What is Sum of Squares: SST, SSR, SSE?

Authored by

stats writer

December 9, 2025

RECOMMENDED CITATION

stats writer (2025). *What is Sum of Squares: SST, SSR, SSE?*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=106737

ARABPSYCHOLOGY.COM

The Sum of Squares (SS) is a fundamental measure of variation used extensively in statistical analysis, particularly within regression modeling and ANOVA. It quantifies the dispersion of data points around a mean value within a given statistical population or sample. Understanding SS is crucial because it forms the basis for determining how well a predictive model fits the observed data, essentially partitioning the total variation into components that are explained by the model and those that remain unexplained. This measure is indispensable when assessing model performance and significance in hypothesis testing.

The overall variability described by the Sum of Squares is systematically decomposed into three essential, additive components: the **Sum of Squares Total (SST)**, the **Sum of Squares Regression (SSR)**, and the **Sum of Squares Error (SSE)**. SST represents the total inherent variability present in the response variable. SSR quantifies the portion of that total variability that is successfully captured or explained by the regression model--the structured relationship identified by the line of best fit. Conversely, SSE measures the residual, unexplained variability--the average squared distance between the observed data points and the model's predictions. The interplay between these three metrics provides a comprehensive and mathematically robust view of the model's effectiveness and reliability.

The Goal of Regression and Sum of Squares Components

In the context of linear regression, the primary goal is to find a line that best "fits" a dataset by minimizing the overall prediction error, typically achieved through the Ordinary Least Squares method. This line of best fit allows us to model the expected value of a dependent response variable based on changes in one or more independent predictor variables. The quality of this fit is directly measured by analyzing how the total variation in the data is distributed among the explained and unexplained components.

We utilize the three specific **sum of squares** values--SST, SSR, and SSE--to meticulously measure and evaluate the quality of the fit provided by the established regression line. Each component isolates a distinct source of variation within the response variable, allowing statisticians to precisely attribute variability to either the model structure or random noise.

1. Sum of Squares Total (SST) - The Sum of Squares Total measures the overall variability in the dependent variable (y). It is calculated as the sum of squared differences between each individual observed data point (yi) and the grand mean of the response variable (y). SST establishes the total amount of variation present in the data before the model is applied, providing the baseline against which model performance is judged.

$SST = \Sigma(yi - y)2$

ARABPSYCHOLOGY.COM

2. Sum of Squares Regression (SSR) - Also known as the Sum of Squares Model (SSM), the Sum of Squares Regression quantifies the variation in the dependent variable that is successfully explained by the predictive relationship established in the regression model. It is calculated by summing the squared differences between the predicted data points (?i) generated by the model and the mean of the response variable (y). A higher SSR indicates that the model successfully captures and accounts for a large proportion of the total variability.

SSR = Σ(?i - y)2

3. Sum of Squares Error (SSE) - The Sum of Squares Error, sometimes referred to as Sum of Squares Residual (SSR), represents the unexplained variation in the dependent variable. This error captures the discrepancies, or residuals, between the actual observed data points (yi) and the data points predicted by the model (?i). It is the quantity that the ordinary least squares method actively seeks to minimize, as it reflects the remaining noise, sampling variability, or random error not accounted for by the predictor variables.

SSE = Σ(?i - yi)2

The Fundamental Identity: SST = SSR + SSE

The relationship among these three measures is a fundamental algebraic identity essential for understanding the partition of variance in a regression setting. This identity states unequivocally that the total variability inherent in the response variable (SST) must equal the sum of the variability explained by the model (SSR) and the variability left unexplained (SSE). This concept is the cornerstone of the Analysis of Variance (ANOVA) applied to regression.

The mathematical expression of this partition is:

SST = SSR + SSE

This relationship is highly practical. If we know any two of these measures, the third can be readily calculated using simple algebra. For instance, if a researcher knows the total variation in their dataset (SST) and the error remaining after fitting the model (SSE), they can immediately determine the explanatory power of their model (SSR) without needing to perform the SSR calculations separately. This provides a crucial internal check on the integrity of the statistical analysis.

