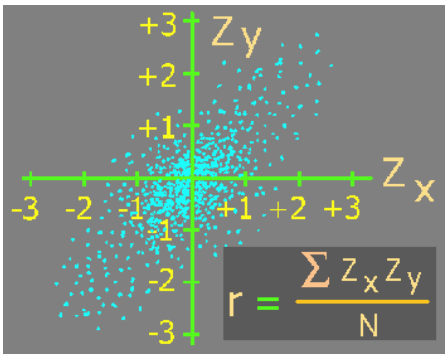
CORRELATION AND REGRESSION

V.K.Sharma and Rajender Parsad
I.A.S.R.I., Library Avenue, New Delhi – 110012
vksharma@iasri.res.in

1. Correlation

Given a pair of related measures (X and Y) on each of a set of items, the correlation coefficient (r) provides an index of the degree to which the paired measures co-vary in a linear fashion. In general r will be positive when items with large values of X also tend to have large values of Y whereas items with small values of X tend to have small values of Y. Correspondingly, r will be negative when Items with large values of X tend to have small values of Y whereas items with small values of X tend to have large values of Y. The value of r is calculated by first converting the Xs and Ys into their respective Z Scores and, keeping track of which Z Score goes with which item, determining the value of the mean Z Score product. Numerically, r can assume any value between -1 and +1 depending upon the degree of the linear relationship. Plus and minus one indicate perfect positive and negative relationships whereas zero indicates that the X and Y values do not co-vary in any linear fashion.

This is also called as *Pearson-product- moment correlation coefficient*. The values of the correlation coefficient have no units. While scatter plot provides a picture of the relation, the value of the correlation is the same if you switch the *Y* (vertical) and *X* (horizontal) measures.



Let $(x_i, y_i), i = 1, 2, \ldots, n$ denote a random sample of n observations from a bivariate population. The sample correlation coefficient r is estimated by the formula

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}.$$

1.1 Test of significance of correlation coefficient

Case 1: $H_0 : \rho = 0$

The variables X, Y follow a bivariate normal distribution. If the population correlation coefficient of X and Y is denoted by ρ, then it is often of interest to test whether ρ is zero or different from zero, on the basis of observed correlation coefficient, r. Thus, if r is the sample correlation coefficient based on a sample of n observations, then the appropriate test statistic for testing the null hypothesis $H_0 : \rho = 0$ against the alternative $H_1 : \rho \neq 0$ is:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Comparison of the computed value of $|t|$, with the table value of *t*-distribution with $(n-2)$ degrees of freedom, and at a given level of significance, say 5 % will indicate the existence or non-existence of correlation. If the computed value of $|t|$ exceeds the table value, then $H_0 : \rho = 0$ is rejected against the alternative $H_1 : \rho \neq 0$.

Case 2: $H_0 : \rho = \rho_0$

Sometimes in a bivariate normal population in which $\rho \neq 0$, one may be interested in testing $H_0 : \rho = \rho_0$ against the alternative $H_1 : \rho \neq \rho_0$. Then we compute, the following quantity

$$\frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right)$$

which is the value of a random variable that follows approximately the normal distribution with mean $\frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right)$ and variance $1/(n-3)$. Thus the test procedure is to compute

$$Z = \frac{\sqrt{n-3}}{2}\left(\log_e\left(\frac{1+r}{1-r}\right) - \log_e\left(\frac{1+\rho_0}{1-\rho_0}\right)\right) = \frac{\sqrt{n-3}}{2}\log_e\left[\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0}\right]$$

and compare to the critical points of the standard normal distribution. For example, if the absolute value of Z, $|Z| > 1.96$, then the null hypothesis $H_0 : \rho = \rho_0$ against the alternative $H_1 : \rho \neq \rho_0$ is rejected at 5% level of significance. The alternative hypotheses $\rho < \rho_0$ or $\rho > \rho_0$ can also be tested using one tailed critical points.

Exercise 1.1: The following data was collected through a pilot sample survey on Hybrid Jowar crop on yield and biometrical characters. The characters were average Plant Population (PP), average Plant Height (PH), average Number of Green Leaves (NGL) and Yield (kg./plot). Compute bivariate correlations among yield, PP, PH and NGL.

