What is the concept of Sum of Squares and how is it broken down into SST, SSR, and SSE?

Authored by

stats writer

April 26, 2024

The concept of Sum of Squares (SS) refers to a statistical method used to measure the variation or discrepancy in a set of data points. It is commonly used in regression analysis to evaluate the relationship between variables. The SS is broken down into three components - the Total Sum of Squares (SST), the Regression Sum of Squares (SSR), and the Error Sum of Squares (SSE).

SST represents the total variation in the data and is calculated by subtracting the mean of the data from each data point, squaring the differences, and then summing them all together. SSR measures the variation that can be attributed to the regression model and is calculated by subtracting the predicted values from the mean of the data, squaring the differences, and then summing them. SSE represents the unexplained or random variation in the data and is calculated by subtracting the predicted values from the actual values, squaring the differences, and then summing them.

Overall, the concept of Sum of Squares allows for the decomposition of the total variation in the data into different components, providing a more comprehensive understanding of the relationship between variables and the accuracy of the regression model.

A Gentle Guide to Sum of Squares: SST, SSR, SSE

is used to find a line that best "fits" a dataset.

We often use three different sum of squares values to measure how well the regression line actually fits the data:

1. Sum of Squares Total (SST) - The sum of squared differences between individual data points (yi) and the mean of the response variable (y).

SST = Σ(yi - y)2

2. Sum of Squares Regression (SSR) - The sum of

squared differences between predicted data points (?i) and the mean of the response variable(y).

SSR = Σ(?i - y)2

3. Sum of Squares Error (SSE) - The sum of squared differences between predicted data points (?i) and observed data points (yi).

SSE = Σ(?i - yi)2

The following relationship exists between these three measures:

SST = SSR + SSE

Thus, if we know two of these measures then we can use some simple algebra to calculate the third.

SSR, SST & R-Squared

, sometimes referred to as the coefficient of determination, is a measure of how well a linear regression model fits a dataset. It represents the proportion of the variance in the that can be explained by the predictor variable.

The value for R-squared can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all. A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable.

Using SSR and SST, we can calculate R-squared as:

R-squared = SSR / SST

For example, if the SSR for a given regression model is 137.5 and SST is 156 then we would calculate R-squared as:

This tells us that 88.14% of the variation in the response variable can be explained by the predictor variable.

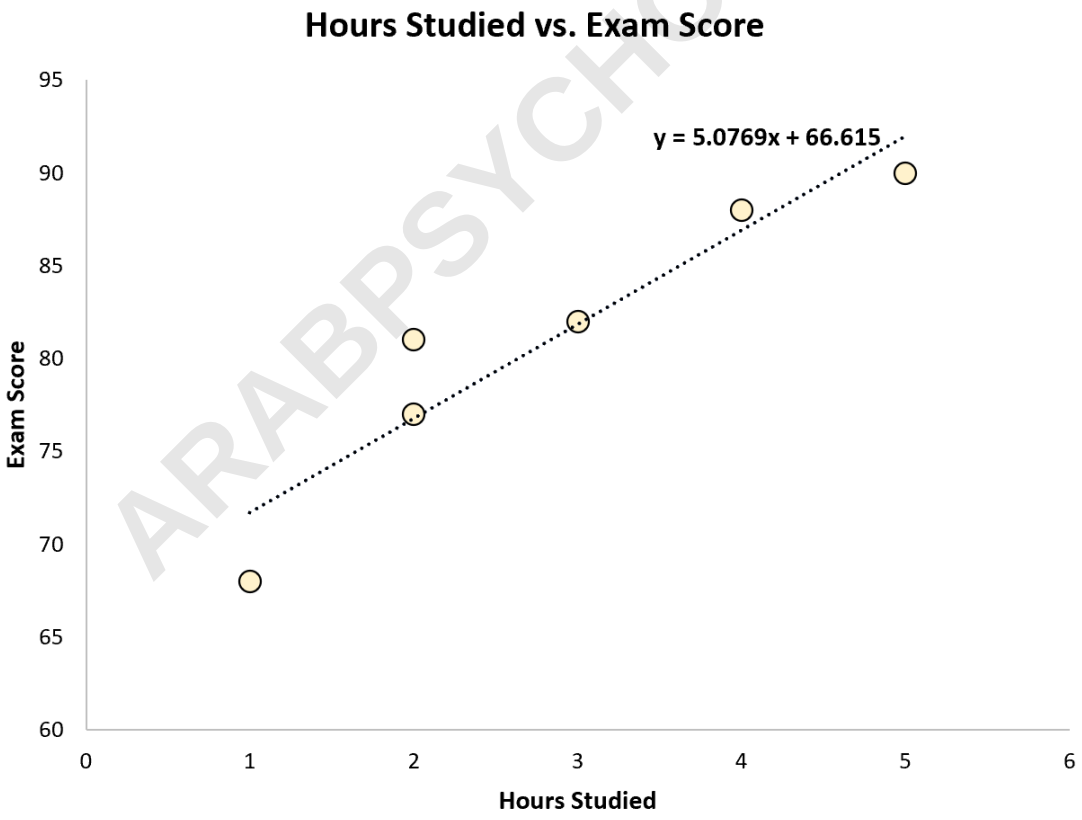
Calculate SST, SSR, SSE: Step-by-Step Example