R-Squared: Quantifying Model Explanatory Power

The ratio derived from the explained variation (SSR) and the total variation (SST) forms the basis for calculating the **Coefficient of Determination**, which is universally known as R-squared (R^2). R-squared is a pivotal metric used to assess the goodness of fit of a linear regression model to a

ARABPSYCHOLOGY.COM

dataset, offering an intuitive interpretation of model performance.

Conceptually, R-squared represents the proportion of the total variance in the dependent variable that can be predicted or accounted for by the independent variable(s). If the regression model perfectly predicts all observed values, meaning all data points lie exactly on the line of best fit, then the SSE would be zero and R-squared would equal one. Conversely, if the model explains none of the variability beyond what the mean itself explains, SSR would be zero, resulting in an R-squared of zero.

The value for R-squared is always constrained to the interval . A higher value, closer to 1, suggests a strong fit, indicating that the response variable is largely explained by the predictor variables, with minimal residual error. Using SSR and SST, we formally calculate R-squared as:

R-squared = SSR / SST

Consider an example where the SSR for a given regression model is calculated as 137.5 and the SST is 156. The resulting R-squared calculation would be: 137.5 / 156 ≈ 0.8814. This result provides a clear, actionable insight, telling us that 88.14% of the variation in the response variable can be systematically explained by the predictor variable included in the model.

Step-by-Step Example: Calculating SST, SSR, and SSE

To demonstrate the practical application of these concepts, we will walk through a detailed example focusing on a simple scenario. Suppose we have the following dataset that tracks the number of hours studied by six different students along with their resulting final exam scores:

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

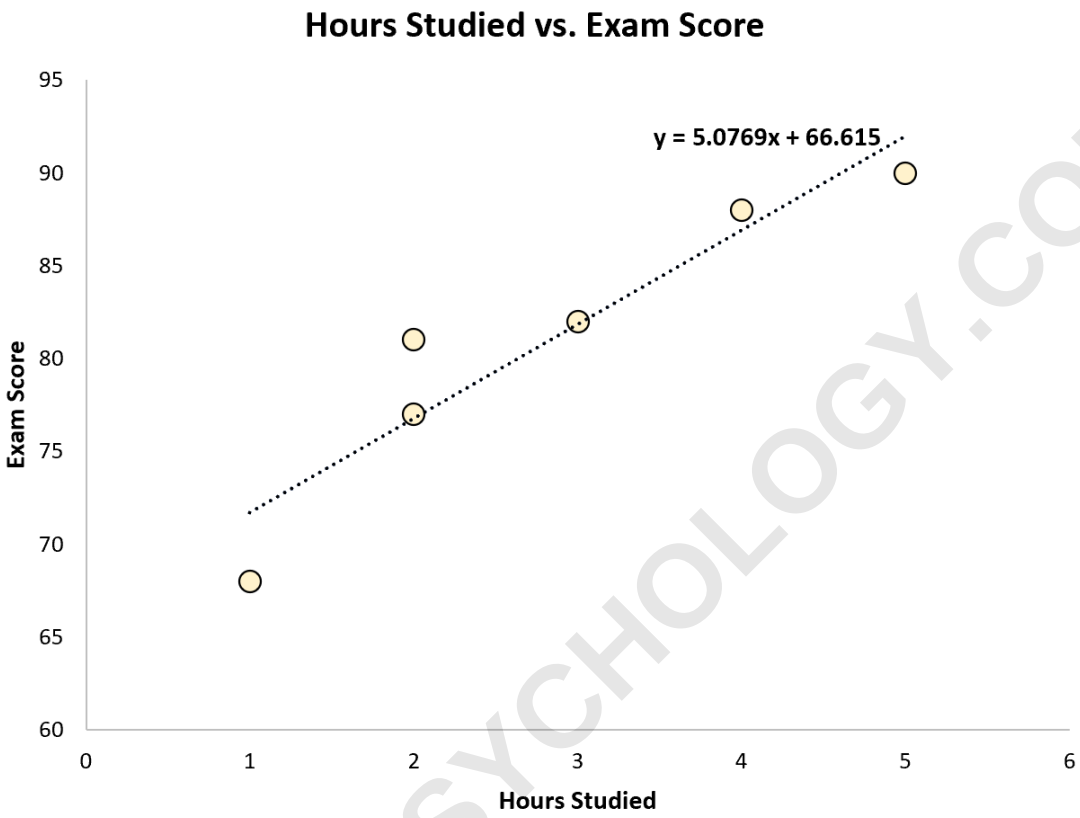
We aim to determine how well 'Hours Studied' predicts 'Exam Score'. Using standard statistical software, such as JMP, the R programming language, or even complex functionalities within spreadsheet software, we first establish the line of best fit using the least squares method.