PP	PH	NGL	Yield
142	0.53	8.2	2.47
107	0.66	9.3	3.31
78	0.66	7.5	1.97
100	0.46	5.9	1.34
86.5	0.35	6.4	1.14

103.5	0.86	6.4	1.5
155.99	0.33	7.5	2.03
80.88	0.28	8.4	2.54
61.77	0.27	8.3	2.91
79.11	0.66	11.6	2.76
155.99	0.42	8.1	0.59
61.81	0.34	9.4	0.84
74.5	0.63	8.4	3.87
93.14	0.68	6.4	3.31
37.43	0.67	8.4	1.57
36.44	0.28	7.4	0.53
51	0.28	7.4	1.15
104	0.28	9.8	1.08
49	0.49	4.8	1.83
54.66	0.39	5.5	0.76
55.55	0.27	5	0.43
88.44	0.98	5	4.08
99.55	0.65	9.6	2.83
63.99	0.64	5.6	2.57
138.66	0.72	9.9	2.62
90.22	0.63	8.4	2
76.92	1.25	7.3	1.99
126.22	0.58	6.9	1.36
80.36	0.61	6.8	0.68
56.5	0.36	9.7	2.12
144.5	0.61	9.8	3.12
157.33	0.61	8.8	2.07
91.99	0.38	7.7	1.17
121.5	0.55	7.7	3.62
64.5	0.32	5.7	0.67
116	0.46	6.8	3.05
77.5	0.72	11.8	1.7
70.43	0.63	10	1.55
133.77	0.54	9.3	3.28
89.99	0.49	9.8	2.69

1.2 Rank Correlation

The Spearman rank correlation coefficient is usually calculated on occasions when it is not convenient, economic, or even possible to give actual values to variables, but only to assign a rank order to instances of each variable. It may also be a better indicator that a relationship exists between two variables when the relationship is non-linear.

The rank correlation is the Pearson's Product moment correlation coefficient and is defined as the correlation between ranks of individuals with respect to two characters. This is also known

as *Spearman's Rank correlation coefficient* and lies between –1 and +1. If d_i denotes the difference between the ranks of i^{th} individual and *n* denotes the number of individuals, then the Spearman's Rank Correlation Coefficient is given by

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

If there is a tie in the ranks, then the ranks assigned is the average of the ranks assigned to these individuals had there been no tie. In case of ties, the rank correlation coefficient is given by

$$r = 1 - \frac{6\left(\sum_{i=1}^{n} d_i^2 + T_X + T_Y\right)}{n(n^2-1)},$$

where $T_X = \frac{1}{12}\sum_{i=1}^{s}(m_i^3 - m_i)$ and $T_Y = \frac{1}{12}\sum_{j=1}^{t}(m_j'^3 - m_j')$. Here , there are *s* ties in the X-series and m_i individuals in the i^{th} tie; similarly, there are *t* ties in the Y-series and j^{th} tie has m_j' individuals.

1.3 Partial Correlation

Let Y, X_1 and X_2 be three variables, the *correlation* between the two variables Y and X_1 after removing the linear effect of variable X_2 is called the partial correlation, denoted by the symbol $r_{Y1.2}$, and is estimated as follows:

- Regress variable Y on X_2.
- Regress variable X_1 on X_2.
- Compute residuals for each of the regression equations.
- Compute the usual Pearson correlation between the two sets of residuals.

If we write the ordinary correlation coefficients for Y and X_1, Y and X_2, and X_1 and X_2 as r_{Y1}, r_{Y2}, and r_{12}, respectively, the partial correlation coefficient for Y and X_1 with X_2, held fixed can also be obtained as follows:

$$r_{Y1.2} = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{\left(1-r_{Y2}^2\right)\left(1-r_{12}^2\right)}}.$$

The partial correlation coefficients obtained after removing the effect of one variable as discussed above are called partial correlation coefficients of order one. In some situations, however, we may have to obtain the partial correlation coefficients after eliminating the effects of two or more variables. The number of variables that are used for eliminating the effects is known as the order of the partial correlation coefficient.

Test of Significance of Partial Correlation Coefficient

To test $H_0 : \rho_{ij.12\cdots k+2} = 0$ against $H_1 : \rho_{ij.12\ldots k+2} \neq 0$ compute

$$t = \frac{r_{ij.12\cdots k+2}}{\sqrt{1 - r^2_{ij.12\ldots k+2}}} \sqrt{n-k-2}$$

where k is the order of the coefficient. This statistic follows t-distribution with $n-k-2$ degrees of freedom. Reject H_0 if $|t| > t_{\alpha/2, n-k-2}$.

