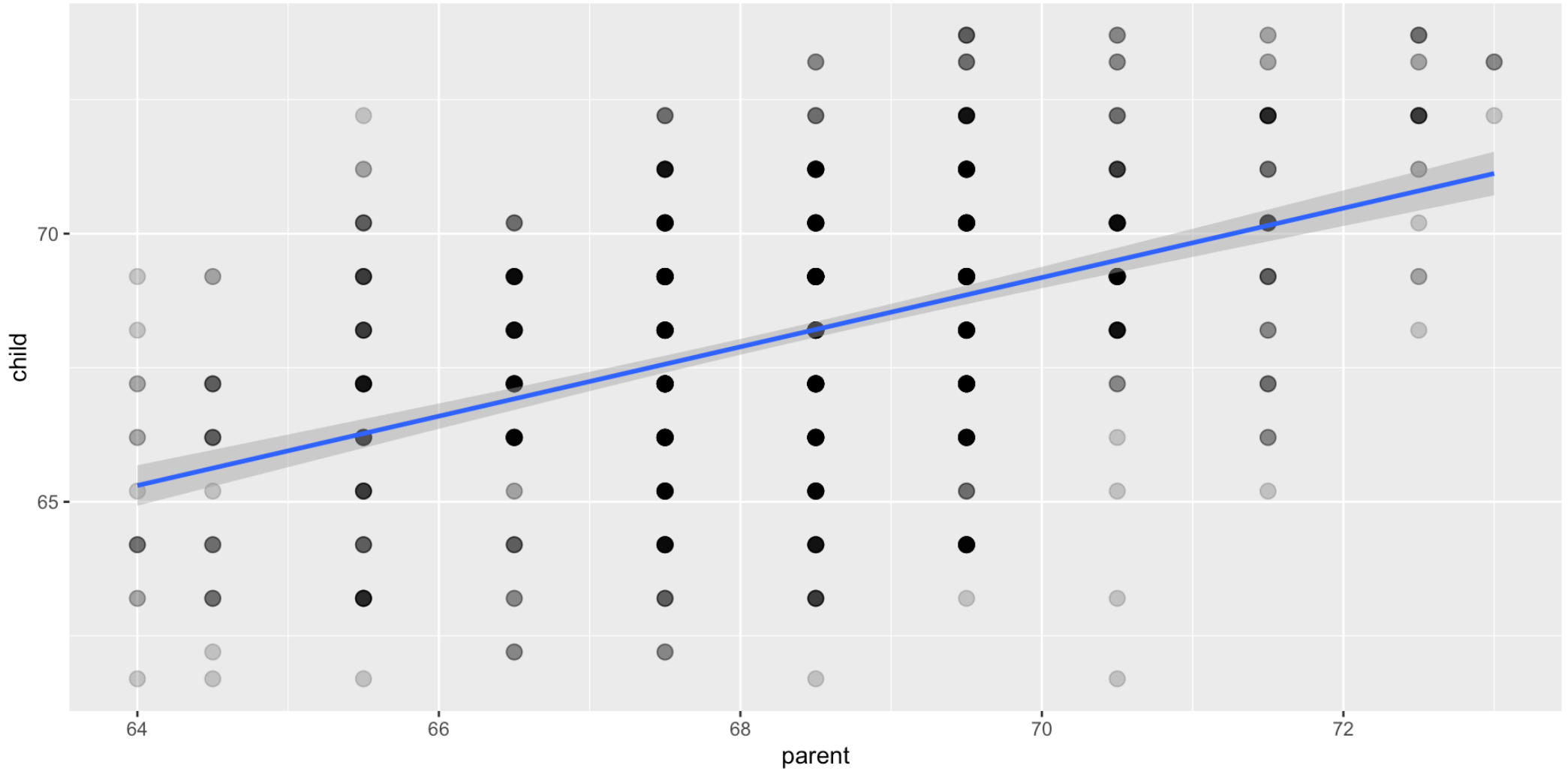
> **Note**
>
> Many other people contributed to the development of regression analysis, but these three are the "most" well-known.

Galton's data

```
# A tibble: 928 × 2
   parent child
    <dbl> <dbl>
 1   70.5  61.7
 2   68.5  61.7
 3   65.5  61.7
 4   64.5  61.7
 5   64    61.7
 6   67.5  62.2
 7   67.5  62.2
 8   67.5  62.2
 9   66.5  62.2
10   66.5  62.2
# ℹ 918 more rows
```

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


child
70
65
64
66
68
70
72
parent

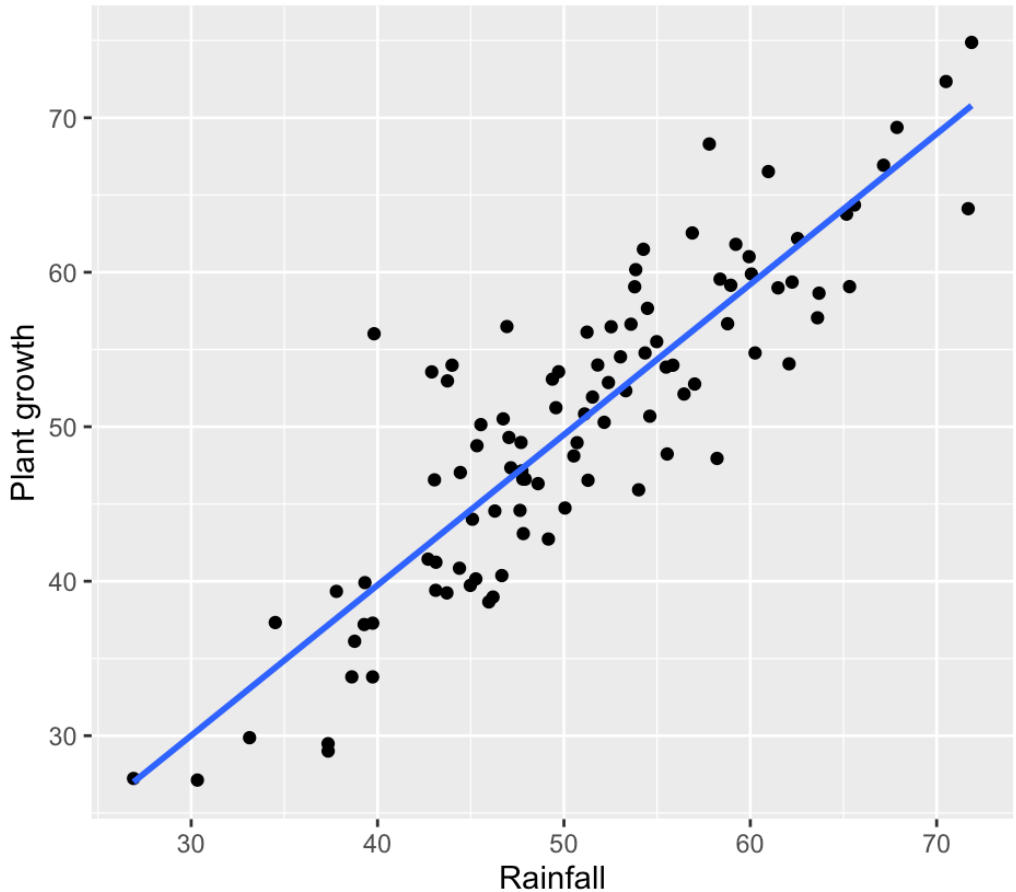

child
70
65
64
66
68
70
72
parent

Why do we care?

- Correlations are useful for describing relationships between two **continuous** variables
  - **Direction**:
    - positive - both variables increase together
    - negative - one variable increases as the other decreases
  - **Strength**: weak, moderate, strong – *subjective*, but useful for *describing* the relationship

