# Units 5: Correlation and Regression Analysis

**Bbsnotes.com**

**What is correlation?**

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

**How is correlation measured?**

The sample correlation coefficient, r, quantifies the strength of the relationship. Correlations are also tested for statistical significance.

i. **Direct method: When actual data are given,**

$r_{12}$ = **Correlation coefficient between $X_1$ and $X_2$**

$$= \frac{n\sum X_1X_2 - \sum X_1 \sum X_2}{\sqrt{n\sum X_1^2 - (\sum X_1)^2}\sqrt{n\sum X_2^2 - (\sum X_2)^2}}$$

$r_{23}$ = **Correlation coefficient between $X_2$ and $X_3$**

$$= \frac{n\sum X_2X_3 - \sum X_2 \sum X_3}{\sqrt{n\sum X_2^2 - (\sum X_2)^2}\sqrt{n\sum X_3^2 - (\sum X_3)^2}}$$

$r_{13}$ = **Correlation coefficient between $X_1$ and $X_3$**

$$= \frac{n\sum X_1X_3 - \sum X_1 \sum X_3}{\sqrt{n\sum X_1^2 - (\sum X_1)^2}\sqrt{n\sum X_3^2 - (\sum X_3)^2}}$$

**Example 7:**

The information about the simple correlation coefficient between $X_1$ and $X_2$; $X_1$ and $X_3$, and $X_2$ and $X_3$ are as follows: $r_{12} = 0.59$, $r_{13} = 0.46$ and $r_{23} = 0.77$.

**Compute:**

i. Partial correlation coefficient between variables $X_1$ and $X_2$ keeping $X_3$ as constant.

ii. Partial correlation coefficient between variables $X_1$ and $X_3$ keeping $X_2$ as constant.

iii. Partial correlation coefficient between variables $X_3$ and $X_2$ keeping $X_1$ as constant.

iv. Coefficient of partial determination of $r_{12.3}$.

**Solution:**

Here,

Simple correlation coefficient between variables $X_1$ and $X_2$

$r_{12} = 0.59$

Simple correlation coefficient between variables $X_1$ and $X_3$

$r_{13} = 0.46$

Simple correlation coefficient between variables $X_3$ and $X_2$

$r_{23} = 0.77$

Now,

i. Partial correlation coefficient between variables $X_1$ and $X_2$ keeping $X_3$ as constant is given by,

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

$$= \frac{0.59 - 0.46 \times 0.77}{\sqrt{1-(0.49)^2}\sqrt{1-(0.77)^2}}$$

$$= \frac{0.2358}{\sqrt{0.6519}\sqrt{0.4071}}$$

$$\therefore r_{12.3} = 0.4162$$

ii. Partial correlation coefficient between variables $X_1$ and $X_3$ keeping $X_2$ as constant is given by,

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}}$$

$$= \frac{0.46 - 0.59 \times 0.77}{\sqrt{1-(0.59)^2}\sqrt{1-(0.77)^2}}$$

$$= \frac{0.0057}{\sqrt{0.6519}\sqrt{0.4071}}$$

$$\therefore r_{13.2} = 0.0111$$

iii. Partial correlation coefficient between variables $X_2$ and $X_3$ keeping $X_1$ as constant is given by,

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}}$$

$$= \frac{0.77 - 0.59 \times 0.46}{\sqrt{1-(0.59)^2}\sqrt{1-(0.46)^2}}$$

$$= \frac{0.4986}{\sqrt{0.6519}\sqrt{0.7884}}$$

$$\therefore r_{23.1} = 0.6955$$

iv. Coefficient of partial determination of $r_{12.3}$ is

$$r_{12.3}{}^2 = 0.173 = 17.3\%$$

It means 17.3% total variation on $X_1$ is explained by the variable $X_2$ and controling the effect of variable $X_3$.

**relations** ... lationship (association)

... corresponding values of $X_1$ and $X_2$.