3. Regression

The general purpose of regression is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable.

Suppose $X_1, X_2, \ldots, X_p$ are the cause of variation in Y, we fit multiple regression of y on $x's$ to account for this variation. Multiple regression of y on $x's$ is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

where β_0 denotes intercept and $\beta_i's \ \ (i = 1,2,\ldots, p)$ are called regression coefficients. ε is random error. β_i gives average change in y per unit change in x_i keeping other $x's$ constant.

In the above equation, y is the *dependent* or *outcome* or *predicted* variable, the one you are trying to predict; $x_1, x_2, \ldots, x_p$ are the *independent* or *predictor* variables. If the model is a good descriptor of the relation between the variables, one can use the estimates of the coefficients to predict the value of the dependent variable for new cases.

Fitting of multiple regression model

Suppose n observation are made on y and $x's$. Then for each observation we have our unobserved error term, ε_i. We make the following assumptions regarding the $\varepsilon_i's$, which are random variables (i) errors are independent (ii) errors have zero mean and constant variance (σ^2). These assumptions can also be written as

$$E(\varepsilon_i) = 0, \ \ Var(\varepsilon_i) = \sigma^2 \ \ \text{for all } i = 1,2,\cdots, n.$$
$$Cov(\varepsilon_i, \varepsilon_{i'}) = 0 \qquad \text{for all } i \neq i' = 1,2,\ldots, n$$

We assume that x's are linearly independent. In order to estimate the unknown parameters $\beta_0, \beta_1, \beta_{2,\ldots}, \beta_p$, we use the method of least squares which requires minimization of the error sum of squares, given by

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_p x_{pi}\right)^2$$

Differentiating S w.r.t. $\beta_0, \beta_1, \beta_{2,\ldots}, \beta_p$ and equating the derivatives to zero, we get a set of $p+1$ equations, known as normal equations, in $p+1$ unknowns as

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{1i} + \beta_2 \sum_{i=1}^{n} x_{2i} + \ldots + \beta_p \sum_{i=1}^{n} x_{pi}$$

$$\sum_{i=1}^{n} x_{1i} y_i = \beta_0 \sum_{i=1}^{n} x_{1i} + \beta_1 \sum_{i=1}^{n} x_{1i}^2 + \beta_2 \sum_{i=1}^{n} x_{1i} x_{2i} + \ldots + \beta_p \sum_{i=1}^{n} x_{1i} x_{pi}$$

$$\sum_{i=1}^{n} x_{2i} y_i = \beta_0 \sum_{i=1}^{n} x_{2i} + \beta_1 \sum_{i=1}^{n} x_{1i} x_{2i} + \beta_2 \sum_{i=1}^{n} x_{2i}^2 + \ldots + \beta_p \sum_{i=1}^{n} x_{2i} x_{pi}$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{pi} y_i = \beta_0 \sum_{i=1}^{n} x_{pi} + \beta_1 \sum_{i=1}^{n} x_{1i} x_{pi} + \beta_2 \sum_{i=1}^{n} x_{2i} x_{pi} + \ldots + \beta_p \sum_{i=1}^{n} x_{pi}^2$$

These normal equations can be solved simultaneously to get $p+1$ unknowns. However, it is better to solve these equations by inverting the matrix of moments on the left hand side as this enables us to test significance of $\beta's$ in a straightforward manner. The above equations can be written as

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{pi} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} & \cdots & \sum x_{1i}x_{pi} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 & \cdots & \sum x_{2i}x_{pi} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum x_{pi} & \sum x_{1i}x_{pi} & \sum x_{1i}x_{pi} & \cdots & \sum x_{pi}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{1i} y_i \\ \sum x_{2i} y_i \\ \vdots \\ \sum x_{pi} y_i \end{bmatrix}$$

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} = \begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{pi} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} & \cdots & \sum x_{1i}x_{pi} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 & \cdots & \sum x_{2i}x_{pi} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum x_{pi} & \sum x_{1i}x_{pi} & \sum x_{1i}x_{pi} & \cdots & \sum x_{pi}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum x_{1i} y_i \\ \sum x_{2i} y_i \\ \vdots \\ \sum x_{pi} y_i \end{bmatrix}$$

where b_i's denotes the least squares estimators of β_i's.