The derived equation for the line of best fit that minimizes the sum of squared errors is:

ARABPSYCHOLOGY.COM

Score = 66.615 + 5.0769*(Hours)

This equation allows us to calculate the predicted score (?i) for any given input of study hours. This regression line is visualized below, illustrating how the model attempts to capture the central tendency of the data points.



ARABPSYCHOLOGY.COM

With the established model, we can now proceed to calculate the three essential sum of squares components (SST, SSR, and SSE) sequentially, relying on the mean of the response variable and the predicted values.

Calculating the Core Components of Variability

The calculation process begins by establishing the statistical baseline--the mean of the response variable. This figure represents the best possible prediction if no explanatory variables were considered.

Step 1: Calculate the mean of the response variable.

The mean of the response variable (y) is determined by averaging the six observed exam scores. This calculation yields a grand mean of **81**. This average value acts as the crucial reference point for both the SST and SSR calculations.

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 use the line of best fit equation (Score = 66.615 + 5.0769 * Hours) to calculate the predicted exam score (?i) for each student based on the hours they studied. For instance, the predicted score for the student who studied for one hour is 71.69. This produces a new set of data points that lie exactly on the regression line.

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).

SST is calculated by summing the squared differences between each observed score (yi) and the mean score (y). This quantifies the overall inherent spread of the scores. For the first student, the calculation is: (yi - y)2 = (68 - 81)2 = 169. Summing these squared deviations for all students:

ARABPSYCHOLOGY.COM

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 total sum of squares (SST) for this example dataset is **316**.

Step 4: Calculate the Sum of Squares Regression (SSR).

SSR measures the explained variation by calculating the squared difference between the predicted score (?i) and the mean score (y). This shows the extent to which the model improves upon the mean as a predictor. For the first student, the SSR component is: (?i - y)2 = (71.69 - 81)2 = **86.64**.

We repeat this process for all students to find the aggregate explained variation:

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 total sum of squares regression (SSR) is calculated as **279.23**.

Step 5: Calculate the Sum of Squares Error (SSE).

Finally, SSE measures the unexplained variation, or error. It is the sum of the squared differences between the predicted score (?i) and the actual observed score (yi). For the first student, the SSE

ARABPSYCHOLOGY.COM

component is: (?i - yi)2 = (71.69 - 68)2 = **13.63**.

By summing these residual squared values across the dataset, we quantify the total noise:

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**

The total sum of squares error (SSE) is calculated as **36.77**.

Verification and Final Model Assessment

After calculating the three components, we verify the fundamental partitioning identity to ensure the calculations are sound: SST must equal the sum of SSR and SSE.

SST = SSR + SSE
316 = 279.23 + 36.77
316 = 316.00 (Verification successful)

We can now calculate the R-squared value for this specific regression model using the derived sum of squares figures:

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

This result confirms that **88.36%** of the variation in the final exam scores can be explained by the linear relationship with the number of hours studied. This suggests a highly effective and strong predictive model.

Streamlining Calculations with Statistical Tools

While the manual, step-by-step approach detailed above is essential for conceptual understanding,

ARABPSYCHOLOGY.COM

practical statistical analysis involving large datasets necessitates the use of specialized software. These environments, whether open-source platforms like R or commercial software, automatically handle the computationally intensive process of least squares estimation and the subsequent derivation of the sum of squares components.

Statistical software rapidly generates the Analysis of Variance (ANOVA) table, which summarizes the calculated SST, SSR, and SSE, along with other critical diagnostic statistics. Utilizing these tools ensures precision, handles complex models efficiently, and allows analysts to dedicate their time to interpreting the derived relationships rather than manual data processing. For smaller or exploratory analyses, functions within advanced spreadsheet programs like Excel can also be configured to automate these sum of squares calculations.

ARABPSYCHOLOGY.COM