Two Variable Statistics – A Simple Summary

X and Y are two variables

$x_1\ x_2\ x_3\ x_4 \ldots x_i \ldots x_n$

$y_1\ y_2\ y_3\ y_4 \ldots y_i \ldots y_n$

$\sum$ is the summation sign

$\sum_{i=1}^{n} = x_1 + x_2 + x_3 \ldots + x_i \ldots + x_n$

or more simply $\sum x$

$\bar{x}$ (x bar) is the sample mean of x

$\bar{x} = \{ \sum x \} / n$ and $\bar{y} = \{ \sum y \} / n$

S_{xx} is the variance of x using $\bar{x}$ as the best estimate of the mean of μ.

$S_{xx} = \{ \sum (x - \bar{x})^2 \} / n$

$= \sum (\bar{x})^2 / n \ - 2 \bar{x} \sum x / n + n \bar{x} / n$

$= \{ \sum (x)^2 \} / n \ - \bar{x}^2$

We say the mean of the squares minus the square of the means. This form is simpler to calculate and more stable.

The standard deviation

$S_x = \sqrt{(S_{xx})}$

S_{xy} is the covariance of X and Y. It is the measure of how two variables change together.

The sign shows the tendency in the linear relationship between the two variables.

$S_{xy} \quad = \{ \sum (x - \bar{x}) (y - \bar{y}) \} / n$

$\quad = \sum xy / n - \bar{x}\,\bar{y} - \bar{y}\,\bar{x} + \bar{x}\,\bar{y}$

$\quad = \sum xy / n - \bar{x}\,\bar{y}$

Compare this with S_{xx}

The magnitude of the covariance is difficult to interpret directly so it is normalised to produce the correlation coefficient r a dimensionless measure.

$r = S_{xy} / (S_x S_y)$

Using the two simplifications already shown and cancelling the n's we get

$r = (\sum xy - n\bar{x}\bar{y}) / \sqrt{(\sum x^2 - n\,\bar{x}^2)}\sqrt{\sum (y^2 - n\bar{y}^2)}$

The correlation coefficient is a measure of the degree of linearity between the two variables X and Y

Now if we draw a line of best fit between the plots the equation is given by $y = mx + c$

Now the gradient is given by

$m = S_{xy} / S_{xx}$

Strictly this is y on x and the line passes through the point $(\bar{x} , \bar{y})$

So we have $(y - \bar{y}) = m (x - \bar{x})$

Rearranging we determine

$c = \bar{y} - m\,\bar{x}$

or $\quad c = \bar{y} - (S_{xy} / S_{xx})\,\bar{x}$

This is the y-intercept of the line of best fit.

So the equation for linear regression is

$(y - \bar{y}) = \{ S_{xy} / S_{xx} \} (x - \bar{x})$

The coefficient of determination

r^2

is a measure of the strength of association between the two variables X and Y.
For linear regression r^2 it is the proportion explained by the model.

0	$\le \mathrm{r}^2 < 0.25$	very weak
0.25	$\le \mathrm{r}^2 < 0.5$	weak
0.5	$\le \mathrm{r}^2 < 0.75$	moderate
0.75	$\le \mathrm{r}^2 < 0.9$	strong
0.9	$\le \mathrm{r}^2 < 1.0$	very strong

$\mathrm{r}^2 = 1$ is perfect correlation.
For linear correlation we have

$r^2 = \mathrm{r}^2$

but be careful

r^2 is the coefficient of determination and is a thing all in its own right

whereas

r^2 is the correlation coefficient squared.

Not a lot of people know this!

∞ rg