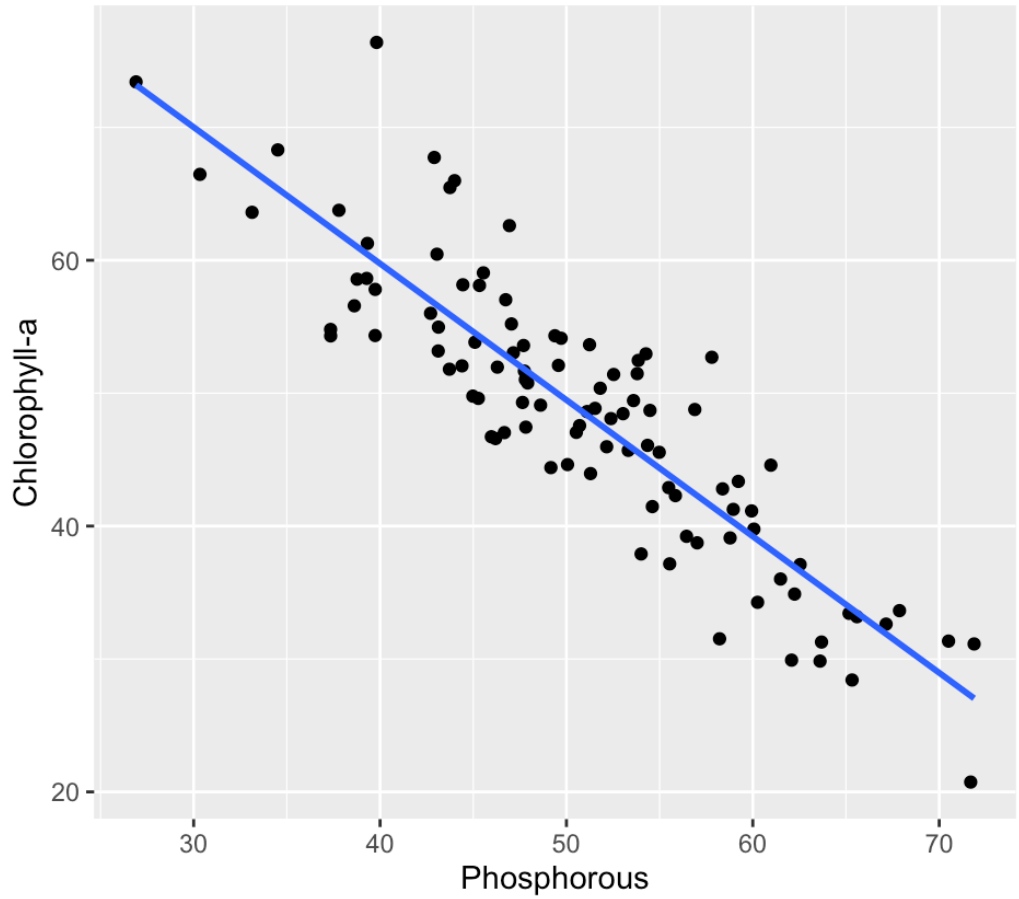
Positive correlation

Correlation coefficient: 0.88



Negative correlation

Correlation coefficient: -0.89

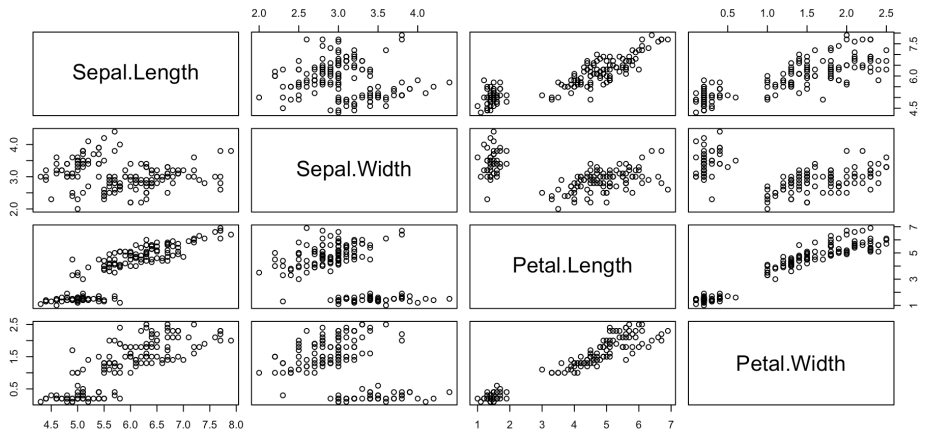


What can correlation analysis be used for?

- **Describing** the *possible* linear relationship between two variables
- Used *extensively* in exploratory data analysis: because it is *fast* and *easy* to calculate.

E.g. what is the correlation between all continuous variables in the `iris` dataset?

```
             Sepal.Length Sepal.Width Petal.Length
Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538
0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401
-0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000
0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654
1.0000000
```



- We can essentially use correlation analysis to *identify* variables that are useful for *predicting* another variable, or if they will present issues for *model fitting* i.e. multicollinearity.

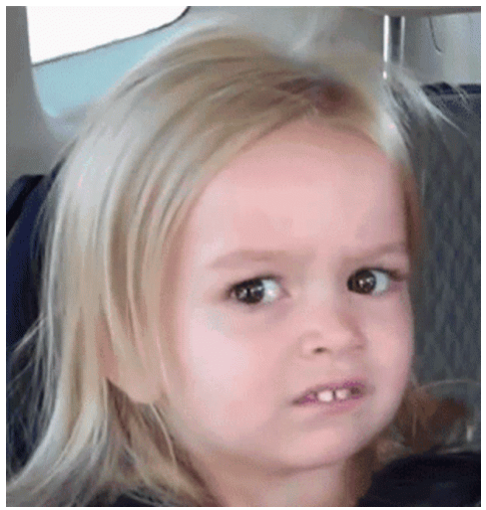
Different types of correlation coeefficients

- Parametric (normally distributed data):
  - Pearson correlation coefficient
    - most commonly used
- Non-parametric (non-normally distributed data):
  - Spearman rank correlation coefficient
  - Kendall's tau
  - more conservative i.e. values are often *smaller*, but more robust to outliers

Pearson correlation coefficient

Formula:

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



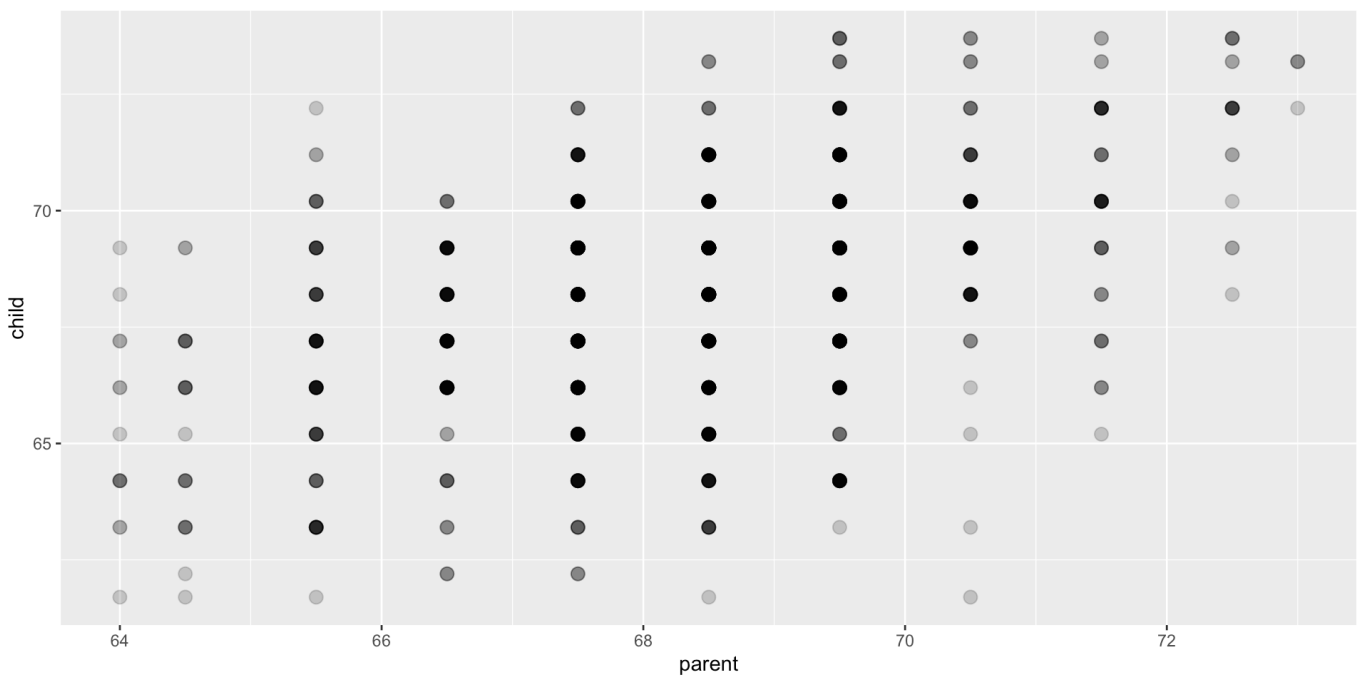
> covariance divided by the product of the standard deviations