Let the inverse matrix on the right hand side of the above be denoted by

$$c = \begin{bmatrix} c_{00} & c_{01} & c_{02} & \cdots & c_{0p} \\ & c_{11} & c_{12} & \cdots & c_{1p} \\ & & c_{22} & \cdots & c_{2p} \\ & & & \cdots & \vdots \\ & & & & c_{pp} \end{bmatrix}$$

Then , $Var(b_j) = c_{jj}\sigma^2$, $j = 0,1,2,\cdots, p$ and $Cov(b_j, b_{j'}) = c_{jj'}\sigma^2$.

The fitted equation is thus given by

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

The 'hat' over y indicates that if we substitute for $x's$ values that is within the observed range of the predictor $x's$, but has not necessarily been observed, then the regression equation gives us the predicated y for that given values of $x's$.

Estimation of σ^2

In addition to estimating $\beta's$ an estimate of σ^2 is required to test hypotheses and construct interval estimates pertinent to the regression model. An estimate of σ^2 is obtained from the residual or error sum of squares.

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \quad = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The residual sum of squares has $(n - p - 1)$ degrees of freedom, because $p + 1$ degrees are associated with the estimates $\beta's$ involved in obtaining $\hat{y}_i$. Now the expected value of SSE is $E(SSE) = (n - p - 1)\sigma^2$ so an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n - p - 1} = MSE$$

The quantity MSE is called the error mean square or the residual mean square.

The splitting of the total sum of squares due to $y's$ into two components can be formally put in an Analysis of variance table, as below:

Analysis of Variance for Multiple Linear Regression

Source of Variation	**d.f.**	**S.S.**	**M.S.**
Regression	p	$\sum b_j S_{x_j y}$	MSR
Deviation form Regression (Residual)	$n - p - 1$	SSE	$s^2 = SSE/(n - p - 1) = MSE$
Total (corrected mean)	$n - 1$	S_{yy}	

We have partitioned total variation in y in two parts as variations due to regression and deviation from regression. For testing the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ against H_1: at least one of $\beta_i \neq 0$, we assume that the errors are normally distributed. The test statistic for testing the above hypothesis is given by

$F_0 = \dfrac{MSR}{MSE}$. F_0 follows $F_{p,n-p-1}$ under H_0.

Test of Significance of β' s

Now if we are interested in testing the hypothesis $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$ for some j. The appropriate test statistic for testing this is

$$t = \frac{|b_j|}{SE(b_j)} = \frac{|b_j|}{\sqrt{c_{jj} s^2}} \text{ as } \hat{Var}(b_j) = c_{jj} s^2$$

Which follows, a t-distribution with $(n - p - 1)$ d.f, if H_0 is true. Reject H_0 if $t > t_{\alpha/2, n-p-1}$.

Multiple Correlation Coefficient (R)

The correlation coefficient between the observed values y_i and predicted values $\hat{y}_i$ is called as multiple correlation coefficient (R). Note that $0 \leq R \leq 1$. R is obtained as

$$R = \sqrt{\frac{\text{Sum of Squares due to regression} \mid \beta_0}{\text{Total corrected sum of squares of } y}}$$

$$= \sqrt{\frac{b_0 \sum_{i=1}^{n} y_i + b_1 \sum_{i=1}^{n} x_{1i} y_i + b_2 \sum_{i=1}^{n} x_{2i} y_i + b_p \sum_{i=1}^{n} x_{pi} y_i - \left(\sum_{i=1}^{n} y_i \right)^2 \frac{1}{n}}{\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y \right)^2 \frac{1}{n}}}$$

$$= \sqrt{\frac{\sum_{j=1}^{p} b_j S_{x_j y}}{S_{yy}}}$$

Test of Significance of R

The test of the null hypothesis that multiple correlation coefficient in the population is zero is identical to the F-test of the null hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. The relation is

$F = \dfrac{R^2}{1-R^2} \quad \dfrac{n-p-1}{p}$. This F follows F-distribution with p and $(n-p-1)$ d.f. Reject H_0 if $F > F_{\alpha, p, n-p-1}$.

