# R

There is a formula for the correlation coefficient and most things have the capability of calculating this. It explains how correlated values are, R^2 is better tho.

# R^2

R is correlation value, if its 1 or -1 then the values are strongly related if its close to zero there is low correlation.

What is R^2^ tho? it is more linear correlation than R, R2 of 0.7 is 1.4 times better than R2 of 0.5, this is not true for R

the variation of the data is sum of the squared y value minus mean value for each data point

Chart, line chart

Description automatically generated

We could then order the mice by size and then have a line that tells us about the mice weight based on size.

Chart, line chart

Description automatically generated

Var(line) = Sum((data - line) ^ 2) / n // note the line could also just be the data mean value variance is the average sum of squares per data point from the goal

R2 = variation of y explained by x / variation of y without taking into account x

R^2 is a way to look at the variation in the line with the variation of the dataset removed from consideration.

R^2 is the variation of the line as a percent of a variation of the mean. This means that we can figure out how much less variation the line has rather than the mean value. This means we can extract variation from relationship from general noise variation.

R^2 is the percent less variation around the fitted line versus the variation in the mean. Tells us how important a variation is.

R^2 is just the squared value of R

We want a p value tho to tell us if our R2 value is statistically significant. The p value for R2 comes from F. F is the variation in y explained by x over the variation in y not explained by x.

F =

The p values here are the degrees of freedom. P fit would be the number of parameters in the line, in 2 variable linear regression this would be 2 since we have the slope and the y-intercept. P mean would be the number of parameters in the mean line, this is usually just the y intercept. This means p fit – p mean would be 1 in this case.

This means the numerator for the F calculation is the variance in y explained by the extra parameter, if we had two extra parameters then we would want this number to be lower because one parameter in itself is not fully responsible for the reduction in variance. The numerator is the variance in y explained by all the parameters in X.

The denominator is the variation in y not explained by the fit.

If we took random data aka just noise and calculated the F value of it ( potentially millions of times ) and put it on a histogram. Then we calculate F value of our dataset. The P value is the number of more extreme values divided by all values of F. We could just use an F distribution line that is precalculated to calculate our p value. The number of degrees of freedom determine the shape of our graph.

A picture containing graphical user interface

Description automatically generated

P is larger when there are more samples in the dataset versus parameters used.