Luckily we have Excel and R

- Excel: `=CORREL()` formula, or use the analysis toolpak
- R: `cor()` function

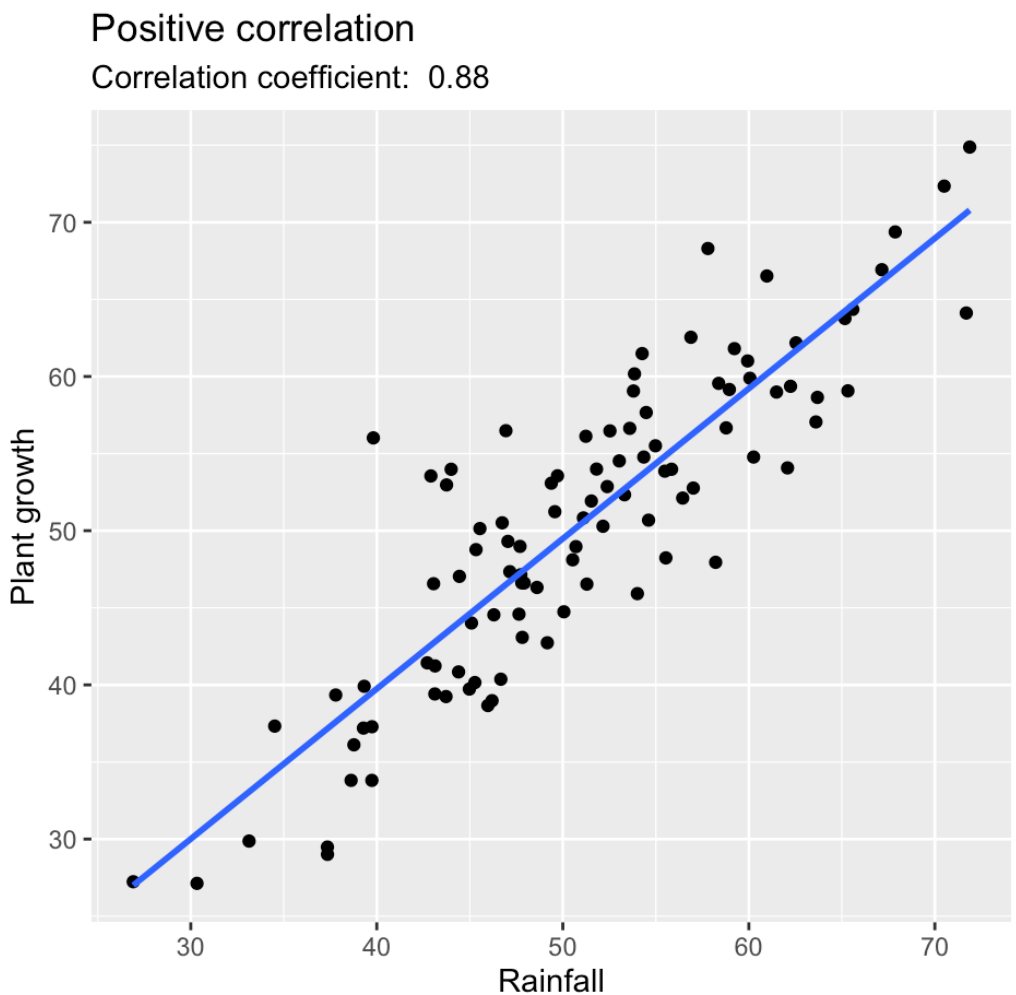
```
[1] 0.4587624
```

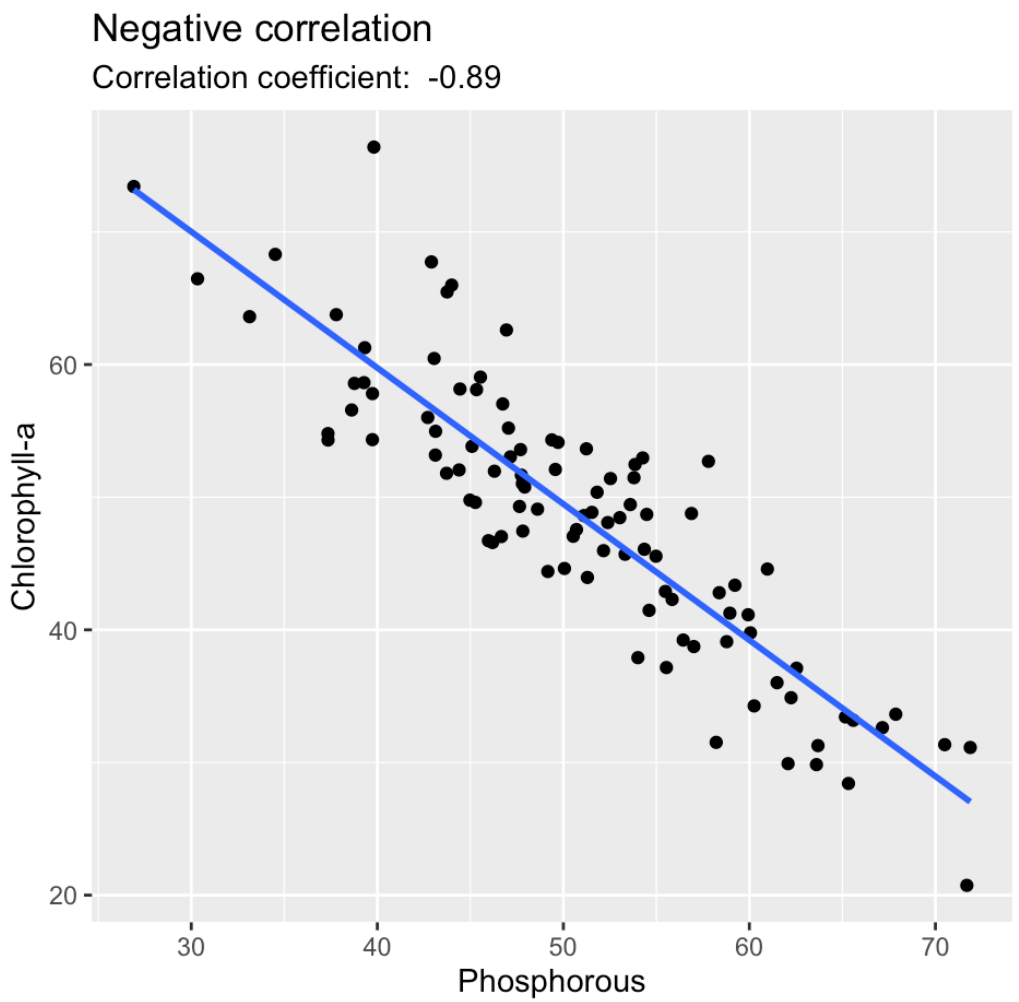
We can also visually inspect the relationship between the two variables using a scatterplot:



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

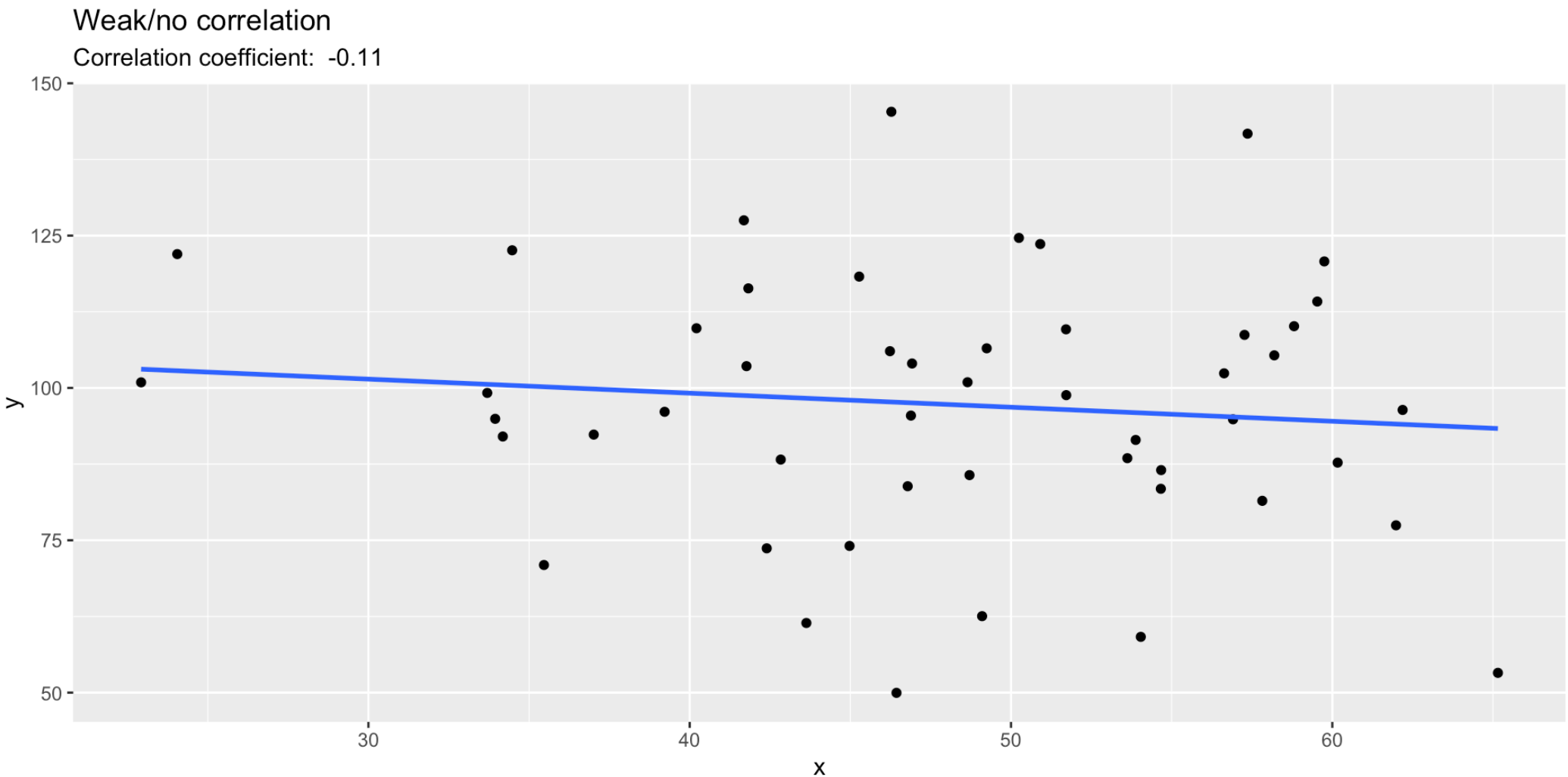
Interpretation: strong





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

Interpretation: weak



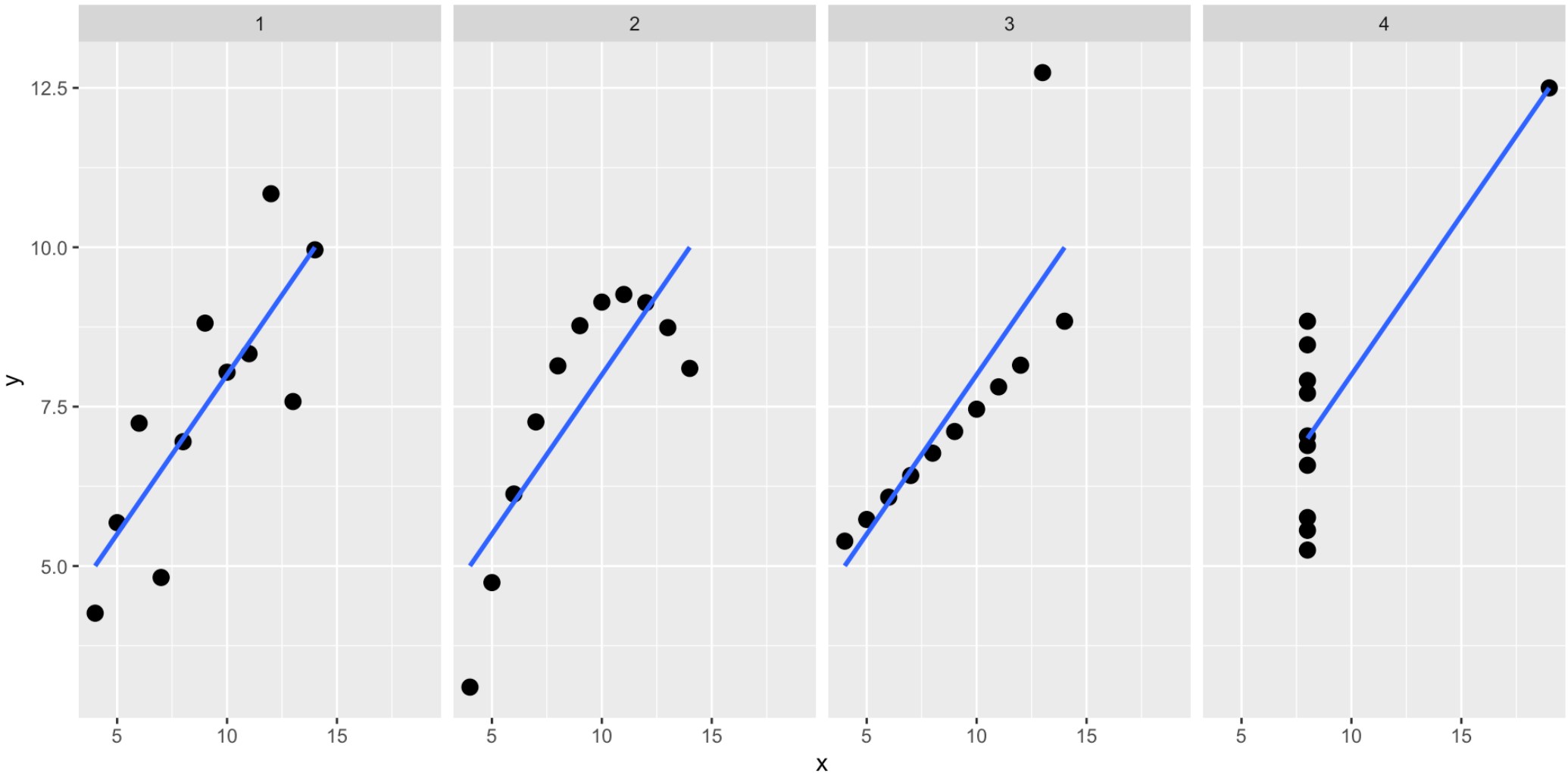
A **weak** or **nonexistent** relationship is one where the correlation coefficient is close to 0.

Interpretation: numbers

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

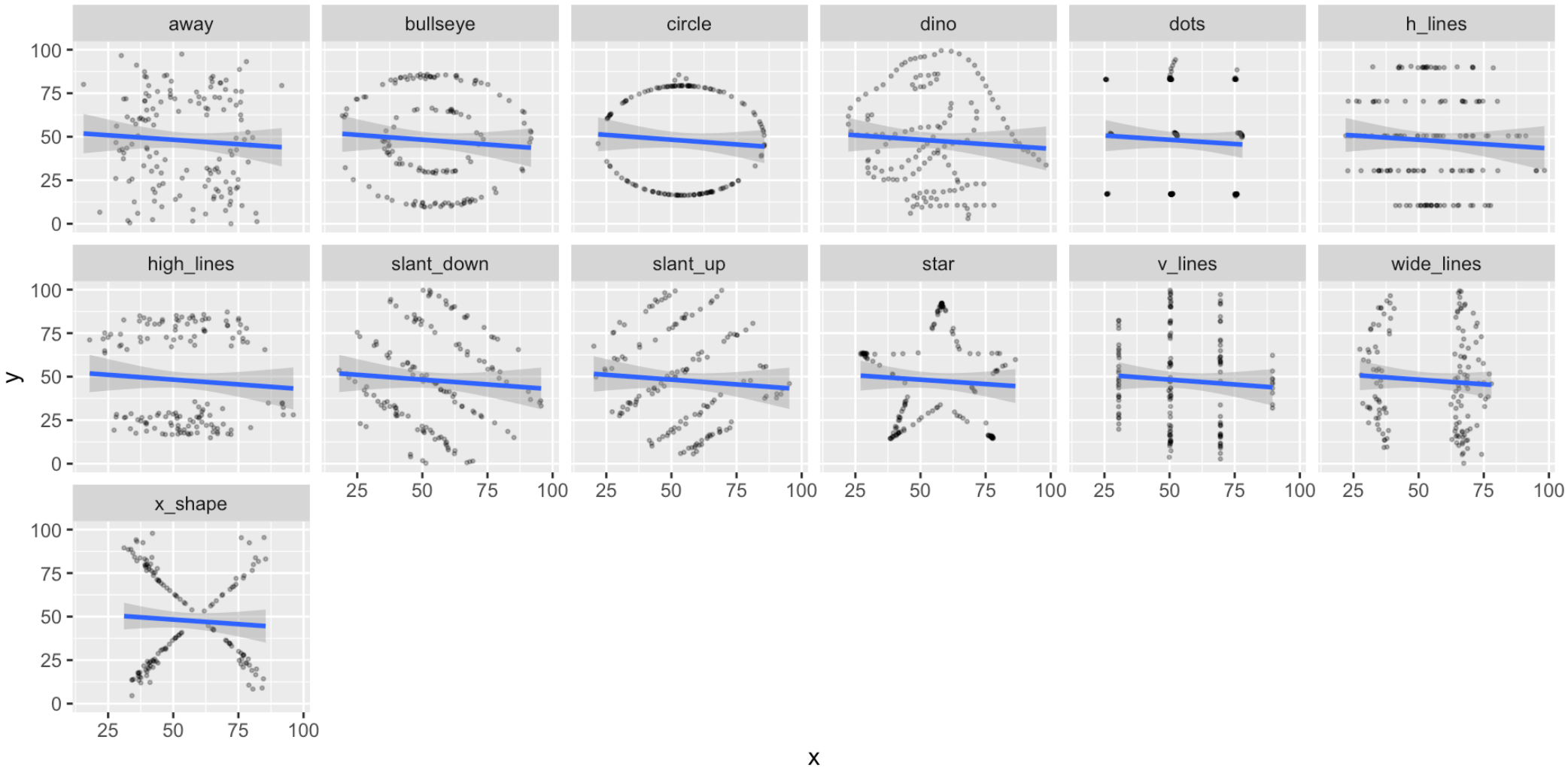
- It is enough to deduce the *strength* of the relationship from the correlation coefficient value(s) alone.
- **Not reliable** for *inference* about the relationship between variables. For this **you must visualise**.