Coefficient of Determination $\left(R^2\right)$

The sample coefficient of multiple determination, denoted by $R^2_{Y.12\ldots p}$, is given by

$$R^2_{Y.12\ldots p} = 1 - \frac{SSE}{S_{yy}}$$

where $SSE = S_{yy} - b_1 S_{x_1 y} - b_2 S_{x_2 y} - \cdots - b_p S_{x_p y}$

One can easily see that the coefficient of determination is the square of multiple correlation coefficient and is denoted by $\left(R^2\right)$. This concept is very important as $R^2 \times 100$ gives percentage of variation in y explained by regressors. Obviously, R^2 lies between 0 and 1. However, a large value of R^2 should not alone be taken as a measure of goodness of fitted regression model.

For p=1, the above description reduces to that of simple linear regression.

Example 3:

Observation No.	y	x_1	x_2	x_3	x_4
1	78.5	7	26	6	60
2	74.3	1	29	15	22
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	119.4	10	68	8	12

Fit a simple linear regression equation between y and x_1. Also fit a multiple linear regression equation with y as dependent and x_1, x_2, x_3 and x_4 as independent variables.

Solution: Simple Linear Regression

Model to be fitted is $y = \beta_0 + \beta_1 x_1 + \varepsilon$

Normal equations for estimation of parameters are

$$13\beta_0 + 97\beta_1 = 1250.5$$
$$97\beta_0 + 1139\beta_1 = 10132.0$$

These can also be written as

$$\begin{bmatrix} 13 & 97 \\ 97 & 1139 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 1250.5 \\ 10132.0 \end{bmatrix}$$

Parameter estimates are

$$b_0 = 81.792, \qquad b_0 = 1.930, \qquad s^2 = 140.01$$
$$(5.437) \qquad (0.581)$$

The figures in the parenthesis denote the SE of the estimated parameter.

Fitted model is $y = 81.792 + 1.930x_1 \qquad \left(r^2 = 0.501\right)$

Test of significance of $\beta's$

(a) $H_0 : \beta_0 = 0, \qquad H_1 : \beta_0 \neq 0,$

$$t = \frac{81.792}{5.437} = 15.043 \qquad \left(t_{.05'11} = 2.201\right)$$

(b) $H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0,$

$$t = \frac{1.93}{.581} = 3.322$$

Estimated mean response at $x_1 = 4$

$$\hat{y}_0 = 81.792 + 1.930(4) = 89.512$$
$$\hat{V}\left(\hat{y}_0\right) = 14.829$$

From the above example, it can be seen that 97.7% of the variation in y is explained by x_1, x_2, x_3 and x_4. Coefficients of x_1 and x_2 are significantly different from zero whereas that of x_3 and x_4 are not.

4. Remarks

One should first explore the variables graphically in scatter plots to ascertain if a linear model is appropriate for describing the relationship and to identify any possible rogue values (outliers) that might distort results.

The assumption of normality is not required for the estimation of the coefficients by least squares method. To make tests and estimate confidence intervals, however, these assumptions are required:

- The errors are normally distributed with mean 0.
- The errors have constant variance.
- The errors are independent of each other.

These assumptions are checked by studying the residuals from the model. The **Durbin-Watson** statistic can be used to test for the serial correlation of adjacent error terms.

To identify problems, always look at plots of y versus x before the regression and plots of residuals and diagnostics after the analysis. Non-linearity and outliers can distort the results of regression analysis. Relationships among the dependent and independent variables may be *masked or falsely enhanced* if outliers are present.

Exercise: For the data given in Exercise 1.1, perform the following:

1. Plot a simple scatter diagram between
   (i) yield and PP
   (ii) yield and PH
   (iii) yield and NGL.
2. Plot a scatter diagram using matrix option using the variables yield, PP, PH and NGL.
3. Compute bivariate and partial correlations among yield, PP, PH and NGL.
4. Fit a simple linear regression by taking yield as dependent variable and NGL as independent variable.
5. Fit a multiple linear regression equation by taking yield as dependent variable and biometrical characters as explanatory variables.
6. Compute various statistics viz. Estimates, Confidence intervals, Covariance matrix, Model fit, R-squared change, Descriptives, Part and partial correlations, diagnostics and Residuals.
7. Identify the outliers in the data set.
8. Fit the model without intercept.