Proof of $SS_T = SS_R + SS_E$ In The Context of OLS

BIO210 Biostatistics

Extra reading material for Lecture 38

Xi Chen
School of Life Sciences
Southern University of Science and Technology

May 2022

During the lecture, we demonstrated that for each observation, the total deviation of y_i from its mean $\bar{y}$ consists of two parts: unexplained deviation due to error and deviation explained by the regression line. That is:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Once we collect the deviation for all observation and sum them up, we have:

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

We want to prove that $SS_T = SS_E + SS_R$.

Proof. We start with:

$$\begin{aligned}
SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\
&= \sum_{i=1}^{n} [(y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})] \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
&= SS_E + SS_R + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})
\end{aligned}$$

Now we only need to prove that $\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$. Expand the terms, we have:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(\beta_0 + \beta_1 x_i - \bar{y})$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}[(y_i - \beta_0 - \beta_1 x_i)(\beta_0 - \bar{y}) + (y_i - \beta_0 - \beta_1 x_i)\beta_1 x_i] \\
&= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(\beta_0 - \bar{y}) + \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\beta_1 x_i \\
&= (\beta_0 - \bar{y})\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) + \beta_1\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i
\end{aligned}
$$

Now, we are going to prove that both the brown and the purple terms are equal to 0.

Since we are doing OLS, the SE_{line} should take minimum value. By the definition of SE_{line}:

$$
SE_{line} = \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2
$$

Because SE_{line} is taking the minimum value. We should have:

$$
\frac{\partial SE_{line}}{\partial \beta_0} = 0\,, \text{and } \frac{\partial SE_{line}}{\partial \beta_1} = 0
$$

Now, let's first re-write SE_{line} using β_0 as the variable:

$$
\begin{aligned}
SE_{line} &= \sum_{i=1}^{n}[y_i^2 - 2y_i(\beta_0 + \beta_1 x_i) + (\beta_0 + \beta_1 x_i)^2] \\
&= \sum_{i=1}^{n}[y_i^2 - 2y_i\beta_0 - 2y_i\beta_1 x_i + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2] \\
&= \sum_{i=1}^{n}[\beta_0^2 + (2\beta_1 x_i - 2y_i)\beta_0 + (y_i^2 - 2y_i\beta_1 x_i + \beta_1^2 x_i^2)]
\end{aligned}
$$

Now we let $\dfrac{\partial SE_{line}}{\partial \beta_0} = 0$, we have:

$$
\frac{\partial SE_{line}}{\partial \beta_0} = \sum_{i=1}^{n}[2\beta_0 + (2\beta_1 x_i - 2y_i)] = 0
$$

Divide by 2 at both sides, we have:

$$\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - y_i) = 0 \Rightarrow \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

We have now proved that the brown term is 0. Similarly, re-write SE_{line} using β_1 as the variable:

$$SE_{line} = \sum_{i=1}^{n}[x_i^2\beta_1^2 + (2\beta_0 x_i - 2x_i y_i)\beta_1 + (y_i^2 - 2y_i\beta_0 + \beta_0^2)]$$

Now, we let $\frac{\partial SE_{line}}{\partial \beta_1} = 0$, we have:

$$\frac{\partial SE_{line}}{\partial \beta_1} = \sum_{i=1}^{n}(2x_i^2\beta_1 + 2\beta_0 x_i - 2x_i y_i) = 0$$

Divide by 2 at both sides, we have:

$$\sum_{i=1}^{n}(x_i^2\beta_1 + \beta_0 x_i - x_i y_i) = 0 \Rightarrow \sum_{i=1}^{n}(y_i - \beta_0 - x_i\beta_1)x_i = 0$$

Now we have proved the purple term is also 0. $\square$