Anscombe’s quartet



All of these data have a correlation coefficient of about 0.8.

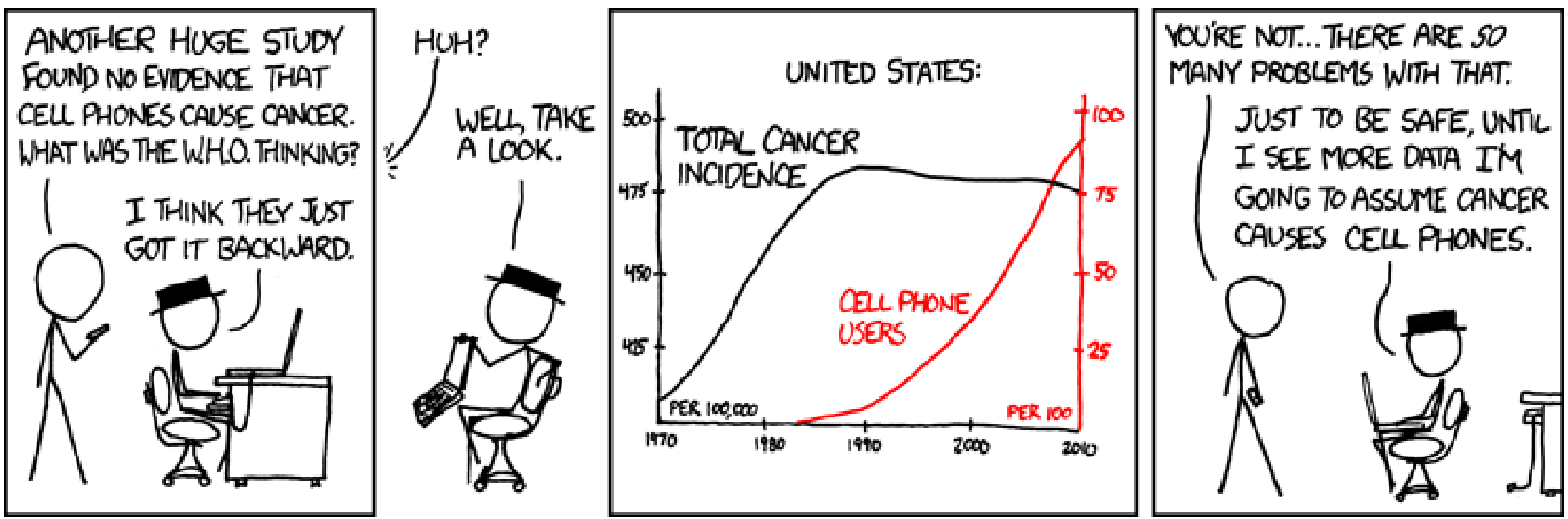
Datasaurus Dozen



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

Correlation ≠ causation

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



What comes after correlation?

- First, a summary:
  - Correlation analysis is a *fast* and *easy* way to *describe* the *possible* linear relationship between two variables
  - But, we can't infer causation: **domain knowledge is required to interpret the results**
    - Are we expecting 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 analysis is a *statistical* method for *predicting* an outcome based on a predictor variable
- We can *also* use regression analysis to **test our hypothesis** about the relationship between two variables

Why regression?

Describe the relationship between two variables

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

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?

A gateway to the world of modelling

Many types of regression models exist:

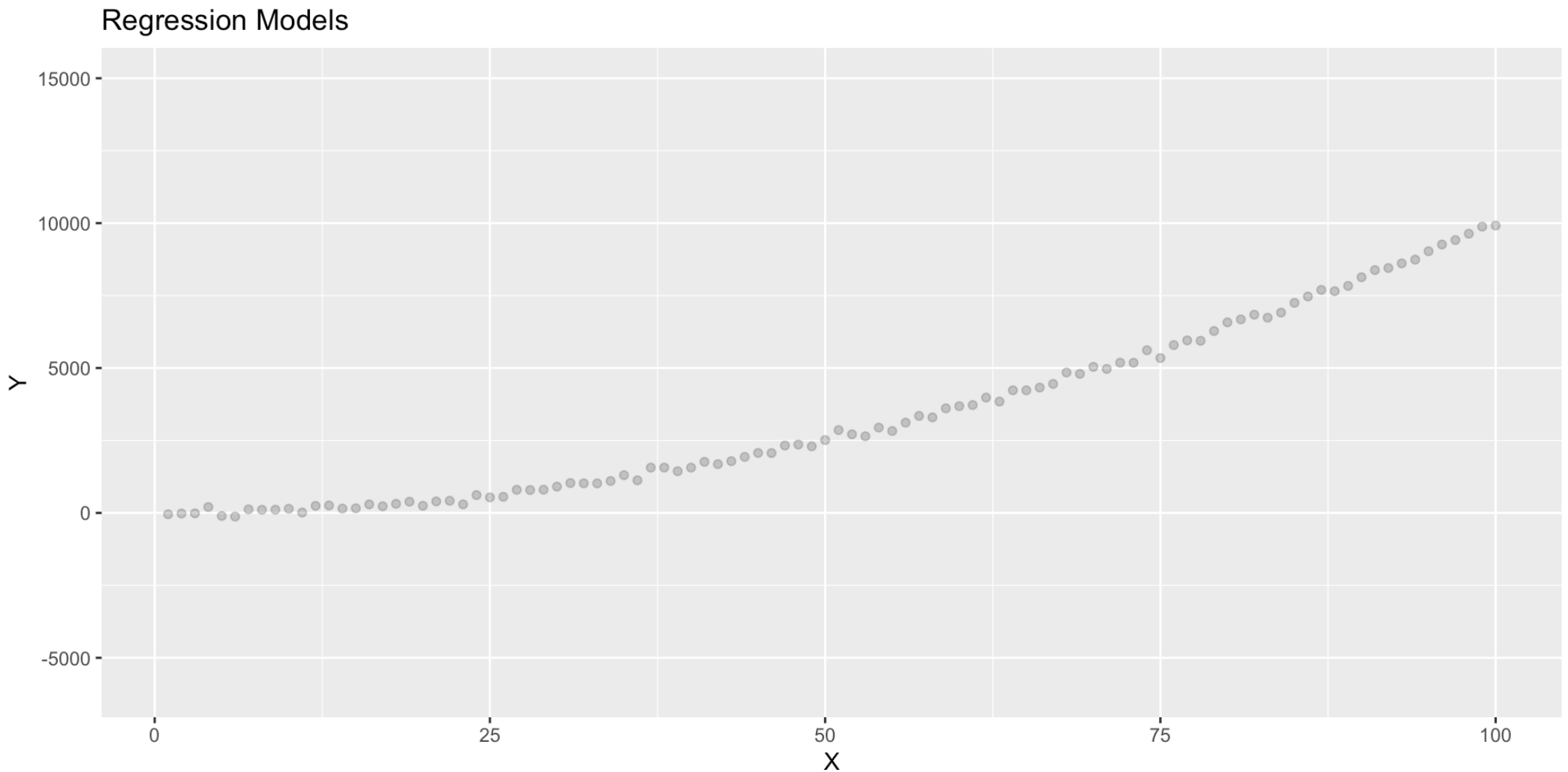
- Simple linear regression
- Multiple linear regression
- Non-linear regression, using functions such as polynomials, exponentials, logarithms, etc.

