Correlation and Regression

Mirza Hasanuzzaman
Professor, Department of Agronomy
Sher-e-Bangla Agricultural University

Correlation

When there is a relationship between quantitative measures between two sets of phenomena, the appropriate statistical tool for discovering and measuring the relationship and expressing it on a precise way is known as correlation.

In an experiment, if the changes of one variable affect the changes of the other variable, then the variables are said to be correlated. e.g. yield and tiller number, length of panicles and grains/panicles, etc. If the measurements of the variables are in the same direction, then the variables are said to be directly correlated or positively correlated, e.g. test grain weight and grain size, leaf area and leaf size. If the movements of variables are in opposite direction, then the variables are said to be negatively correlated, e.g. yield and pest incidence, germination percentage and ageing of seeds. To know the direction of movement of the variables, scattered diagram method is used.

When two variables are involved, the correlation is termed as 'simple correlation'. If more than two variables are involved, the correlation is known as 'multiple correlation'.

To measure the degree of relationship between the correlated variables x and y the following formula is used:

$$r_{xy} = \frac{Cov\,(xy)}{\sqrt{Var(x).\,Var(y)}}$$

$$= \frac{SP(xy)}{SS(x).SS(y)}$$

$$= \frac{\frac{1}{n}\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\left\{\frac{1}{n}\sum\left(x-\bar{x}\right)^2\right\}\left\{\frac{1}{n}\sum\left(y-\bar{y}\right)^2\right\}}}$$

$$= \frac{\sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}}{\sqrt{\left\{\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right\}\left\{\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right\}}}$$

Example:

Data of number of siliquae per plant and seed yield of mustard in a field experiment are as follows:

Here, number of siliquae is independent variable is denoted by '***x***' and seed yield is dependent variable is denoted as '***y***'.

We can calculate the correlation coefficient manually by following way-

No. of siliquae plant^{-1} (x)	**Seed yield (t ha^{-1}) (y)**	x^2	y^2	xy
30	0.5	900	0.25	15
40	0.7	1600	0.49	28
50	0.9	2500	0.81	45
60	1.1	3600	1.21	66
70	1.3	4900	1.69	91
80	1.5	6400	2.25	120
90	1.7	8100	2.89	153
100	1.9	10000	3.61	190
110	2.1	12100	4.41	231
120	2.5	14400	6.25	300
$\sum x$ = 750	**$\sum y$ = 14.2**	**$\sum x^2$ = 64500**	**$\sum y^2$ = 23.86**	**$\sum xy$ = 1239**

Calculation:

$$r = \frac{\sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}}{\sqrt{\left\{\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right\}\left\{\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right\}}}$$

$$= \frac{1239 - \frac{750 \times 14.2}{10}}{\sqrt{\left\{64500 - \frac{750^2}{10}\right\}\left\{23.86 - \frac{14.2^2}{10}\right\}}}$$

$$= \frac{1239 - 1065}{\sqrt{\{64500 - 56250\}\{23.86 - 20.164\}}}$$

$$= \frac{174}{\sqrt{8250 \times 3.696}}$$

$$= \frac{174}{\sqrt{30492}}$$

$$= \frac{174}{174.61}$$

= 0.996

Test of significance of Correlation coefficient

Let us suppose that r be the correlation coefficient from a sample of size n from a bivariate normal population. We are to test null hypothesis that the population correlation coefficient is zero, i.e. H_0: ρ =0

The required test statistic is

t = $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, which is distributes as t with (n-2) d. f

If the calculated value of t with (n-2) d. f is seen to be smaller than the tabulated value of t with same d. f. at 5% levels of significance then the calculated value of | t | is insignificant and the hypothesis may be accepted. On the other hand, if the calculated value of t is greater than the tabulated value of t than the hypothesis is rejected and the calculated value of | t | is significant.

Regression

Regression can be defined as a method that estimates the value of one variable when that of other variable is known, provided the variables are correlated.

