Random Variables and Stochastic Process (AI5030)

Soumyajit Chatterjee
AI22MTECH02005

February 6, 2022

Question 56 (2019)

There are two sets of observations on a random vector (X, Y). Consider a simple linear regression model with an intercept for regressing Y on X. Let $\hat{\beta_i}$ be the least square estimate of the regression coefficient obtained from the ith (i=1, 2) set consisting of n_i observations $(n_1, n_2) > 2$. Let $\hat{\beta_0}$ be the least square estimate obtained from the pooled sample size $n_1 + n_2$. If it is known that $\hat{\beta_1} > \hat{\beta_2} > 0$, which of the following statements is true ?

1. $\hat{\beta_2} < \hat{\beta_0} < \hat{\beta_1}$
2. $\hat{\beta_0}$ may lie outside $(\hat{\beta_2}, \hat{\beta_1})$ but cannot exceed $\hat{\beta_1} + \hat{\beta_2}$
3. $\hat{\beta_0}$ may lie outside $(\hat{\beta_2}, \hat{\beta_1})$ but cannot be negative
4. $\hat{\beta_0}$ can be negative

Solution

Derivation of Least Square estimator

Let a straight line be estimated on a vector (X, Y) using an estimator such that:

$$\hat{Y} = \hat{\beta}X \tag{1}$$

The error in approximation with respect to the actual distribution can be then given as,

$$E = \sum(\hat{Y} - Y)^2 \tag{2}$$

The above equation can be re-written in the matrix form as:

$$E = (\hat{Y} - Y)^T(\hat{Y} - Y) \tag{3}$$

$$= \hat{Y}^T\hat{Y} - Y^T\hat{Y} - \hat{Y}^TY + Y^TY \tag{4}$$

We know that $\hat{Y} = \hat{\beta}X$, therefore substituting in equation 4:

$$E = Y^TY - 2\hat{\beta}^TX^TY + \hat{\beta}^TX^TX\hat{\beta} \tag{5}$$

Differentiating the above equation w.r.t $\hat{\beta}$ and equating it to 0 to obtain the $\hat{\beta}$ that minimizes the error.

$$\frac{dE}{d\hat{\beta}} = -2X^TY + 2X^TX\hat{\beta} = 0 \tag{6}$$

$$X^TX\hat{\beta} = X^TY \tag{7}$$

$$\hat{\beta} = (X^TX)^{-1}X^TY \tag{8}$$

We know that the term X^TX is the term $\sum X^2$ which is always positive. Therefore, the term that determines the sign of $\hat{\beta}$ is X^TY

We now take two cases which eliminates our options:

Case 1:

Sample n_1,

$$\text{X} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad \text{Y} = \begin{bmatrix} -2 \\ -4 \\ -6 \end{bmatrix} \quad \hat{\beta}_1 = -2$$

Sample n_2,

$$\text{X} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} \quad \text{Y} = \begin{bmatrix} -8 \\ -10 \\ -12 \end{bmatrix} \quad \hat{\beta}_2 = -2$$

Sample $n_1 + n_2$,

$$\text{X} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix} \quad \text{Y} = \begin{bmatrix} -2 \\ -4 \\ -6 \\ -8 \\ -10 \\ -12 \end{bmatrix} \quad \hat{\beta}_0 = -2$$

For the above case, $\hat{\beta}_0$ is negative and $\hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_0$, therefore, the options (1) and (3) are invalid.

Case 2:

Sample n_1,

$$X = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad Y = \begin{bmatrix} -2 \\ -4 \\ -6 \end{bmatrix} \quad \hat{\beta}_1 = -2$$

Sample n_2,

$$X = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} \quad Y = \begin{bmatrix} 12 \\ 15 \\ 18 \end{bmatrix} \quad \hat{\beta}_2 = 3$$

Sample $n_1 + n_2$,

$$X = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix} \quad Y = \begin{bmatrix} -2 \\ -4 \\ -6 \\ 12 \\ 15 \\ 18 \end{bmatrix} \quad \hat{\beta}_0 = 5$$

Here, $\hat{\beta}_1$ is -2 and $\hat{\beta}_2$ is 3. Now, $\hat{\beta}_1 + \hat{\beta}_2 = -2 + 3 = 1$. But we see that $\hat{\beta}_0 = 5$ which is outside the range of $\hat{\beta}_1 + \hat{\beta}_2$

Therefore, option (2) is also incorrect. The only option which remains i.e option (4) is correct.