**Notes:**

**Linear Regression Models:**

A graph of a graph of a line

AI-generated content may be incorrect.

**Calculating a line:**

* Draw a horizontal line through the data
* Measure the distance from each point to the line (measuring the residual), square each distance then sum them
* Rotate the line, measure residuals, square then sum
* Rotate again, measure residuals, square then sum multiple times, then plot a graph of ∑ squared residuals against different rotations
* Locate the rotation that has the least sum of squares, then use that rotation to fit the line, this is called least squares method

**Calculating R²:**

* Calculate the average for one variable, then sum the squared residuals from the mean to the point (known as SS(mean))
* SS(mean) = (data-mean)² and Var(mean) = SS(mean)\*n¯¹
* Sum the squared residuals from least-squares fit (known as SS(fit))
* SS(fit) = (data-line)² and Var(fit) = ((data-line)²)\*n¯¹ or Var(fit) = SS(fit)\*n¯¹
* R² tells us how much of a variation in data can be explained by its correlation to another variable

A black text on a white background

AI-generated content may be incorrect.

**R² in multivariable models:**

* A 3D graph is needed to plot the data
* Fit a least-squares plane instead of a line
* Measure residuals and sum to find R²
* If one of the parameters is useless in terms of predicting an outcome, aka it has no direct correlation, then we should ignore it by making it equal to 0
* The more parameters added, the more opportunities for random events, so SS(fit) gets smaller, leading to a better R²
* This means people report an adjusted R² value which scales R² by the number of parameters

**Calculating a p-value:**

* R² does not work for only 2 measurements
* The p-value comes from something called F
* F is the variation in a variable explained by another variable divided by the variation in a variable not explained by another variable

A red oval with black text

AI-generated content may be incorrect.

* Pfit is the number of parameters in the fit line and Pmean is the number of parameters in the mean line
* Get a set of data, calculate F, plot the value on a histogram then repeat for multiple sets of data, until you end up with an F distribution curve