**Example 12:**

For the following data on sales of product, the advertising expenses and the price of the product, estimate the sales with advertising expenses Rs.100 and price Rs. 100. Also calculate the coefficient of multiple determination.

| Year | Y (Sales "oo") | $X_1$(advertising expenses ,00) | $X_2$price"00) |
|---|---|---|---|
| 1 | 2 | 1 | 2 |
| 2 | 3 | 2 | 2 |
| 3 | 5 | 3 | 2 |
| 4 | 4 | 5 | 3 |
| 5 | 1 | 7 | 4 |
| 6 | 2 | 6 | 4 |
| 7 | 3 | 4 | 5 |
| 8 | 2 | 5 | 6 |
| 9 | 5 | 3 | 6 |
| 10 | 3 | 4 | 6 |
| Total | 30 | 40 | 40 |

**Solution:**

Let, the regression line of Y on $X_1$ and $X_2$ is

$Y = a + b_1X_1 + b_2X_2$ ........................(i)

The normal equations are

$\sum Y = na + b_1 \sum X_1 + b_2 \sum X_2$ ........ (ii)

$\sum YX_1 = a\sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1X_2$ ...... (iii)

$\sum X_2Y = a\sum X_2 + b_1 \sum X_2X_1 + b_2 \sum X_2^2$ ....... (iv)

| Y | $X_1$ | $X_2$ | $YX_1$ | $YX_2$ | $X_2X_1$ | $X_1^2$ | $X_2^2$ | $Y^2$ |
|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 2 | 2 | 4 | 2 | 1 | 4 | 4 |
| 3 | 2 | 2 | 6 | 6 | 4 | 4 | 4 | 9 |
| 5 | 3 | 2 | 15 | 10 | 6 | 9 | 4 | 25 |
| 4 | 5 | 3 | 20 | 12 | 15 | 25 | 9 | 16 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 7 | 4 | 7 | 4 | 28 | 49 | 16 | 1 |
| 2 | 6 | 4 | 12 | 8 | 24 | 36 | 16 | 4 |
| 3 | 4 | 5 | 12 | 15 | 20 | 16 | 25 | 9 |
| 2 | 5 | 6 | 10 | 12 | 30 | 25 | 36 | 4 |
| 5 | 3 | 6 | 15 | 30 | 18 | 9 | 36 | 25 |
| 3 | 4 | 6 | 12 | 18 | 24 | 16 | 36 | 9 |
| 30 | 40 | 40 | 111 | 119 | 171 | 190 | 186 | 106 |

From equations (i), (ii), and (iv), we get

$30 = 10a + 40b_1 + 40 b_2$ ... (v)

$111 = 40a + 190b_1 + 171b_2$ ... (vi)

And, $119 = 40 a + 172 b_1 + 186 b_2$ ... (vii)

Multiplying equation (iv) by4 and solving with (v), we get

$9 = -30 b_1 - 11b_2$ ... (viii)

Again, solving (v) and (vii), we get

$-8 = 19 b_1 - 15b_2$ .... (ix)

Multiplying equation (viii) by 19 and equation (ix) by 30 on Solving, we get,

$-69 = -659 b_2$

Or, $b_2 = 0.105$

From equation (ix), we have

$-8 = 19 b_1 - 15 \times 0.105$

Or, $b_1 = - 0.338$

From equation (v), we have

$30 = 10a + 40 \times (- 0.338) + 40 \times 0.105$

Or, $a = 3.932$

Now the regression equation of Y on $X_1$ and $X_2$ is

$\hat{Y} = 3.932 - 0.338X_1 + 0.105 X_2$

If $X_1 = 10$ and $X_2 = 10$, the estimated value of Y is

$\hat{Y} = 3.932 - 0.338 \times 10 - 0.105 \times 10 =$

Now, the multiple correlation coefficient is given by,

$$R_{Y.X_1X_2} = \sqrt{\frac{a\sum Y + b_1 \sum YX_1 + b_2 \sum YX_2 - n\bar{Y}^2}{\sum Y^2 - n\bar{Y}^2}}$$

$$= \sqrt{\frac{3.932\times30-0.338+111+0.10\times119-10\times3^2}{106-10\times3^2}}$$

$$= \sqrt{\frac{2.932}{16}} = 0.428$$

Now, the coefficient of determination is given as

$$R_{Y.X_1X_2}{}^2 = (0.428)^2 = 0.183$$

**Example 13:**

Estimate the expenditure on the food of a family with an annual income of Rs. 60,000 and 4 family size of the following data

| Expenditure on food (Rs. '000) (Y) | Annual income (Rs. '000) ($X_1$) | Family size ($X_2$) |
|---|---|---|
| 7 | 30 | 3 |
| 9 | 45 | 2 |
| 10 | 35 | 4 |
| 11 | 55 | 5 |
| 13 | 30 | 1 |

