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 $\beta_i$ 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 $\beta_0$ 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.** $\beta_0$ may lie outside ($\beta_2$, $\beta_1$) but cannot exceed $\beta_1+\beta_2$
**3.** $\beta_0$ may lie outside ($\beta_2$, $\beta_1$) but it cannot be negative
**4.** $\beta_0$ can be negative

**Sol.**

Derivation for least square estimator:

Let our original data be $y_i$ which is approximated by $\beta_0$ and $\beta_1$ such that $\hat{Y}_i = \beta_0 + \beta_1 x_i$

Therefore, the error in measurement as measured by the least squares metric would be

$$E = \sum_{i-1}^{N} (yi - \hat{Y}i)^2$$

To get the minimum error E we have to differentiate the above equation and set the 1st derivative to 0.

Therefore,

$$\frac{dE}{d\beta 0} = \sum_{i=1}^{N} - 2(y_i - \beta_0 - \beta_1 x_i) = 0$$

Which is:

$$\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \beta_0 - \sum_{i=1}^{N} \beta_1 x_i = 0$$

Now, we know that,

$$\frac{\sum_{i=i}^{N} yi}{N} = \bar{y}$$

Therefore,

$$\sum_{i=1}^{N} y_i = N\bar{y}$$

Substituting above equation in:

$$\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \beta_0 - \sum_{i=1}^{N} \beta_1 x_i = 0$$

The above equation becomes $N\bar{y} - N\beta_0 - N\beta_1 \bar{x} = 0$ which on further simplification becomes

$$\bar{y} - \beta_0 - \beta_1 \bar{x} = 0$$

Therefore, finally

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Similarly, we can differentiate for $\beta_1$ as:

$$\frac{dE}{d\beta 1} = \sum_{i=1}^{N} -2x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

Rearranging the equation:

$$\sum_{i=i}^{N} ( x_i y_i - \beta_0 x_i - \beta_1 x_i^2) = 0$$

Substituting the value of $\beta_{0,}$

$$\sum_{i=i}^{N} ( x_i y_i - (\bar{y} - \beta_1 \bar{x}) x_i - \beta_1 x_i^2) = 0$$

Simplifying,

$$\sum_{i=i}^{N} x_i y_i - \bar{y} \sum_{i=i}^{N} x_i - \beta_1 \bar{x} \sum_{i=i}^{N} x_i - \beta_1 \sum_{i=i}^{N} x_i^2 = 0$$

Again using the property that $\sum_{i=1}^{N} y_i = N\bar{y}$, we get

$$\sum_{i=i}^{N} x_i y_i - N\bar{y}\bar{x} - \beta_1 N\bar{x}^2 - \beta_1 \sum_{i=i}^{N} x_i^2 = 0$$

Which is,

$$\sum_{i=i}^{N} x_i y_i - N\bar{y}\bar{x} - \beta_1( N\bar{x}^2 - \sum_{i=i}^{N} x_i^2 ) = 0$$

Therefore $\beta_1$ becomes,

$$\beta_1 = \frac{\sum_{i=i}^{N} xi\, yi - N\bar{y}\bar{x}}{\sum_{i=i}^{N} xi^2 - N\bar{x}^2}$$

Writing $\bar{y}$ and $\frac{\sum_{i=i}^{N} yi}{N}$ and $\frac{\sum_{i=i}^{N} xi}{N}$ again we get,

$$\beta_1 = \frac{(\sum xy) - N(\frac{\sum x}{N})(\frac{\sum y}{N})}{(\sum x^2) - N(\frac{\sum X}{N})^2}$$

Multiplying by N we get the final expression.

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

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, $\beta_0$ is the estimator of the combined samples of $n_1$ and $n_2$. But even then, the denominator of $\beta_0$ will be positive.

**2.** The numerator of $\beta_0$ 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 $\beta_0$ can be very well greater than even $\beta_1$ or $\beta_2$. Therefore, $\beta_0$ can become greater than $\beta_1$ and $\beta_2$.

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 $\beta_0$ negative.

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