Two Variable Statistics – A Simple Summary

X and Y are two variables $x_1\ x_2\ x_3\ x_4 \ldots x_i \ldots x_n$ and $y_1\ y_2\ y_3\ y_4 \ldots y_i \ldots y_n$

$\sum$ is the summation sign $_{i=1}\sum^n = x_1 + x_2 + x_3 \ldots + x_i \ldots + x_n$ or more simply $\sum x$

$\bar{x}$ (x bar) is the sample mean of x so $\bar{x} = \{ \sum x \} / n$ 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.

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

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

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 equation of the line of best fit 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 $c = \bar{y} - (S_{xy} / S_{xx})\,\bar{x}$

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

So the equation for linear regression is $(y - \bar{y}) = \{ S_{xy} / S_{xx} \} (x - \bar{x})$

0	$\leq r^2 < 0.25$	$\equiv$	0 $\leq \lvert r \rvert <$	0.5	very weak
0.25	$\leq r^2 < 0.5$	$\equiv$	0.5 $\leq \lvert r \rvert <$	0.7	weak
0.5	$\leq r^2 < 0.75$	$\equiv$	0.7 $\leq \lvert r \rvert <$	0.87	moderate
0.75	$\leq r^2 < 0.9$	$\equiv$	0.87 $\leq \lvert r \rvert <$	0.95	strong
0.9	$\leq r^2 < 1.0$	$\equiv$	0.95 $\leq \lvert r \rvert <$	1.0	very strong

The Correlation Coefficient and the Coefficient of Determination

For linear regression, squaring the **correlation coefficient** - ie r^2 - gives a measure of the degree of linearity between the two variables X and Y.
It is the proportion explained by the model. $r^2 = 1$ is perfect correlation

However the **coefficient of determination** is r^2 is a measure of the strength of association between the two variables X and Y and may not necessarily be linear. It is the proportion of the variation in the dependent variable that is predictable from the independent variable(s).
(r^2) is not any measure squared – it is a measure in its own right.
This summary sheet does not detail exactly how this is calculated but in general it uses the residual sum of squares. Normally values range from 0 to 1 but circumstances can arise when it is negative.

While for linear correlation we have $\mathit{r}^2 = r^2$

do not confuse (r^2) with (r)^2 *rg*

Two Variable Statistics – A Simple Summary

X and Y are two variables $x_1\ x_2\ x_3\ x_4 \ldots x_i \ldots x_n$ and $y_1\ y_2\ y_3\ y_4 \ldots y_i \ldots y_n$

$\sum$ 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 so $\bar{x} = \{ \sum x \} / n$ 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.

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

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

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 equation of the line of best fit 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 $c = \bar{y} - (S_{xy} / S_{xx}) \bar{x}$

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

So the equation for linear regression is $(y - \bar{y}) = \{ S_{xy} / S_{xx} \} (x - \bar{x})$

0	$\leq r^2 < 0.25$	$\equiv$	0	$\leq \vert r \vert <$	0.5	very weak
0.25	$\leq r^2 < 0.5$	$\equiv$	0.5	$\leq \vert r \vert <$	0.7	weak
0.5	$\leq r^2 < 0.75$	$\equiv$	0.7	$\leq \vert r \vert <$	0.87	moderate
0.75	$\leq r^2 < 0.9$	$\equiv$	0.87	$\leq \vert r \vert <$	0.95	strong
0.9	$\leq r^2 < 1.0$	$\equiv$	0.95	$\leq \vert r \vert <$	1.0	very strong

The Correlation Coefficient and the Coefficient of Determination

For linear regression, squaring the **correlation coefficient** - ie r^2 - gives a measure of the degree of linearity between the two variables X and Y.
It is the proportion explained by the model. $r^2 = 1$ is perfect correlation

However the **coefficient of determination** is r^2 is a measure of the strength of association between the two variables X and Y and may not necessarily be linear. It is the proportion of the variation in the dependent variable that is predictable from the independent variable(s).
(r^2) is not any measure squared – it is a measure in its own right.
This summary sheet does not detail exactly how this is calculated but in general it uses the residual sum of squares. Normally values range from 0 to 1 but circumstances can arise when it is negative.

While for linear correlation we have $\mathit{r}^2 = r^2$

do not confuse (r^2) with $(r)^2$ xo *rg*