Modelling fundamentals – a linear model with a single, continuous X

BIOL2022 – **B**iology **E**xperimental **D**esign and **A**nalysis (**BEDA**)

Dr Januar Harianto

The University of Sydney

Semester 2, 2025

Learning objectives

You should:

- □ Understand the concept of a (general) linear model.
- □ Be able to come up with a linear model equation and interpret the coefficients and error term.
- □ Be able to interpret the results of a linear model.
- □ Understand that a linear model needs to be validated by checking assumptions.
- □ Be able to design a study with linear modelling in mind.

Introduction to linear modelling

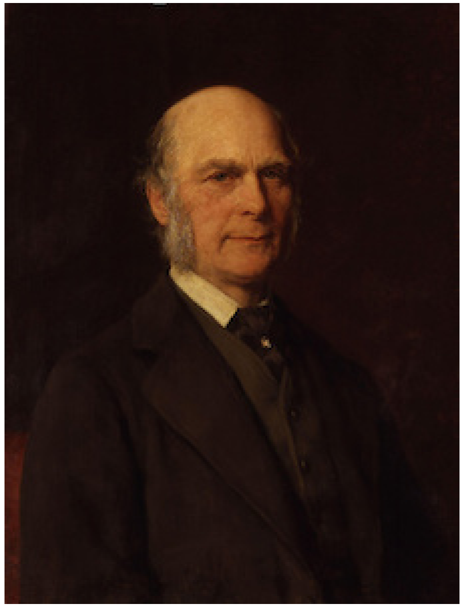
> “Cars with flames painted on the hood might get more speeding tickets. Are the flames making the car go fast? No. Certain things just go together. And when they do, they are correlated. **It is the darling of all human errors to assume, without proper testing, that one is the cause of the other.**”

— Barbara Kingsolver, Flight Behavior (2012)

Origin

Adrien-Marie Legendre (1752 – 1833), French Mathematician, first introduced the method of least squares in 1806. No known portrait of Legendre exists. Source: Wikipedia

Carl Friedrich Gauss (1777 – 1855), German mathematician, astronomer, and physicist. Published the method of least squares in 1809. Source: Wikipedia

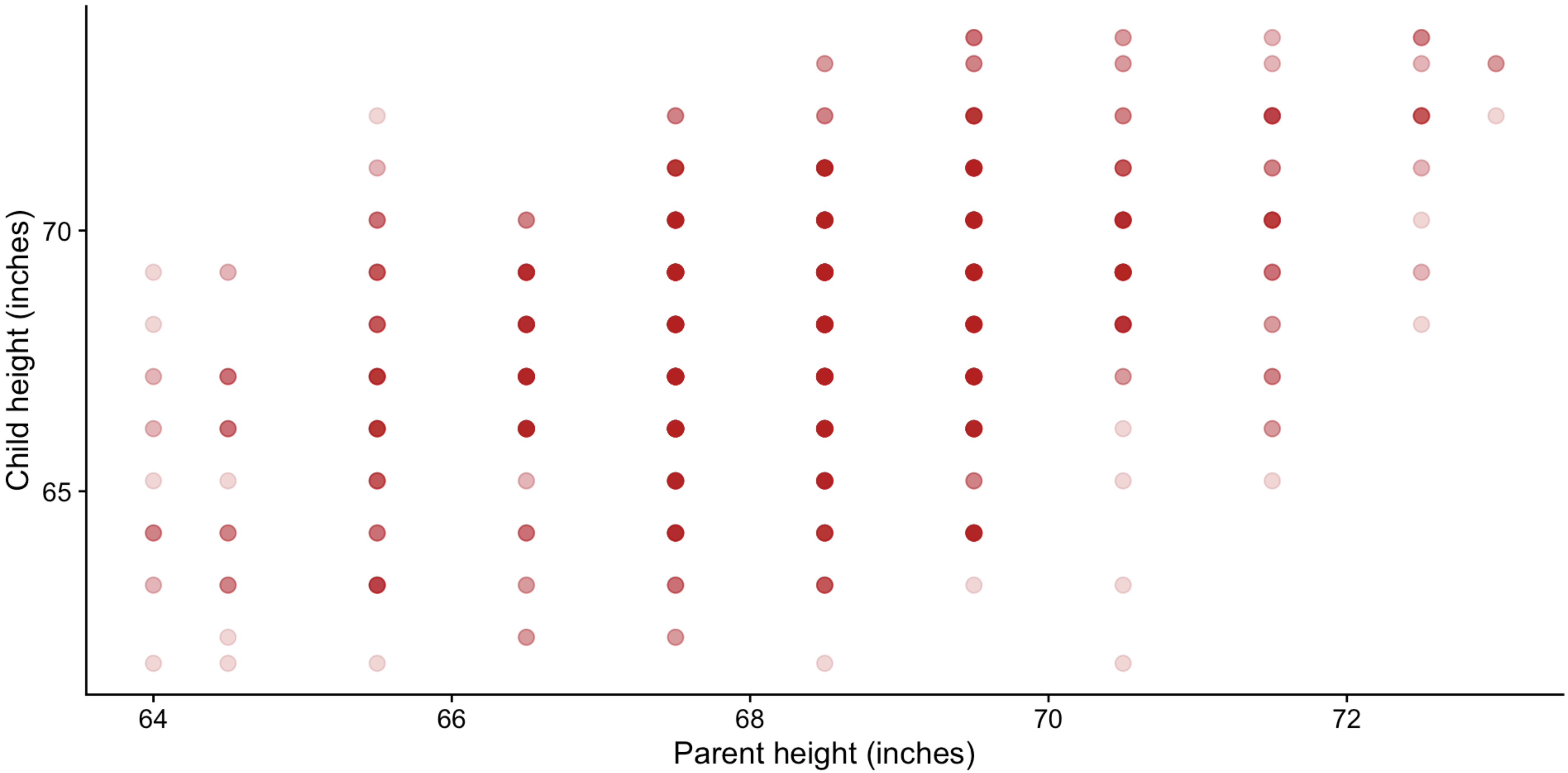


Francis Galton (1822 – 1911), cousin of Charles Darwin, inventor of the regression line and the correlation coefficient. Source: Wikipedia

Galton’s data

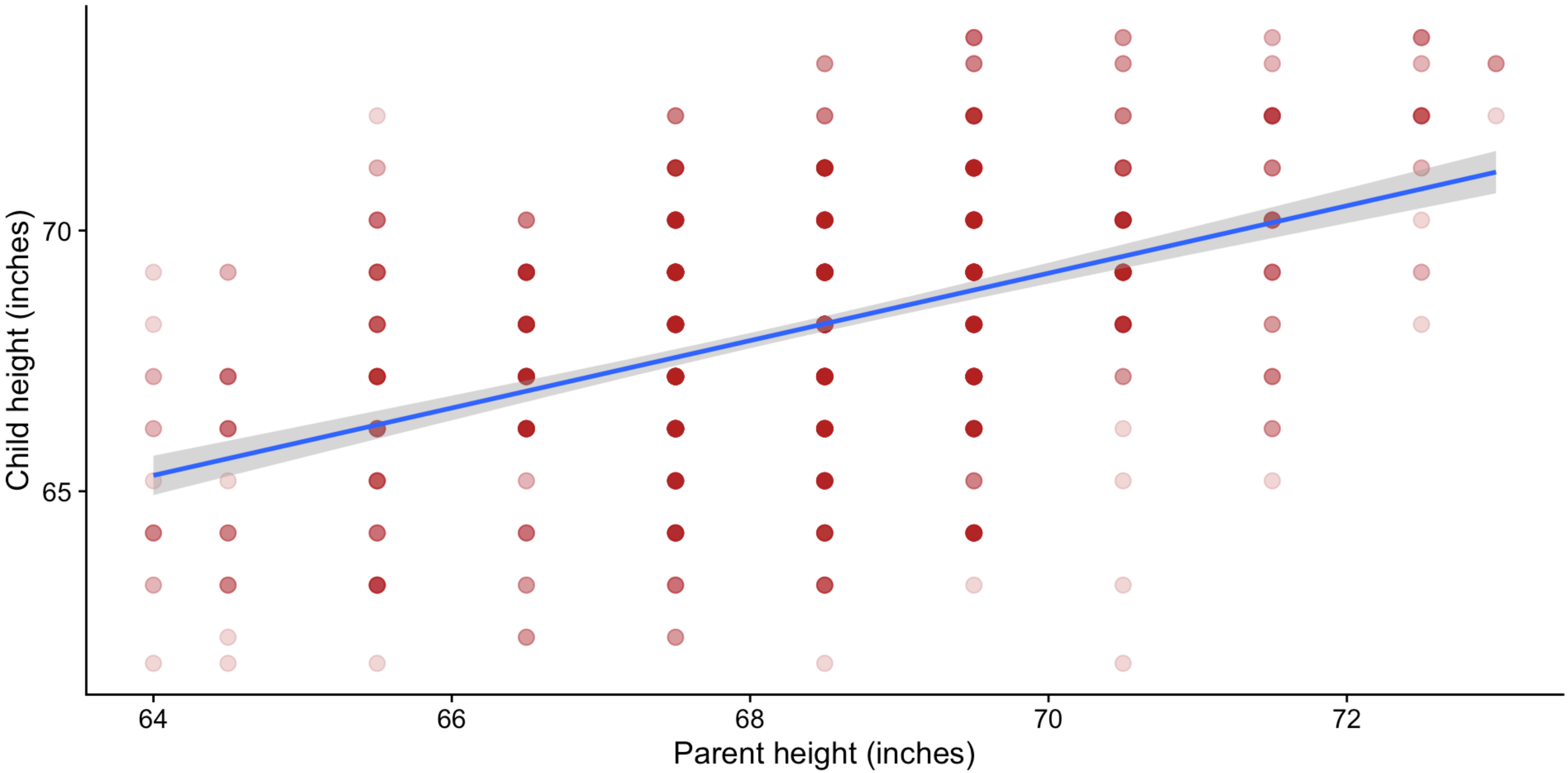
- 928 children of 205 pairs of parents.
- Height of parents and children measured in inches.
- Size classes were binned (hence data looks discrete).

Scatterplot



We want to explain this *noisy* relationship…

Regression line



We want to explain this *noisy* relationship… **using a deterministic model, i.e. a fixed equation.**

How does linear modelling work?

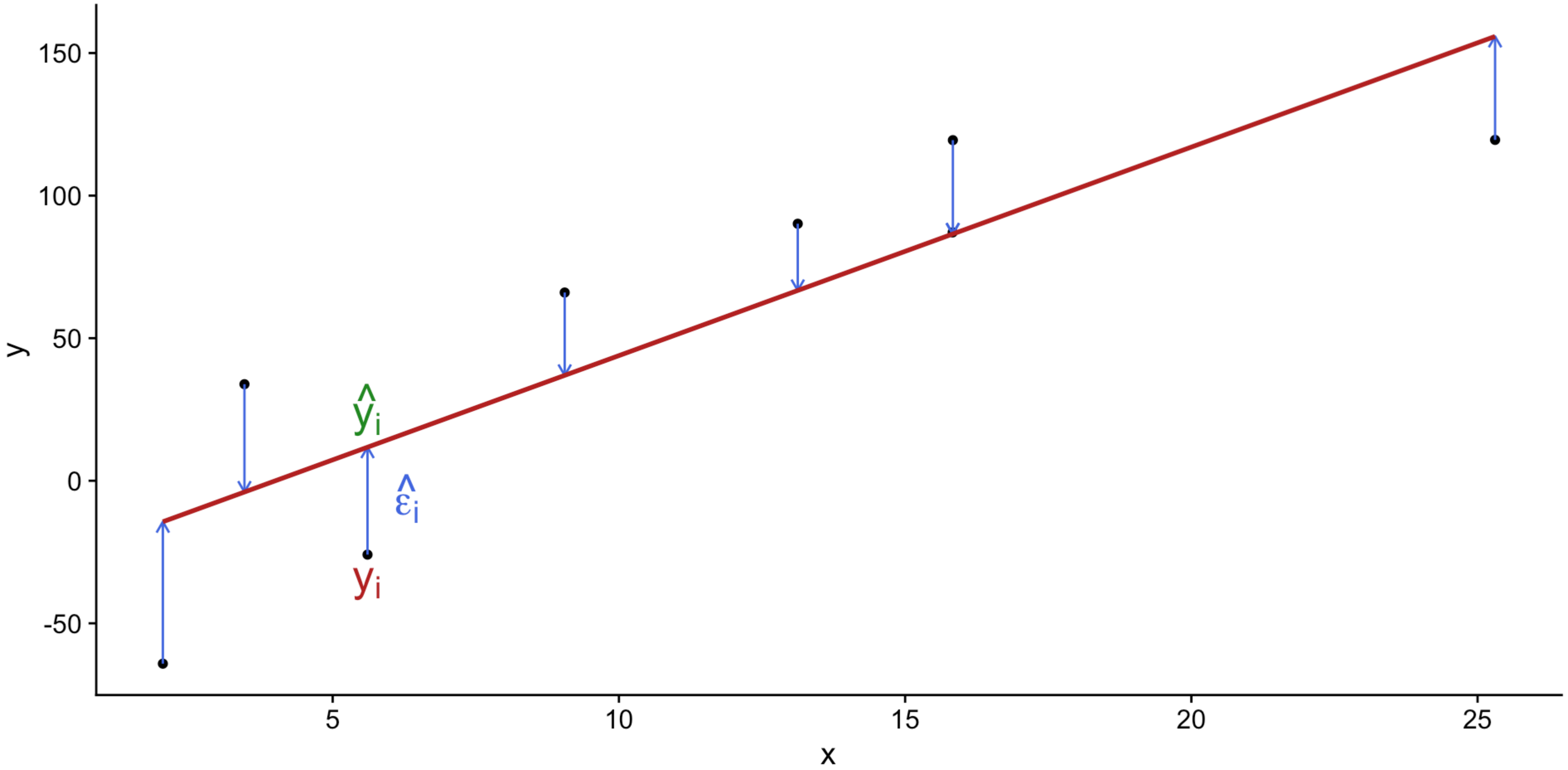
Least squares

> The method of least squares is the **automobile of modern statistical analysis**: despite its limitations, ocassional accidents and incidental pollution, it and its numerous variations, extensions, and related conveyances **carry the bulk of statistical analyses**, and are known and valued by nearly all.

– Stigler, 1981 (emphasis added)

Residuals, $\hat{\epsilon}$

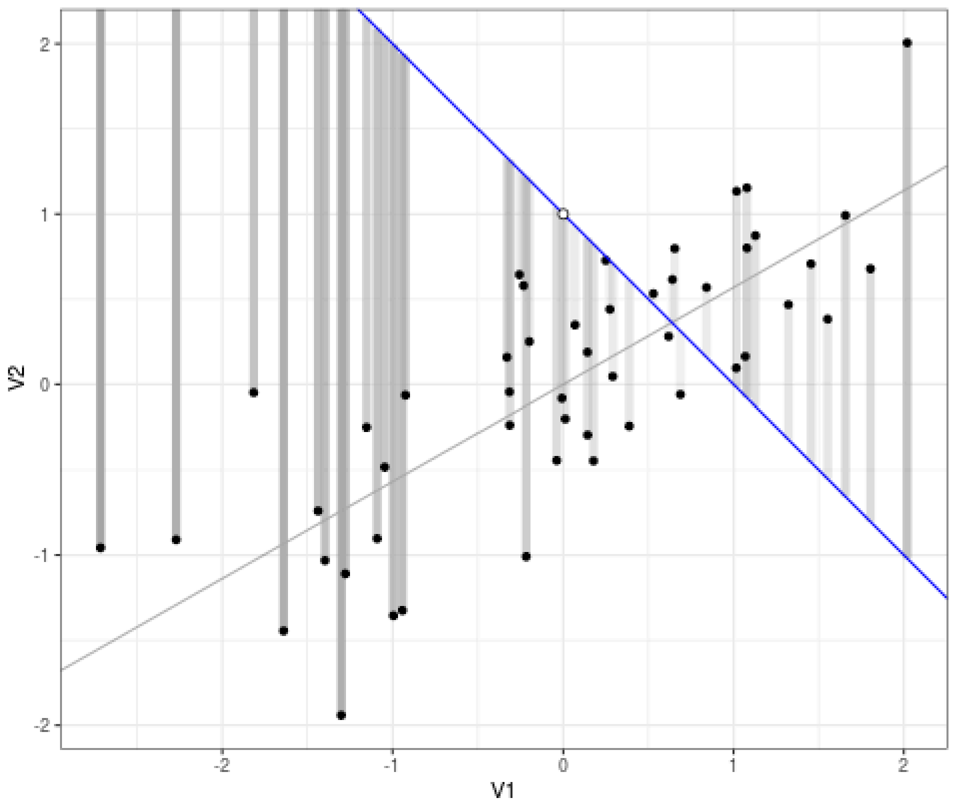
$$\hat{\epsilon_i} = y_i - \hat{y_i}$$



How do we fit a line?

Minimise the sum of the squared residuals:

$$argmin_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$



Source

1. Draw a line.
2. Calculate the residuals for each point.
3. Square the residuals, sum them up.
4. Repeat for all possible lines.
5. Choose the line with the smallest sum of squared residuals.
6. **Calculate the slope and intercept of that line.**

The linear model

Linear regression

A statistical method that fits a linear model equation to explain the relationship between two variables, x and y. The relationship is a **general linear model** that has the following form:

$$y = c + mx + \epsilon$$

where ϵ is the error term that accounts for the variability in y that is not explained by x.

In other words, we want to explain that changes in y can be estimated by the slope m and the intercept c.

Modelling

It would be useful to understand that in:

$$y = c + mx + \epsilon$$

we are fitting a *deterministic* straight line equation $c + mx$ to the data, with an *error* term ϵ that accounts for the variability in y that is not explained by x. Here are more ways to think about it:

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

What are we checking?

$$y = c + mx + \epsilon$$

- Estimate the coefficients (intercept c and slope m).
- **Once fitted, the line is fixed**; only the errors vary.
- Assessment of the model rests on assessing the errors – *because nothing else changes*.

It’s just about the errors… source: Tim & Eric Awesome Show, Great Job! Episode 3 of season 4.

Anatomy of a linear model

$$y = c + mx + \epsilon$$

can be expressed as

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

where y_i is the i-th observation of y, x_i is the i-th observation of x, and ϵ_i is the i-th observation of the error term.

The model is a **general linear model** with coefficients β_0 and β_1 that are estimated from the data. These coefficients represent the intercept and slope of the linear equation, respectively.

Fitting Galton’s data

Fitting a linear model will generally produce a result that looks like this:

▶ Code

Characteristic	Beta	SE	Statistic	95% CI	p-value
(Intercept)	24	2.81	8.52	18, 29	<0.001
parent	0.65	0.041	15.7	0.57, 0.73	<0.001
R²	0.210				
Adjusted R²	0.210				

Abbreviations: CI = Confidence Interval, SE = Standard Error

And this results in the following model:

$$\widehat{\text{child}} = 23.94 + 0.65(\text{parent})$$

- **Beta** coefficients are the **estimated** coefficients of the linear model.
- **SE** tells us how much the coefficient estimate might vary from the true value.
- **95% CI** gives us a range of values that we are 95% confident contains the true coefficient.
- **R-squared** tells us how much of the variability in the response variable is explained by the model. The **adjusted R-squared** adjusts for the number of predictors in the model (in this case, just one).

Interpretation of the model

$$\widehat{\text{child}} = 23.94 + 0.65(\text{parent})$$

$$\text{child} = \beta_0 + \beta_1(\text{parent}) + \epsilon$$

$$y = c + mx + \epsilon$$

- The **intercept** β_0 is the expected value of y when $x = 0$.
- The **slope** β_1 is the change in y for a one-unit change in x.
- The **error term** ϵ_i is the variability in y that is not explained by x.

Results

▶ Code

Characteristic	Beta	SE	Statistic	95% CI	p-value
(Intercept)	24	2.81	8.52	18, 29	<0.001
parent	0.65	0.041	15.7	0.57, 0.73	<0.001
R^2	0.210				
Adjusted R^2	0.210				

Abbreviations: CI = Confidence Interval, SE = Standard Error

The linear regression analysis revealed a significant positive relationship between parent height and child height ($\beta = 0.65$, $R^2 = 0.21$, $p < 0.001$). This indicates that for every one-inch increase in parent height, the child’s height is estimated to increase by 0.65 inches.

Example 2: penguins

▶ Code

Characteristic	Beta	95% CI	p-value
(Intercept)	-2,536	-4,442, -629	0.009
flipper_length_mm	33	23, 43	<0.001
R^2	0.219		
Adjusted R^2	0.214		

Abbreviation: CI = Confidence Interval

The results indicate a significant positive relationship between flipper length and body mass in Adelie penguins ($\beta = 33$, $R^2 = 0.22$, $p < 0.001$) indicating that for every 1 mm increase in flipper length, the body mass is expected to increase by 33 grams. Based on the R^2 value he model explains approximately 22% of the variance in body mass.

Example 3: iris

▶ Code

Characteristic	Beta	95% CI	p-value
(Intercept)	1.2	0.73, 1.6	<0.001
Sepal.Width	0.08	-0.05, 0.21	0.2
R²	0.032		
Adjusted R²	0.011		

Abbreviation: CI = Confidence Interval

The analysis found no significant relationship between sepal width and petal length in *Iris setosa* ($\beta = 0.08$, $p = 0.2$). The model explains only about 3% of the variance in petal length ($R^2 = 0.032$), and the confidence interval for the slope includes zero. This indicates that sepal width is not a meaningful predictor of petal length in this species.

Importantly:

- **Non-significant results are still valid results** – they tell us something important about the data
- **Report the actual p-value** (don't just say "$p > 0.05$")
- **Acknowledge the low explanatory power** (low R^2)
- **Avoid over-interpreting** – don't force biological explanations for non-significant relationships

Example 4: air quality

▶ Code

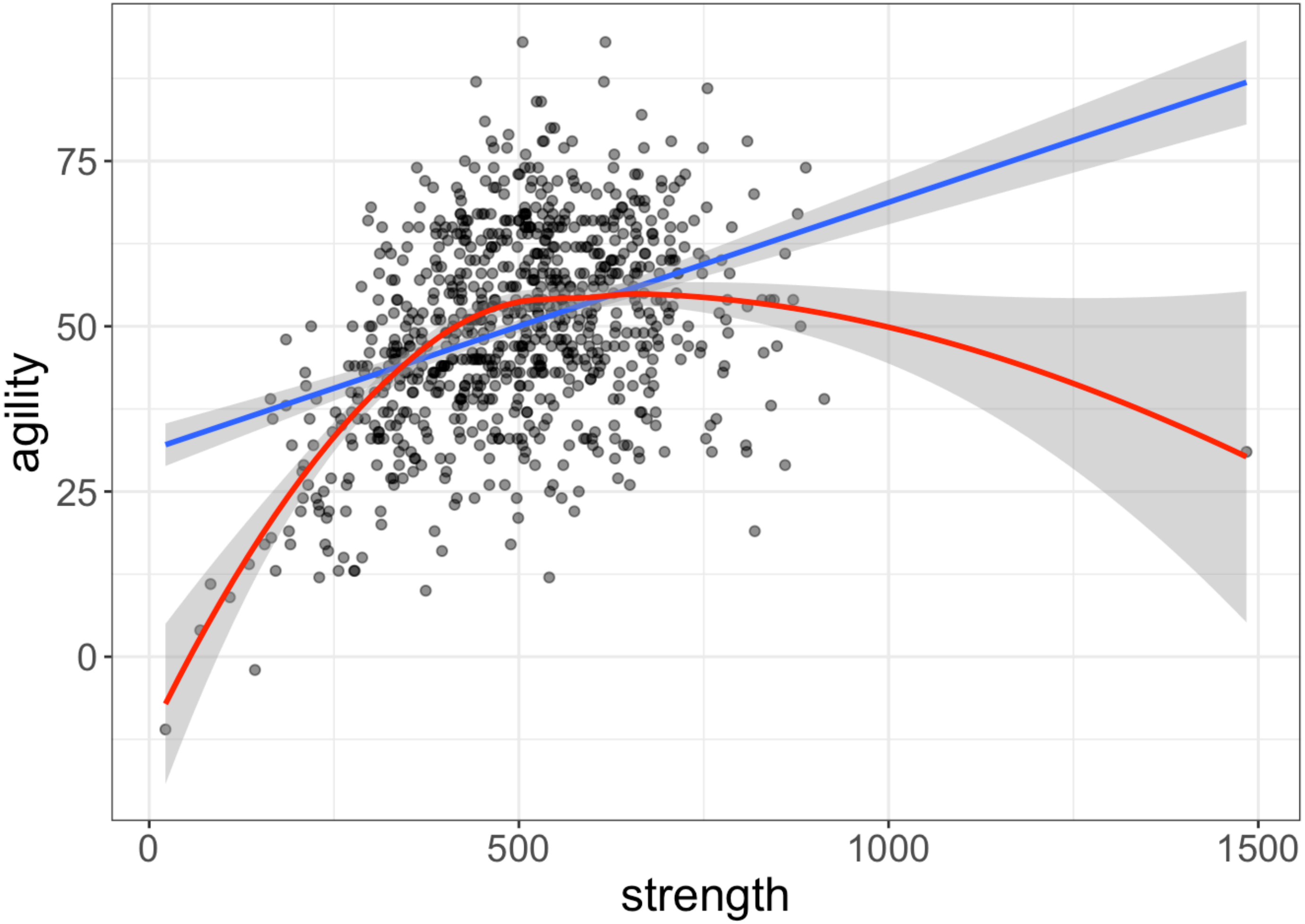
Characteristic	Beta	95% CI	p-value
(Intercept)	10	9.1, 12	<0.001
Solar.R	0.00	-0.01, 0.00	0.5
R^2	0.003		
Adjusted R^2	-0.004		

Abbreviation: CI = Confidence Interval

The analysis shows no significant relationship between solar radiation and wind speed (β = 0.00, 95% CI: -0.01 to 0.00, p = 0.5). The model does not explain much of the variance in wind speed (R^2 = 0.003). This result indicates that solar radiation and wind speed are influenced by different meteorological processes and are not expected to be directly related.

BUT WAIT! How do we know if the model is any good?

- Before we can interpret the results of a linear model, we need to check if the model is a good “fit”.
- Nothing will stop you from fitting a linear model to data, just as nothing can stop **me** from fitting a non-linear model to the same data.



Assessing the validity of the linear model

Assumptions of linear regression

Remember the acronym **LINE** for the key assumptions:

- **L**inearity: Errors (ϵ) are randomly scattered around the predicted values.
- **I**ndependence: Errors (ϵ) are independent of each other.
- **N**ormality: Errors (ϵ) follow a normal distribution.
- **E**qual Variance: Errors (ϵ) have constant variance (homoscedasticity).

These assumptions apply to the errors (ϵ), not directly to the response variable (y) or the relationship between x and y.

How assumptions help "validate" a model

- If the assumptions are met, then we can be confident that the model is a good representation of the data.
- If they are *not* met, the results are still presented, but our **interpretation** of the model is likely to be **flawed**.
- Importantly, we will never have "perfect" models, so assessing their validity is an ongoing process (and requires experience).

We will explore this next week (when we check **model assumptions**).

Thanks