Asking ChatGPT for help with the next slide:

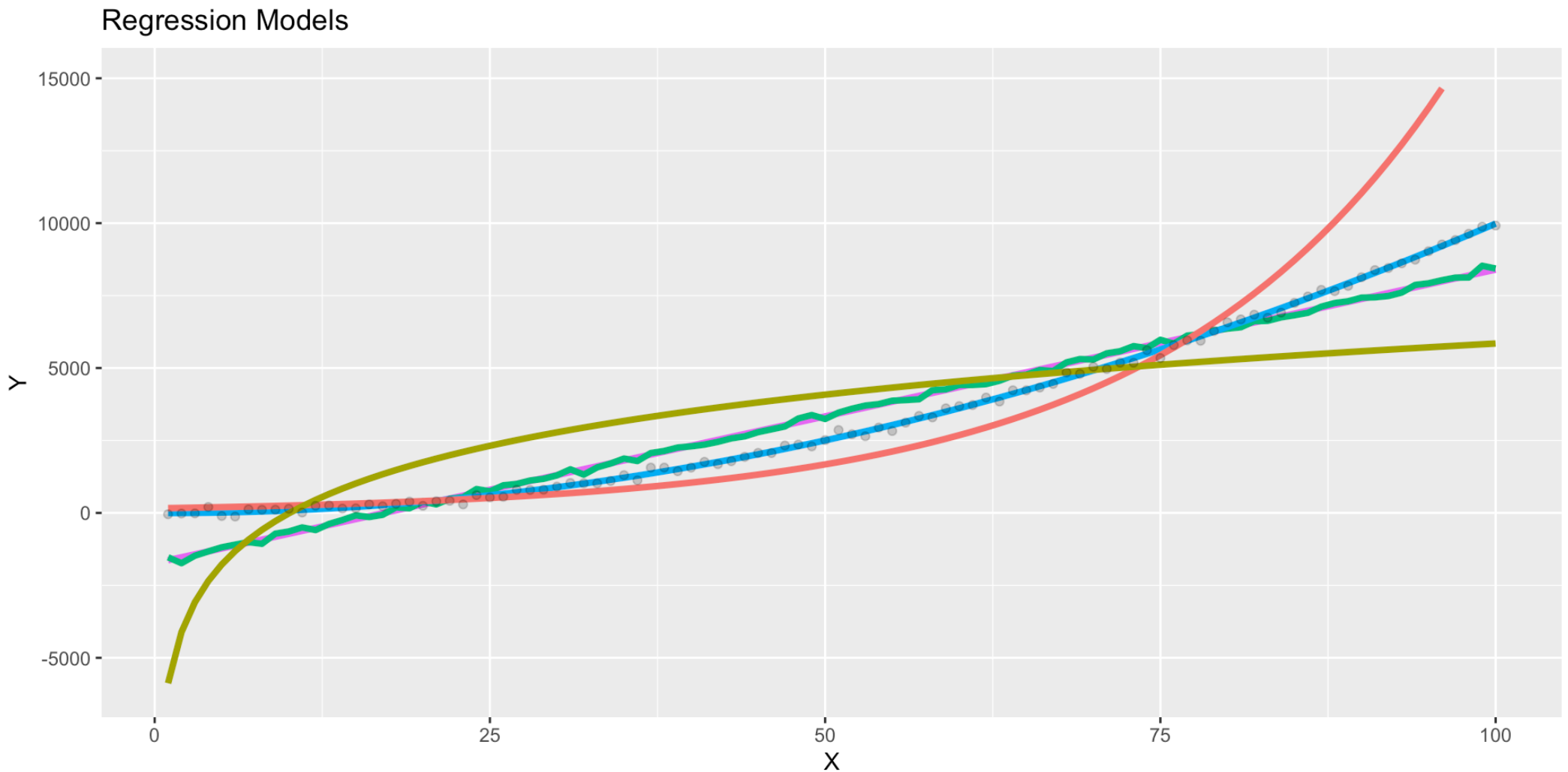
> Using R code, can you generate some data that is useful to demonstrate simple linear regression, multiple linear regression, polynomial, exponential and logarithmic regressions in ggplot2?

