Random Variables and Stochastic Process (AI5030)

Soumyajit Chatterjee
AI22MTECH02005

February 18, 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

Let the straight line estimated on a vector (X, Y) of sample size n_1, n_2 and $n_1 + n_2$ be:

$$y_1 = \hat{\beta}_1 \, x_1 \tag{1}$$

$$y_2 = \hat{\beta}_2 \, x_2 \tag{2}$$

$$y = \hat{\beta}_0 \, x \tag{3}$$

The linear regression coefficient on the vector (X, Y) is given by the formula:

$$\hat{\beta} = (X^T \, X)^{-1} \, (X^T \, Y)$$

The term $(X^T X)^{-1}$ can be written as the square of L_2 norm of X. Therefore, the above equation can be re-written as:

$$\hat{\beta} = \frac{X^T \, Y}{||X||^2}$$

Therefore, from equation 1 and 2 we can write that:

$$\hat{\beta}_1 = \frac{X_1^T Y_1}{||X_1||^2} = \frac{p_1}{q_1} \tag{4}$$

$$\hat{\beta}_2 = \frac{X_2^T Y_2}{||X_2||^2} = \frac{p_2}{q_2} \tag{5}$$

Since, $\hat{\beta}_0$ is the estimator for stacked sample size $n_1 + n_2$ and can be given by:

$$\hat{\beta}_0 = \frac{[X_1^T \; X_2^T] \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}}{||X_1||^2 + ||X_2||^2} \tag{6}$$

Upon simplification we get:

$$\hat{\beta}_0 = \frac{X_1^T Y_1 \; + \; X_2^T Y_2}{||X_1||^2 + ||X_2||^2} \tag{7}$$

Since, all the terms in the numerator are scalars, we can represent the above equation as:

$$\hat{\beta}_0 = \frac{p_1 + p_2}{q_1 + q_2}$$

In the question it is given that $\hat{\beta}_1 > \hat{\beta}_2 > 0$ which implies:

$$\frac{p_1}{q_1} > \frac{p_2}{q_2} > 0$$

Since, q_1 and q_2 is the square of norm, it is always positive. Therefore, from the above equation we can say that $p_1 > 0$ and $p_2 > 0$.
In the expression for $\hat{\beta}_0$,

$$\hat{\beta}_0 = \frac{p_1 + p_2}{q_1 + q_2}$$

All the terms p_1, p_2, q_1, q_2 are positive, therefore, $\hat{beta}_0$ is always positive and cannot be negative. Therefore, option (4) is false.
Let us take an example.

p_1	p_2	q_1	q_2	$\hat{\beta}_1 = p_1/q_1$	$\hat{\beta}_2 = p_2/q_2$	$\hat{\beta}_0 = (p_1 + p_2)/(q_1 + q_2)$	$\hat{\beta}_1 + \hat{\beta}_2$
12	4	3	2	4	2	3.2	6

From the above example for all positive values of $\hat{\beta}_1$ and $\hat{\beta}_2$ we can see that $\hat{\beta}_0$ always lies between $\hat{\beta}_1$ and $\hat{\beta}_2$ and $\hat{\beta}_2 < \hat{\beta}_0 < \hat{\beta}_1$.
That is option (1) is the correct answer.