Regression is a simple and more useful approach to the study of simultaneous variation of two (or more) characters. In a field experimentation, we may examine the relationship between levels of nitrogen and crop yield. The levels of N may be fixed arbitrarily and may not have any distribution. Under such condition, the method of regression is more appropriate than correlation. Of the two variables, one is known as independent variable, denoted by $\boldsymbol{x}$. The other variable is called dependent variable, denoted by $\boldsymbol{y}$. The underlying relation between $\boldsymbol{x}$ and $\boldsymbol{y}$ in a bivariate population can be expressed as a function. Such functional relationship between two variables is termed as regression.

(i) Regression line of y on $\boldsymbol{x}$ is $\boldsymbol{y = a + bx}$

(ii) Regression line of x on **y** is $\boldsymbol{x = a + by}$

Regression coefficient, $\boldsymbol{b_{yx}}$ $= \frac{SP(xy)}{SS(x)}$

$$= \frac{\frac{1}{n}\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\left\{\frac{1}{n}\sum\left(x-\bar{x}\right)^2\right\}}$$

Regression coefficient, b_{xy} $= \frac{SP(xy)}{SS(y)}$

$$= \frac{\frac{1}{n}\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\left\{\frac{1}{n}\sum\left(y-\bar{y}\right)^2\right\}}$$

In case of two set of data, r = $\sqrt{b_{xy} \times b_{yx}}$

Necessity of regression

The regression is used for the following purpose:

(i) To predict the average relationship between the dependent variable and independent variables.

(ii) To determine the contribution of each independent variable on the dependent variables.

(iii) To estimate the value of dependent variable for a given value of independent variables.

An agriculturist may be interested to study the dependence of yield of mustard on irrigation, plant spacing, fertilizers etc. Such an analysis may enable the estimate the average yield of mustard on the basis of information about the independent variables.

Coefficient of determination

Coefficient of determination is the ratio of explained variation and Total variation.

R^2 $= \frac{Explained\ Variation}{Total\ Variation}$

R^2 $= = \frac{\sum(y_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}$

or, Proportion of total variation explained by regression.

R-squared value

A number from 0 to 1 that reveals how closely the estimated values for the trendline correspond to your actual data. A trendline is most reliable when its R-squared value is at or near 1. Also known as the coefficient of determination.

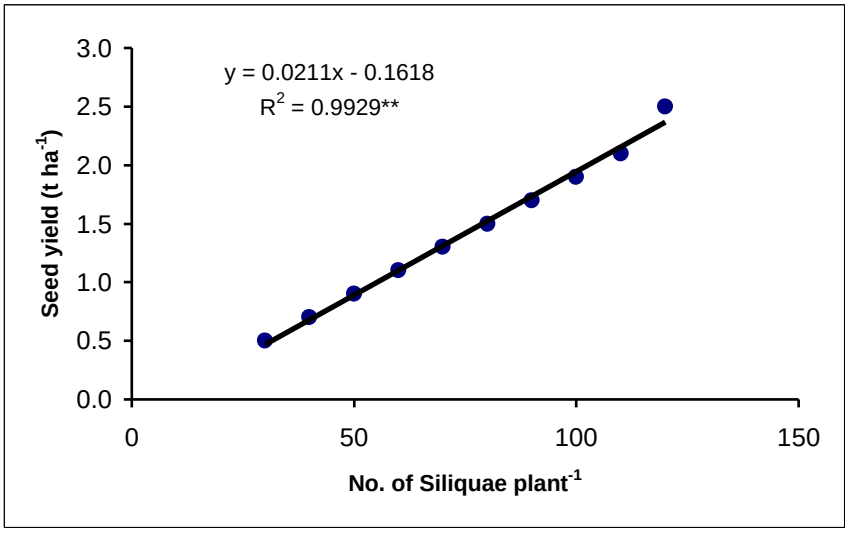

Fig. Relationship between siliquae plant^{-1} and seed yield of mustard

Utility of coefficient of determination

The coefficient of variation is useful in regression analysis to inspect the degree of linear correlation (r) between variables, whether the variables are dependent or independent.

Range of coefficient of determination

The range of coefficient of determination lies between 0 and +1, symbolically $0 \leq R^2 \leq 1$.