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

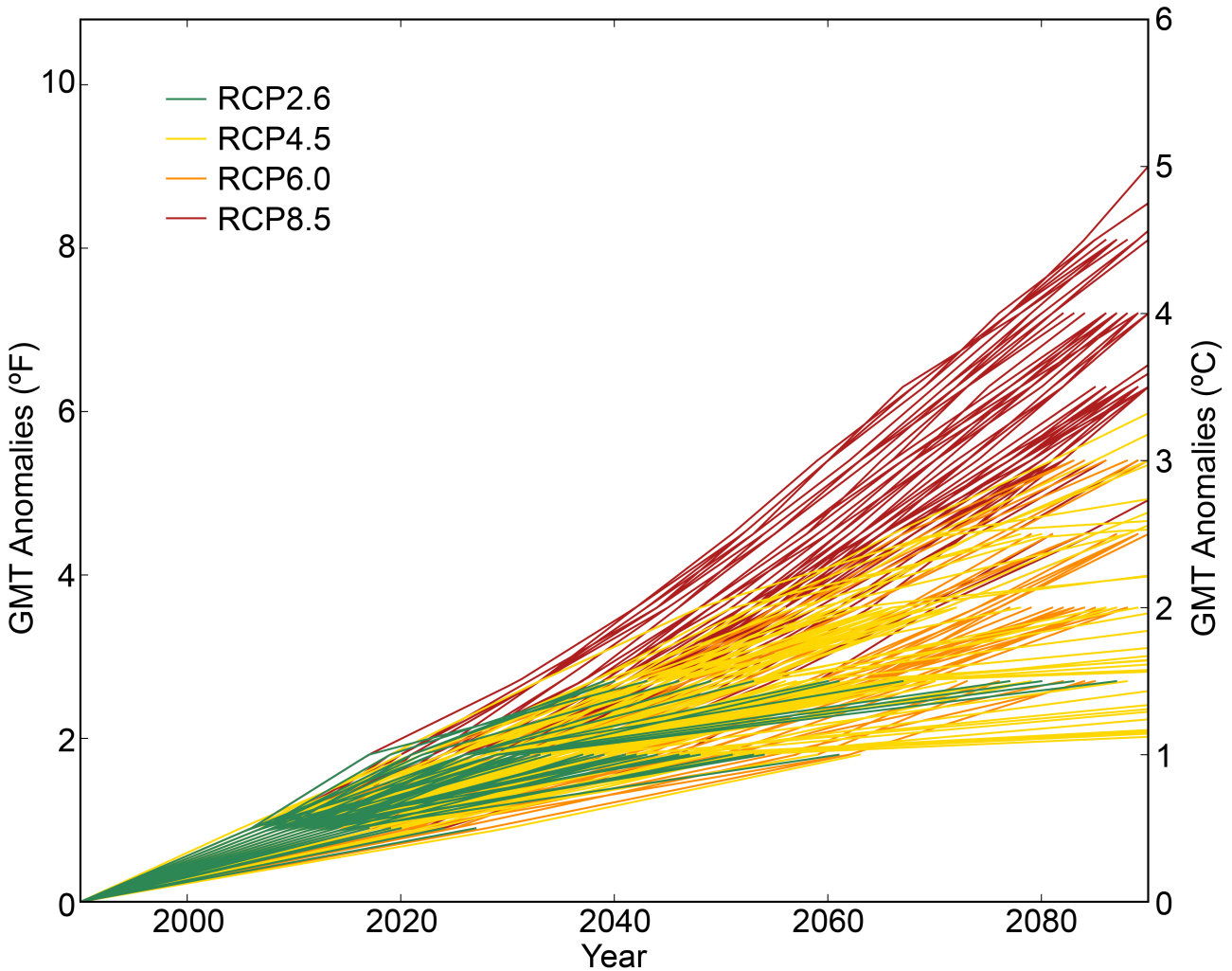
Visualising regression models



Visualising regression models

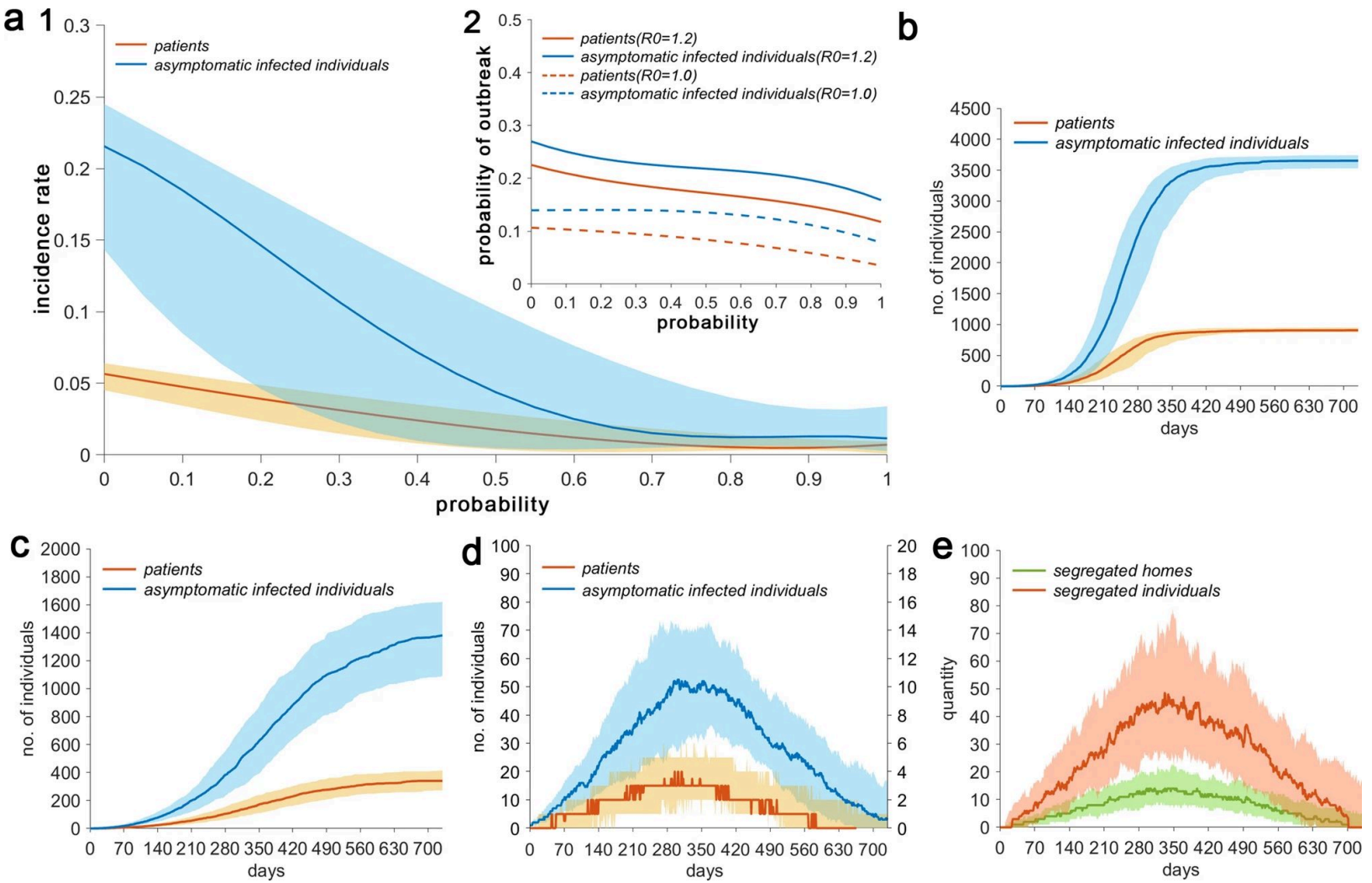


Example: climate change modelling



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

Example: COVID-19 transmission modelling

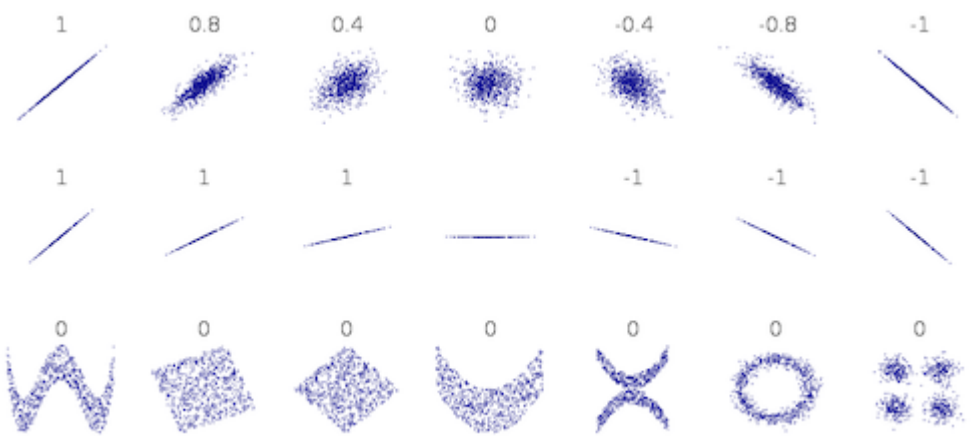


Source: https://www.nature.com/articles/s41598-021-84893-4/figures/1

Simple linear modelling

Defining a linear relationship

- Pearson correlation coefficient measures the linear correlation between two variables
- Does not distinguish different *patterns* of association, only the *strength* of the association



- Not quite usable for *predictive* modelling, or for *inference* about the relationship between variables

Simple linear regression modelling

We want to predict an outcome Y based on a predictor x for i number of observations:

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

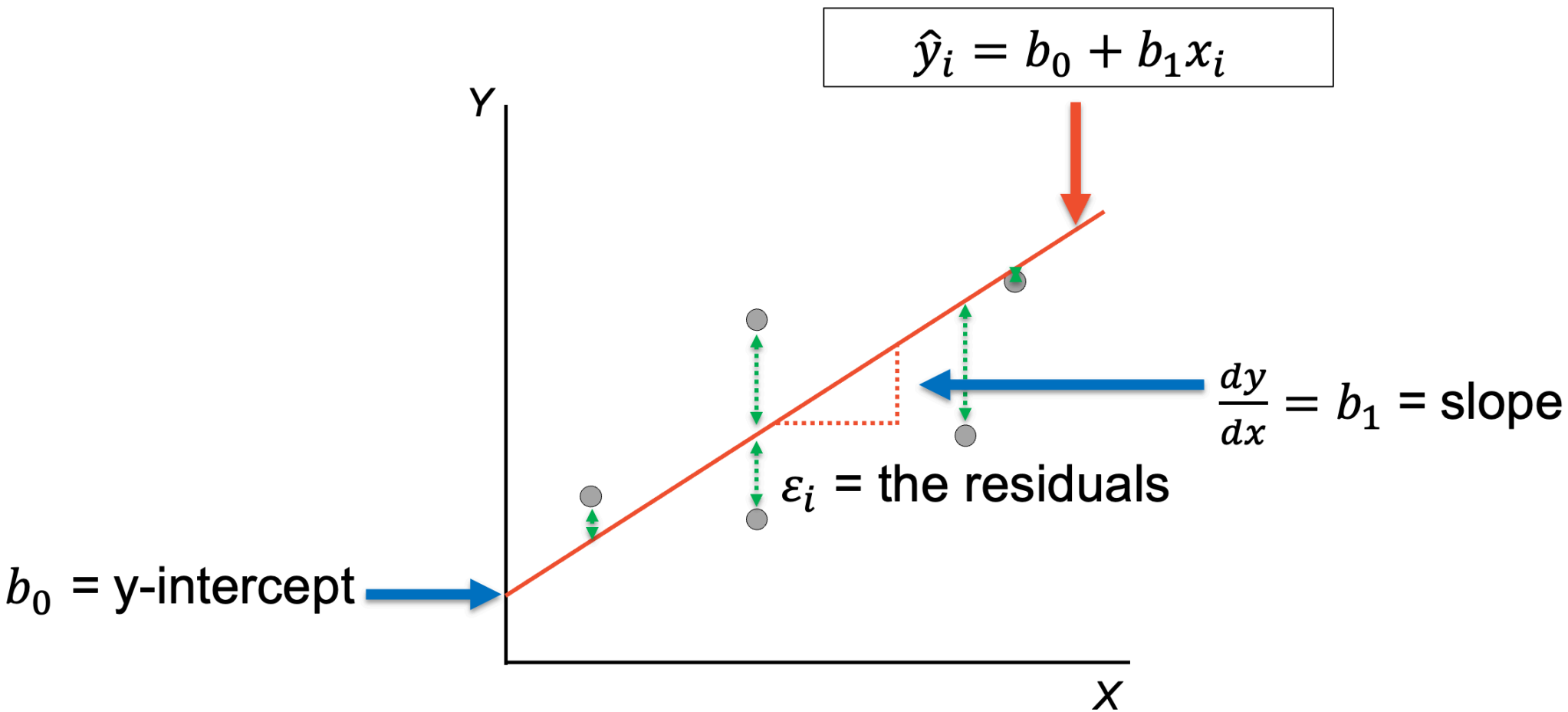
where

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

- Y_i, the *response*, is an observed value of the dependent variable.
- β_0, the *constant*, is the population 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*.