**Solution:**

Let, the regression line of Y on $X_1$ and $X_2$ is

$Y = a + b_1X_1 + b_2X_2$ ..........................(i)

The normal equations are

$\sum Y = na + b_1 \sum X_1 + b_2 \sum X_2$ ........ (ii)

$\sum YX_1 = a\sum X_1 + b_1 \sum X_1{}^2 + b_2 \sum X_1X_2$ ...... (iii)

$\sum X_2Y = a\sum X_2 + b_1 \sum X_2X_1 + b_2 \sum X_2{}^2$ ....... (iv)

| Y | $X_1$ | $X_2$ | $YX_1$ | $YX_2$ | $X_2X_1$ | $X_1^2$ | $X_2^2$ | $\hat{Y}$ |
|---|---|---|---|---|---|---|---|---|
| 7 | 30 | 3 | 210 | 21 | 90 | 900 | 9 | 28.531 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 9 | 45 | 2 | 405 | 18 | 90 | 2025 | 4 | 39.765 |
| 10 | 35 | 4 | 350 | 40 | 140 | 1225 | 16 | 31.497 |
| 11 | 55 | 5 | 605 | 55 | 275 | 3025 | 25 | 45.113 |
| 13 | 30 | 1 | 390 | 13 | 30 | 900 | 1 | 29.699 |
| **195** | **15** | **50** | **1960** | **147** | **625** | **8075** | **55** | **174.605** |

From equations (i), (ii), and (iv), we get

$50=5a+195b_1+15b_2 \Rightarrow 10=a+39b_1+3b_2$ ....... (v)

$1960=195a+8075b_1+625b_2 \Rightarrow 392=39a+1615b_1+125b_2$........ (vi)

And, $147=15a+625b_1+55b_2$ ........ (vii)

Solving (vi) and (v), we get

$1=47b_1+4b_2$ .......... (viii)

Again, solving (v) and (vii), we get

$3=-40b_1-10b_2$ ......... (ix)

Solving (vii) and (ix), we get $b_1 = 0.071$

substituting this value of $b_1$ in equation (viii) we get, $b_2 = -0.584$

Now, again putting these values in equation (v), we get a = 8.983

Now the regression equation becomes,

$Y = 8.983 + 0.071X_1 - 0.584X_2$

If $X_1 = 60$ and $X_2 = 4$, the estimated value of Y is

$\hat{Y} = 8.983 + 0.071 \times 60 - 0.584 \times 4 = 10.907$

Hence, expected expenditure in food is Rs. 10907

**Example 14:**

Estimate the multiple regression equation of yield of the crop on the amount of rainfall and fertilizer from the following data.

| | | | | | |
|---|---|---|---|---|---|
| Yields of crops (in '000' kg) | 5 | 7 | 8 | 4 | 9 |
| Amount of rainfall (in inches) | 2 | 3 | 4 | 3 | 4 |
| Amount of fertilizer (in kg) | 2 | 0 | 3 | 1 | 2 |

i. Estimate the yield of crops when the amount of rainfall is 6 and the amount of fertilizer is 5 kg.

ii. How much variation in yield of crops is explained by rainfall and fertilizer?

iii. Find the standard error of the estimate.

**Solution:**

Y = Yield of crops (in '000' kg), dependent variable

$X_1$ = Amount of rainfall (in inches)

$X_2$ = Amount of fertilizer (in kg)

The multiple regression equation is given by,

$$Y = a + b_1X_1 + b_2X_2 \quad \ldots \quad (i)$$

Where, a = Sample y- intercept

$b_1$ = Sample regression coefficient of Y on $X_1$ Keeping effect of $X_2$ constant.

$b_2$ = Sample regression coefficient of Y on $X_2$ Keeping effect of X constant.

The normal equations are given as

$$\Sigma Y = n\,a + b_1\Sigma X_1 + b_2\Sigma X_2 \quad \text{... (ii)}$$

$$\Sigma X_1Y = a\Sigma X_1 + b_1\Sigma X^2{}_1 + b_2\Sigma X_1X_2 \quad \text{... (iii)}$$

$$\Sigma X_2Y = a\Sigma X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X^2{}_2 \quad \text{... (iv)}$$

**Calculation**

| **Y** | **$X_1$** | **$X_2$** | **$X^2{}_1$** | **$X^2{}_2$** | **$X_1X_2$** | **$YX_1$** | **$YX_2$** | **$Y^2$** |
|---|---|---|---|---|---|---|---|---|
| 5 | 2 | 2 | 4 | 4 | 4 | 10 | 10 | 25 |
| 7 | 3 | 0 | 9 | 0 | 0 | 21 | 0 | 49 |
| 8 | 4 | 3 | 16 | 9 | 12 | 32 | 24 | 64 |
| 4 | 3 | 1 | 9 | 1 | 3 | 12 | 4 | 16 |
| 9 | 4 | 2 | 16 | 4 | 8 | 36 | 18 | 81 |
| **$\Sigma Y$ = 33** | **$\Sigma X_1$ = 16** | **$\Sigma X_2$ = 8** | **$\Sigma X^2{}_1$ = 54** | **$\Sigma X^2{}_2$ = 18** | **$\Sigma X_1X_2$ = 27** | **$\Sigma YX_1$ = 111** | **$\Sigma YX_2$ = 56** | **$\Sigma Y^2$ = 235** |

Here, n= 5

From normal equations, we get

$$33 = 5a + 16b_1 + 8b_2 \quad \text{... (v)}$$

$$111 = 16a + 54b_1 + 27b_2 \quad \text{... (vi)}$$

$$56 = 8a + 27b_1 + 18b_2 \quad \text{... (vii)}$$

Solving equations (v) and (vi) {multiplying equation (v) by 16 and equation (vi) by 5 and subtracting}, we get

$$528 = 80 + 256b_1 + 128b_2$$

$$555 = 80a + 270b_1 + 135b_2$$

$$- \quad - \quad - \quad -$$

$$-27 = -14b_1 - 7b_2 \quad \text{... (viii)}$$

Solving equations (vi) and (vii) {multiplying equation (vii) by 2 and then subtracting}, we get

$$111 = 16a + 54b_1 + 27b_2$$
$$112 = 16a + 54b_1 + 36b_2$$
$$- \quad - \quad - \quad -$$

$$-1 = -9b_2$$
$$\therefore b_2 = 0.111$$

From equation (viii) we get

$$27 = 14b_1 + 7 \times 0.111$$

or, $b_1 = 1.873$

Again, from equation(iv) we get

$$33 = 5a + 16 \times 1.873 + 8 \times 0.111$$

or, $33 = 5a + 30.856$

or, $5a = 2.144$

$\therefore$ $a = 0.4288$

Substituting the values of a, $b_1$ and $b_2$ in equation (i), we get the multiple regression equation as

$$\hat{Y} = 0.429 + 1.873X_1 + 0.111X_2$$

If $X_1 = 6$ and $X_2 = 5$ then,

$$\hat{Y} = 0.429 + 1.873 \; x \; 6 + 0.111 \; x \; 5$$
$$= 12.222\text{kg}$$

ii. The coefficient of multiple determination is

$$R^2_{Y.12} = \frac{a\sum Y + b_1 \sum X_1 Y + b_2 \sum X_2 Y - n\bar{Y}^2}{\sum Y^2 - n\,\bar{Y}^2}, \text{ Where } \bar{Y} = \frac{\sum Y}{n} = \frac{33}{5} = 6.6$$

$$= \frac{0.429 \times 33 + 1.873 \times 111 + 0.111 \times 56 - 5 \times (6.6)^2}{235 - 5 \times (6.6)^2}$$

$$= \frac{10.476}{17.2}$$

$$= 0.6091$$

$$\therefore \quad R^2_{Y.12} = 0.6091 = 60.91\%$$

This indicates that 60.91% of the total variation in yield of crops is explained by the independent variables rainfall and fertilizer and the remaining 39.09% is the effect of other factors (unexplained variation).

iii. The standard error of estimation is given as:

$$S_e^{\,2} = \sqrt{\frac{\sum Y^2 - a\sum Y - b_1\sum X_1\,Y - b_2\sum X_2\,Y}{n-3}}$$

$$= \sqrt{\frac{235 - 0.429\times 33 - 1.873\times 111 - 0.111\times 56}{5-3}}$$

$$= \sqrt{3.362} = 1.8934$$

This indicates the variation of the actual value from the estimated value.

Since, $S_e = 1.8934 \neq 0$, the estimating equation is not a perfect estimator of the dependent variableY.

**Example 16:**

The information about the simple correlation coefficient between $X_1$ and $X_2$, $X_1$ and $X_3$ and $X_2$ and $X_3$ are as follows:

$r_{12} = 0.8$, $r_{13} = 0.5$ and $r_{23} = 0.9$. Find $r_{12.3}$ and $r_{23.1}$

**Solution:**

Here, $r_{12} = 0.8$, $r_{13} = 0.5$ and $r_{23} = 0.9$

$r_{12.3} =?$ and $r_{23.1} =?$

We have,

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{12}^{\,2}}\sqrt{1-r_{23}^{\,2}}}$$

$$= \frac{0.8 - 0.5\times 0.9}{\sqrt{1-(0.5)^2}\sqrt{1-(0.9)^2}}$$

$$= 0.9272$$

$$r_{23.1} = \frac{r_{23} - r_{13}r_{12}}{\sqrt{1-r_{12}^{\,2}}\sqrt{1-r_{13}^{\,2}}}$$

$$= \frac{0.9 - 0.8\times 0.5}{\sqrt{1-(0.5)^2}\sqrt{1-(0.8)^2}}$$

$$= 0.9623$$

## 5.8 Points to Remember

**Correlation:** It is a statistical measure used to study the degree of relationship (association) between two or more variables.

Karl Pearson's correlation coefficient $r_{XY}$ is given by

$$r_{XY} = \frac{\text{cov}(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

- ✓ The Correlation Coefficient is independent of change of origin and scale.
- ✓ It lies between -1 and 1(i.e $-1 \leq r \leq +1$)
- ✓ It is symmetric, i.e. $r_{13} = r_{31}, r_{12} = r_{21}$.

✓ $r = \sqrt{b_{YX} b_{XY}}$, w where $b_{YX}$ and $b_{XY}$ are regression coefficients.

**Rank Correlation:** It measures the relationship between two variables on an ordinal scale. Spearman's formula for the rank correlation coefficient is given by

$R = 1 - \frac{6\sum d^2}{n(n^2-1)}$, Where d is the difference between the rank of two series.

The correlation coefficient of repeated rank is given as

$$R = 1 - \frac{6\left[\sum d^2 + \frac{m_1(m_1^2-1)}{12}\right]}{n(n^2-1)},$$

Where m is the number of repetitions of items.

**Partial Correlation:** In partial correlation, we studied the relationship between two variables taking the effect of other random variables as constant. Among three variables $X_1$, $X_2$and $X_3$, the partial correlation coefficient between two variables $X_1$ and $X_2$ taking the $X_3$ constant is given by

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

Similarly, $r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}}$ and $r_{23.1} = \frac{r_{23} - r_{13}r_{21}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{21}^2}}$

✓ $-1 \leq r_{12.3} \leq +1$

✓ $r_{12.3} = r_{21.3}$

**Multiple Correlation:** In multiple correlations, we study the relationship (association) between a dependent variable and the combined or joint effect of other independent variables. The multiple correlation coefficient between $X_1$ and the joint effect of $X_2$ and $X_3$ is given by

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

Similarly, $R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}}$ and

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$

- ✓ $0 \le R_{1.23} \le 1$ (the value lies between o and 1).
- ✓ $R_{1.23} = R_{1.32}$, $R_{2.13} = R_{2.31}$, $R_{3.21} = R_{3.12}$
- ✓ $R_{1.23} = 0$ if $r_{12} = 0$ and $r_{13} = 0$

Regression: Regression analysis is defined as the mathematical measures of the average relationship between two or more variables in terms of original units of the data. The regression coefficient is given by

The regression coefficient of Y on X, $b_{yx} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}}$

The regression coefficient of X on Y, $b_{XY} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum y^2 - (\sum y)^2}}$

- ✓ The regression coefficient is the geometric mean of two regression coefficients.i.e., $r = \pm\sqrt{\mathbf{b_{yx.}\, b_{xy}}}$
- ✓ The arithmetic mean of regression coefficients is greater than the correlation coefficients.i.e., $\frac{\mathbf{b_{yx} + b_{yx}}}{2} \geq \mathbf{r}$
- ✓ Both the regression coefficients have the same sign and if one is greater than unity, then the other must be less than unity.