Coefficient of Multiple Correlation

AI22MTECH02003 - Shrey Satapara

Department of Artificial Intelligence
IIT Hyderabad

January 31, 2022

Question 57

Suppose $r_{1.23}$ and $r_{1.234}$ are sample multiple correlation coefficients of X_1 on X_2, X_3 and X_1 on X_2, X_3, X_4 respectively. Which of the following is possible? 1. $r_{1.23} = -0.3$ and $r_{1.234} = 0.7$

2. $r_{1.23} = 0.7$ and $r_{1.234} = 0.3$
3. $r_{1.23} = 0.3$ and $r_{1.234} = 0.7$
4. $r_{1.23} = 0.7$ and $r_{1.234} = -0.3$

Data

- Here, as given in the question $X_2, X_3 X_4$ are independent random variables and X_1 is dependent variable.
- $r_{1.23}$ is coefficient of multiple correlation of X_1 on X_2, X_3 and $r_{1.234}$ is coefficient of multiple correlation of X_1 on X_2, X_3, X_4.

Coefficient of Multiple Correlation

- The coefficient of multiple correlation, denoted R, is a scalar that is defined as the Pearson correlation coefficient between the predicted and the actual values of the dependent variable in a linear regression model that includes an intercept.
- The coefficient of multiple correlation(R) is known as the square root of the coefficient of determination(R^2)
- For any multi-linear regression model R^2 can be calculated using the formula given bellow

$$R^2 = 1 - \frac{sumsquaredregression(SSR)}{totalsumofsquares(SST)} \tag{1}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

(2)

Contd...

where,

y_i is actual value,

$\hat{y}_i$ is predicted value,

and $\bar{y}$ is mean of actual y values.

Solution

- So, in our case X_1 is actual value and $\bar{X_1}$ is mean of X_1 values. Here, values of X_1 variable can be predicted by using independent random variables X_2, X_3, X_4 by using multiple linear regression.
- here, lets say $\hat{X_{123}}$ is predicted value of X_1 using X_2, X_3 and $\hat{X_{1234}}$ is predicted value of X_1 using X_2, X_3, X_4 and β_i are parameters of multiple linear regression.
- So

$$\hat{X_{123_i}} = \beta_2 x_2 + \beta_3 x_3 \tag{3}$$

- and

$$\hat{X_{1234_i}} = \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \tag{4}$$

Solution contd...

- So here,

$$r^2_{1.23} = 1 - \frac{\Sigma(X_{1_i} - \hat{X_{123_i}})^2}{X_{1_i} - \bar{X_1}} \tag{5}$$

- and

$$r^2_{1.234} = 1 - \frac{\Sigma(X_{1_i} - \hat{X_{1234_i}})^2}{X_{1_i} - \bar{X_1}} \tag{6}$$

Solution contd...

- Here from equation 2 we can clearly see that coefficient of multiple correlation cannot be negative. Also, if we introduce more variables, the R^2 will always increase, it can never decrease. This follows mathematically from the observation that,

$$(y - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_p - \beta_{p+1} x_{p+1})^2 \leq (y - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_p)^2 \quad (7)$$

- Hence, in our case $\Sigma(X_{1_i} - \hat{X_{1234_i}})^2 \leq \Sigma(X_{1_i} - \hat{X_{123_i}})^2$ is always true. Means, $r_{1.23} \leq r_{1.234}$ is always true.

Conclusion

- If we look at option 1 value of $r_{1.23}$ is negative and in option 4 value of $r_{1.234}$ is negative hence these options are not possible. In option 2, value of $r_{1.23}$ is grater than value of $r_{1.234}$ which is also not possible.
- **Only option 3 ($r_{1.23} = 0.3$ and $r_{1.234} = 0.7$) satisfy both the conditions hence option 3 is the correct answer**

Thank You