Random Variables and Stochastic Process (AI5030)

Soumyajit Chatterjee

AI22MTECH02005

Q. 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 β_i be the least square estimate of the regression coefficient obtained from the ith (i=1,2) set consisting of ni observations (n_1, n_2 > 2). Let β_0 the least square estimate obtained from the pooled sample size $n_1 + n_2$. If it is known that $\beta_1 > \beta_2 > 0$, which of the following statements is true ?

1. $\beta_2 < \beta_0 < \beta_1$
2. β_0 may lie outside (β_2, β_1) but cannot exceed $\beta_1+\beta_2$
3. β_0 may lie outside (β_2, β_1) but it cannot be negative
4. β_0 can be negative

Sol.

Least square estimator β can be given as $\beta = \dfrac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$

We know that X $\in$ R, which means X can take any values including negative values. Now, $\sum X^2$ means the summation of all the squared values of X. Therefore if X contains any negative values then those will become positive due to squaring and then summation of all the positive values are taken. However in $(\sum X)^2$ first the summation is done, then it is squared. If X contains some negative values then the sum would be less.

Therefore in all the cases $\sum X^2$ will be greater than $(\sum X)^2$. Therefore, $\sum X^2$ - $(\sum X)^2$ will always be positive.

1. The term $\sum XY$ signifies the sum of products of individual values of X and Y. Therefore, if the product contains many negative values then the sum of the products may very well be negative.

Then, $n(\sum XY) - (\sum X)(\sum Y)$ can also very well be negative. But in the 3rd option it says that β cannot be negative therefore the 3rd option is false.

We know that n is the sample size. We just established that $\sum XY$ may be negative depending on the product terms. We also established that $\sum X^2$ - $(\sum X)^2$ will always be positive no matter what.

Then $n(\sum X^2)$ - $(\sum X)^2$ will always be positive as n is only the sample size and multiplying $\sum X^2$ with n will make it even larger.

Now, β_0 is the estimator of the combined samples of n_1 and n_2. But even then, the denominator of β_0 will be positive.

2. The numerator of β_0 can be positive if $n(\sum XY)$ term is greater than $(\sum X)(\sum Y)$ and since here n is the combined sample size of n_1 and n_2, then the term $n(\sum XY) - (\sum X)(\sum Y)$ for β_0 can be very well greater than even β_1 or β_2. Therefore, β_0 can become greater than β_1 and β_2.

Therefore, both option 1 and 2 would be incorrect.

3. If the term $n\sum XY$ is negative for the combined samples from n_1 and n_2 than the term $(\sum X)(\sum Y)$ then, $n(\sum XY) - (\sum X)(\sum Y)$ would be negative which will make β_0 negative.

Therefore, the only correct option would be option 4 from logical explanation.