Some OLS Theory: Proofs

2018-09-11

B_1 is a Linear Function of the Observations

We need to show that we can express the slope estimator B_1 as a linear function of the observations, namely,

$$B_1 = \sum w_i Y_i \qquad \text{for some } w_i$$

To do this, we will use the definition of the slope estimator, that

$$B_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Now,

$$\begin{aligned}
B_1 &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} - \frac{\sum(x_i - \bar{x})\bar{y}}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} - \frac{\bar{y}\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}
\end{aligned}$$

Then we make use of the fact that the sum of mean deviations of a random variable is always zero; $\sum(x_i - \bar{x}) = 0$.

$$\begin{aligned}
B_1 &= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} - \frac{\bar{y}(0)}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} - 0 \\
&= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}
\end{aligned}$$

This means that B_1 can be written as a linear function of the observations y_i:

$$B_1 = \sum w_i Y_i \qquad \text{where } w_i = \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2}$$

B_0 is a Linear Function of the Observations

We can carry out a similar process for B_0, but it is far easier. Recall that,

$$B_0 = \bar{y} - B_1\bar{x}$$

Since we just showed that $B_1 = \sum w_i y_i$, we can substitute that in for B_1.

$$B_0 = \bar{y} - \sum w_i y_i \bar{x}$$

Since $\bar{x}$ and $\bar{y}$ are constants, B_0 has been shown to be a linear function of the observations, y_i.

B_1 is Unbiased

To show that the slope estimator is unbiased, we need to show that $\mathbb{E}(B_1) = \beta$. To do this, we will make use of the fact that,

$$\begin{aligned}
\sum(x_i - \bar{x})^2 &= \sum(x_i - \bar{x})(x_i - \bar{x}) \\
&= \sum(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \sum x_i^2 - 2\bar{x}\sum x_i + \sum \bar{x}^2 \\
&= \sum x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 \\
&= \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\
&= \sum x_i^2 - n\bar{x}^2
\end{aligned}$$

and the assumption of linearity in the population (Assumption #2), namely that,

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1(x_i)$$

Since,

$$B_1 = \sum \frac{(x_i - \bar{x})y_i}{\sum x_i^2 - n\bar{x}^2},$$

then the proof is as follows:

$$
\begin{aligned}
\mathbb{E}(B_1) &= \mathbb{E}\Bigg(\sum \frac{(x_i-\bar{x})y_i}{\sum x_i^2 - n\bar{x}^2}\Bigg) \\
&= \mathbb{E}\Bigg(\frac{1}{\sum x_i^2 - n\bar{x}^2} \times \sum \Big[(x_i-\bar{x})y_i\Big]\Bigg) \\
&= \frac{1}{\sum x_i^2 - n\bar{x}^2} \times \mathbb{E}\Bigg(\sum \Big[(x_i-\bar{x})y_i\Big]\Bigg) \\
&= \frac{1}{\sum x_i^2 - n\bar{x}^2} \times \sum \Bigg(\mathbb{E}\Big[(x_i-\bar{x})y_i\Big]\Bigg) \\
&= \frac{1}{\sum x_i^2 - n\bar{x}^2} \times \sum \Bigg((x_i-\bar{x})\mathbb{E}\big[y_i\big]\Bigg) \\
&= \frac{1}{\sum x_i^2 - n\bar{x}^2} \times \sum \Bigg((x_i-\bar{x})\big[\beta_0 + \beta_1(x_i)\big]\Bigg) \\
&= \frac{1}{\sum x_i^2 - n\bar{x}^2} \times \sum \Bigg(x_i\beta_0 - \bar{x}\beta_0 + \beta_1 x_i^2 - \beta_1 x_i \bar{x}\Bigg) \\
&= \frac{1}{\sum x_i^2 - n\bar{x}^2} \times \beta_0 \sum x_i - n\bar{x}\beta_0 + \beta_1 \sum x_i^2 - \beta_1 \bar{x} \sum x_i\Bigg) \\
&= \frac{1}{\sum x_i^2 - n\bar{x}^2} \times \beta_0 n\bar{x} - n\bar{x}\beta_0 + \beta_1\Bigg(\sum x_i^2 - \bar{x}n\bar{x}\Bigg) \\
&= \frac{1}{\sum x_i^2 - n\bar{x}^2} \times \beta_1\Bigg(\sum x_i^2 - n\bar{x}^2\Bigg) \\
&= \beta_1
\end{aligned}
$$

B_0 is Unbiased

We can also prove the unbiasedness of the B_0 estimator. To do this, we need to show that $\mathbb{E}(B_0) = \beta_0$. Again, we first start with the definition of the intercept estimator:

$$B_0 = \bar{y} - B_1(\bar{x})$$

We start with the regression equation from Assumption 1, $y_i = \beta_0 + \beta_1(x_i) + \epsilon_i$, and manipulate it to take the sum of both sides of the equation, and then divide by n.

$$
\begin{aligned}
\sum y_i &= \sum \Bigg(\beta_0 + \beta_1(x_i) + \epsilon_i\Bigg) \\
&= \sum \beta_0 + \sum \beta_1(x_i) + \sum \epsilon_i \\
&= n\beta_0 + \beta_1 \sum (x_i) + \sum \epsilon_i \\
\frac{\sum y_i}{n} &= \frac{n\beta_0}{n} + \frac{\beta_1 \sum (x_i)}{n} + \frac{\sum \epsilon_i}{n}
\end{aligned}
$$

This last expression we can re-write as,

$$\bar{y} = \beta_0 + \beta_1(\bar{x}) + \bar{\epsilon}$$

Now we will substitute this into $\bar{y}$ in the definition of B_0.

$$\begin{aligned} B_0 &= \bar{y} - B_1(\bar{x}) \\ &= \beta_0 + \beta_1(\bar{x}) + \bar{\epsilon} - B_1(\bar{x}) \\ &= \beta_0 + \bar{x}\Bigg(\beta_1 - B_1\Bigg) + \bar{\epsilon} \end{aligned}$$

Lastly, we take the expectation (conditional on X).

$$\begin{aligned} \mathbb{E}(B_0) &= \mathbb{E}\Bigg(\beta_0 + \bar{x}(\beta_1 - B_1) + \bar{\epsilon}\Bigg) \\ &= \mathbb{E}(\beta_0) + \mathbb{E}\Bigg(\bar{x}(\beta_1 - B_1)\Bigg) + \mathbb{E}(\bar{\epsilon}) \\ &= \beta_0 + \bar{x}\mathbb{E}(\beta_1 - B_1) + \mathbb{E}(\bar{\epsilon}) \\ &= \beta_0 + \bar{x}\Bigg(\mathbb{E}(\beta_1) - \mathbb{E}(B_1)\Bigg) + \mathbb{E}(\bar{\epsilon}) \\ &= \beta_0 + \bar{x}\Bigg(\beta_1 - \mathbb{E}(B_1)\Bigg) + \mathbb{E}(\bar{\epsilon}) \end{aligned}$$

Using Assumption #2, we find that $\mathbb{E}(\bar{\epsilon}) = 0$. We also just proved that $\mathbb{E}(B_1) = \beta_1$. So,

$$\begin{aligned} \mathbb{E}(B_0) &= \beta_0 + \bar{x}\Bigg(\beta_1 - \beta_1\Bigg) + 0 \\ &= \beta_0 + \bar{x}(0) \\ &= \beta_0 \end{aligned}$$