> **Note**
>
> We tend to associate ϵ_i with the **residual**, which is a positive or negative difference from the "predicted" response, rather than error itself which is a difference from the **true** response

In pictures…


$\hat{y}_i = b_0 + b_1 x_i$
Y
X
$\frac{dy}{dx} = b_1$ = slope
ε_i = the residuals
b_0 = y-intercept

Interpreting the relationship

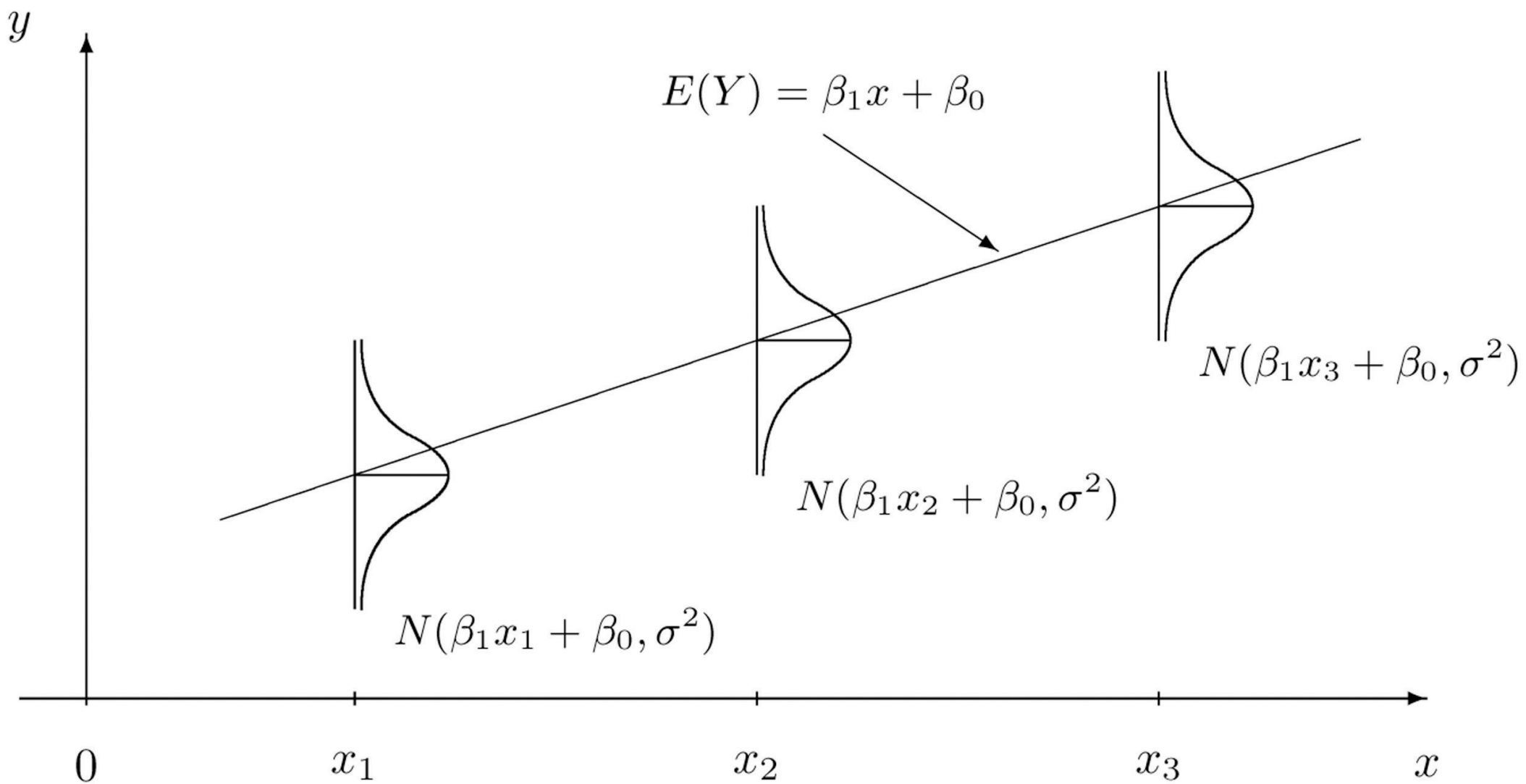
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Basically, a *deterministic* straight line equation $y = c + mx$, with added *random* variation that is normally distributed

- Response = Prediction + Error
- Response = Signal + Noise
- Response = Model + Unexplained
- Response = Deterministic + Random
- Response = Explainable + Everything else

- Y = f(x)
- Dependent variable = f(Independent variable)

The variation in the response



Model fitting

Two approaches; analytical and numerical:

Analytical: equation(s) used directly to find solution, e.g. estimate parameters that minimise residual sum of squares

Numerical: computer uses “random guesses” to find set of parameters to that minimises objective function, in this case residual sum of squares

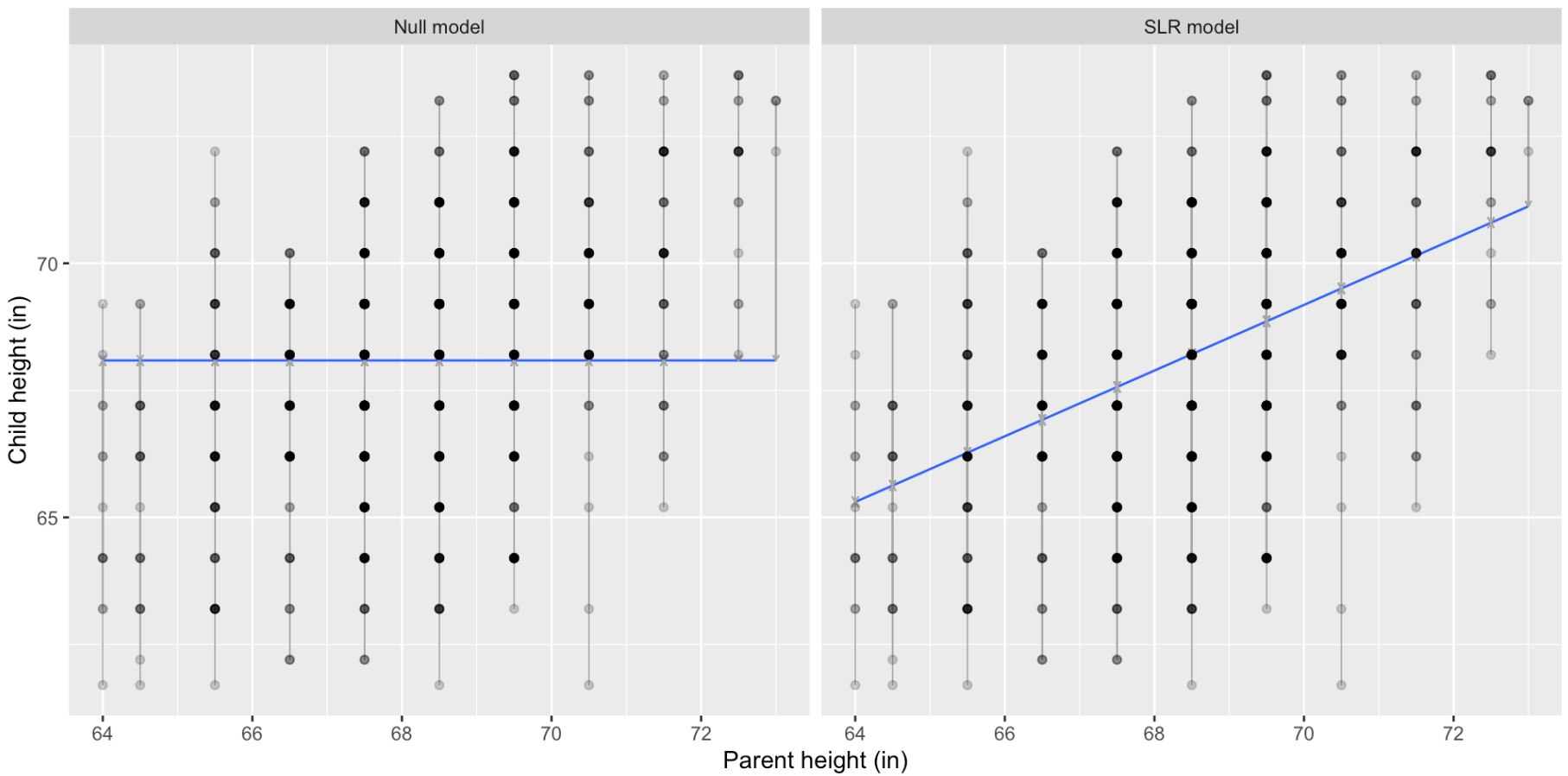
In this week’s practical we will use Solver in Excel to numerically fit linear models.

Inference: back to Galton's data

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

Hypothesis testing

- How does our null ($H_0 : \beta_1 = 0$) model compare to the linear ($H_0 : \beta_1 \neq 0$) model?
- Null model thinks the data can be summarised by the mean $\bar{y}$, and the linear model thinks the data can be summarised by the estimate $\hat{y}$.



Simple linear regression in one step

Done!

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

```
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 strength of the linear relationship between two variables.
- Correlation coefficient provides information on strength and direction of the linear relationship.
- Correlation $\neq$ causation.
- Regression models the relationship between a dependent variable and independent variable(s) - fits a line or function to the data.
    - ➠ Using the method of least squares, we can find the line that minimises the sum of the squared residuals.
- Both Excel and R can be used to fit regression models.

Thanks!