

**SSE SST**

In English, *SSE*, or sum of squared errors (or error sum of squares) is the sum of squared residuals (errors). Remember, that for each observation *i*, the residual is simply the vertical distance (i.e., difference) in the observed value of *y*, which we denote by , and the value of *y* that falls on the estimated regression line, which we denote by . Recall further that ; that is, in order to obtain the value of , we need to know the values of , *x*, and .

SSE is the quantity that we aim to minimize in *ordinary least squares* regression.

The way to interpret SSE is the amount of variability in *y* that is *NOT* explained when accounting for *x* in the model.

In English, *SST*, or total sum of squares is the total variability in *y*. (In fact, if we were to delete the *SST* by (*n* – 1), we would obtain the variance of *y*.) That is, for each observation *i*, we’re simply calculating the squared deviation of that observation from the overall mean of *y*, which we denote by , and then summing those squared deviations across all observations *i*, without any regard to the value of *x*.

Two other terms need to be introduced here:

1. *SSR* (regression sum of squares), calculated as: *SSR*= *SST* – *SSE*. The higher the *SSR*, the higher the amount of variability in *y* that is explained when accounting for *x*.
2. Instead of presenting the raw *SSR*, which is not a standardized measure, statisticians prefer another measure, called *R2*, or the *coefficient of determination*. It is defined as the *proportion* of variance in *y* that is explained by the model (i.e., by the predictor *x*). Mathematically, *R2* = *SSR*/*SST* = 1 *– SSE*/*SST*. *R2* always ranges between 0 and 1.

**Question:** What does it mean if *SSE* = *SST*?

**Answer:** *x* and *y* are not strongly related, meaning that the amount of variation in *y* that is not explained by *x* is the same as the total amount of variation in *y*. That is, the inclusion of *x* doesn’t do much to explain the variation in *y*.

Recall that the regression line is an estimate of the true regression line, i.e., the line that shows the true average values of *y* for different values of *x*. However, if *SSE* = *SST*, it means that *y* and *x* are not related, and the line that shows the (estimated) average values of *y* for the different values of *x* is simply the horizontal line . That is, (the estimate of) the true average value of *y* doesn’t change with respect to *x*, and the slope coefficient , meaning that the estimated regression equation is reduced to , where , the *y*-intercept, is simply , the average value of *y* across all observations.

**Question:** What happens when we’re dealing with multiple regression?

**Answer:** In multiple regression with *k* predictors (*xi*…*xk*),

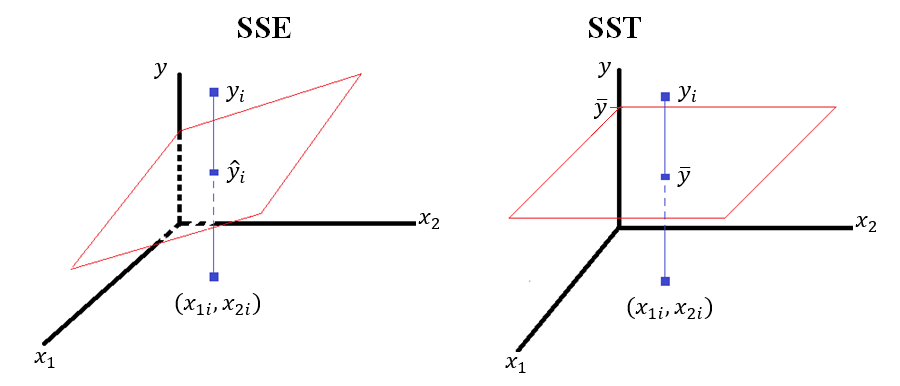
.

For instance, when there are only *k* = 2 predictors, the equation becomes

.

The equation for *SST* remains the same, regardless of the number of predictors:

The graph below shows the *SSE* and *SST* for a regression model with *k* = 2 predictors.

****