Suppose we have the following dataset that shows the number of hours studied by six different students along with their final exam scores:

Hours Studied	Exam Score
1	68
2	77
2	81
3	82
4	88
5	90

Using some statistical software (like , ,) or even , we can find that the line of best fit is:

Score = 66.615 + 5.0769*(Hours)



Once we know the line of best fit equation, we can use

the following steps to calculate SST, SSR, and SSE:

Step 1: Calculate the mean of the response variable.

The mean of the response variable (y) turns out to be 81.

Hours Studied	Exam Score	$\bar{y}$
1	68	81
2	77	81
2	81	81
3	82	81
4	88	81
5	90	81

Step 2: Calculate the predicted value for each observation.

Next, we can use the line of best fit equation to calculate the predicted exam score () for each student.

For example, the predicted exam score for the student who studied one hours is:

Score = 66.615 + 5.0769*(1) = 71.69.

We can use the same approach to find the predicted

score for each student:

Hours Studied	Exam Score	$\bar{y}$	$\hat{y}$
1	68	81	71.69
2	77	81	76.77
2	81	81	76.77
3	82	81	81.85
4	88	81	86.92
5	90	81	92.00

Step 3: Calculate the sum of squares total (SST).

Next, we can calculate the sum of squares total.

For example, the sum of squares total for the first student is:

(yi - y)2 = (68 - 81)2 = 169.

We can use the same approach to find the sum of squares total for each student:

Hours Studied	Exam Score	$\bar{y}$	$\hat{y}$	$(y_i - \bar{y})^2$
1	68	81	71.69	169
2	77	81	76.77	16
2	81	81	76.77	0
3	82	81	81.85	1
4	88	81	86.92	49
5	90	81	92.00	81
				316
				SST

The sum of squares total turns out to be 316.

Step 4: Calculate the sum of squares regression (SSR).

Next, we can calculate the sum of squares regression.

For example, the sum of squares regression for the first student is:

(?i - y)2 = (71.69 - 81)2 = 86.64.

We can use the same approach to find the sum of squares regression for each student:

Hours Studied	Exam Score	$\bar{y}$	$\hat{y}$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$
1	68	81	71.69	169	86.64
2	77	81	76.77	16	17.90
2	81	81	76.77	0	17.90
3	82	81	81.85	1	0.72
4	88	81	86.92	49	35.08
5	90	81	92.00	81	120.99
				316	279.23
				SST	**SSR**

The sum of squares regression turns out to be 279.23.

Step 5: Calculate the sum of squares error (SSE).

Next, we can calculate the sum of squares error.

For example, the sum of squares error for the first student is:

(?i - yi)2 = (71.69 - 68)2 = 13.63.

We can use the same approach to find the sum of squares error for each student:

Hours Studied	Exam Score	$\bar{y}$	$\hat{y}$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$	$(\hat{y}_i - y_i)^2$
1	68	81	71.69	169	86.64	13.63
2	77	81	76.77	16	17.90	0.05
2	81	81	76.77	0	17.90	17.90
3	82	81	81.85	1	0.72	0.02
4	88	81	86.92	49	35.08	1.16
5	90	81	92.00	81	120.99	4.00
				316	279.23	36.77
				SST	**SSR**	**SSE**

We can verify that SST = SSR + SSE

SST = SSR + SSE316 = 279.23 + 36.77

We can also calculate the R-squared of the regression model by using the following equation:

R-squared = SSR / SSTR-squared = 279.23 / 316R-squared = 0.8836

This tells us that 88.36% of the variation in exam scores can be explained by the number of hours studied.

You can use the following calculators to automatically calculate SST, SSR, and SSE for any simple linear regression line: