

How to Calculate SST, SSR, and SSE in Excel?

Authored by
stats writer

December 9, 2025

RECOMMENDED CITATION

stats writer (2025). # How to Calculate SST, SSR, and SSE in Excel?. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=106730>

In the field of statistical modeling, particularly regression analysis, measuring how well a model fits the observed data is crucial for validation and interpretation. Three foundational metrics used to quantify this fit are the Sum of Squares Total (SST), Sum of Squares Regression (SSR), and Sum of Squares Error (SSE). These statistics are essential components of the overall analysis of variance and provide profound insight into the variation captured by the independent variables versus the unexplained variation.

To accurately calculate these sums of squares metrics in Excel, we typically leverage the capabilities of the Data Analysis ToolPak, which automatically computes these values as part of the standard regression output. However, understanding the underlying calculations requires recognizing that we are summing the squares of differences: differences between actual values and the mean (for SST), differences between predicted values and the mean (for SSR), and differences between actual values and predicted values (for SSE). Mastery of these concepts is vital for any rigorous interpretation of a statistical model.

Understanding the Significance of Sums of Squares

The concept of "Sum of Squares" is fundamental in statistics, serving as a measure of the total deviation or dispersion of data points. When applying this concept to a regression model, these three components--SST, SSR, and SSE--help partition the total variability observed in the response variable (Y). This partitioning allows statisticians to determine how much of the variability is successfully explained by the model (SSR) and how much remains unexplained (SSE).

Understanding the contribution of each sum of squares is the first step toward evaluating model efficacy. SST represents the baseline variability that exists before applying any model; it measures the scatter of the data points around their mean. When we introduce a regression model, we attempt to reduce this total variability, and the extent to which we succeed is quantified by the SSR. Conversely, the SSE captures the noise or random error--the residual variability that the model failed to capture.

These values are intrinsically linked, forming the backbone of the ANOVA (Analysis of Variance) table generated during regression analysis. We use these metrics to assess the strength of the linear relationship between variables and, critically, to calculate the coefficient of determination, or R-squared, which is perhaps the most common metric for model fit.

Defining the Core Metrics: SST, SSR, and SSE

To ensure a clear foundation, it is important to formally define each of these three measures of variation. Each formula describes a different aspect of the data's deviation and their sum provides the complete picture of variability within the dependent variable.

The three key values used to measure how effectively a statistical model fits a dataset are:

Sum of Squares Total (SST) - This metric represents the total variation in the dependent variable (y). It is calculated as the sum of squared differences between each individual data point (y_i) and the mean of the response variable (\bar{y}). SST provides the total baseline variability that the model attempts to explain.

SST Formula: $\text{SST} = \sum (y_i - \bar{y})^2$

Sum of Squares Regression (SSR) - Also known as the Sum of Squares Explained, this measures the variation explained by the regression line or the model. It is the sum of squared differences between the predicted data points (\hat{y}_i) and the mean of the response variable (\bar{y}). A higher SSR relative to SST indicates a better model fit.

SSR Formula: $\text{SSR} = \sum (\hat{y}_i - \bar{y})^2$

Sum of Squares Error (SSE) - Also referred to as the Residual Sum of Squares, this measures the unexplained variation or the residual error. It is the sum of squared differences between the observed data points (y_i) and the predicted data points (\hat{y}_i). This represents the amount of variation left over after the model has done its best to fit the data. Ideally, we want a low SSE.

SSE Formula: $\text{SSE} = \sum (y_i - \hat{y}_i)^2$

These definitions confirm the role of these statistics in decomposing the overall variation. Note that while the original content had SSE as the difference between predicted and observed ($\hat{y}_i - y_i$), the standard definition uses $y_i - \hat{y}_i$. Since we are squaring the difference, the order of subtraction does not affect the final positive result, but consistency in definition is important for clarity.

The Fundamental Identity: $\text{SST} = \text{SSR} + \text{SSE}$

The relationship between these three statistics is not arbitrary; they are linked by a crucial identity: the total variation in the response variable (SST) is exactly equal to the variation explained by the regression model (SSR) plus the unexplained variation (SSE). This identity forms the cornerstone of variance decomposition in linear regression and is the basis for calculating the coefficient of determination.

This fundamental identity ensures that every piece of variation in the dependent variable is accounted for, either by the predictive power of the independent variables within the model or by random error. When performing regression analysis in software like Excel, the results in the ANOVA table are structured specifically to confirm this relationship, providing confidence in the

validity of the computed statistics.

We will now proceed with a practical, step-by-step example demonstrating how to obtain these essential metrics using Excel's powerful regression capabilities, focusing specifically on the Data Analysis ToolPak.

Step 1: Preparing Your Dataset in Excel

The first prerequisite for calculating the sums of squares is having a well-structured dataset. For this demonstration, we will use a hypothetical dataset tracking the relationship between the number of hours a student studied (the independent variable, X) and their corresponding exam score (the dependent variable, Y). This simple linear relationship provides an excellent platform for demonstrating regression analysis.

Ensure your data is organized into clearly labeled columns, where the independent variables (predictors) are distinct from the dependent variable (response). Proper organization minimizes errors when selecting the input ranges for the regression tool.

Below is the dataset structure showing the hours studied and the exam scores for 20 students:

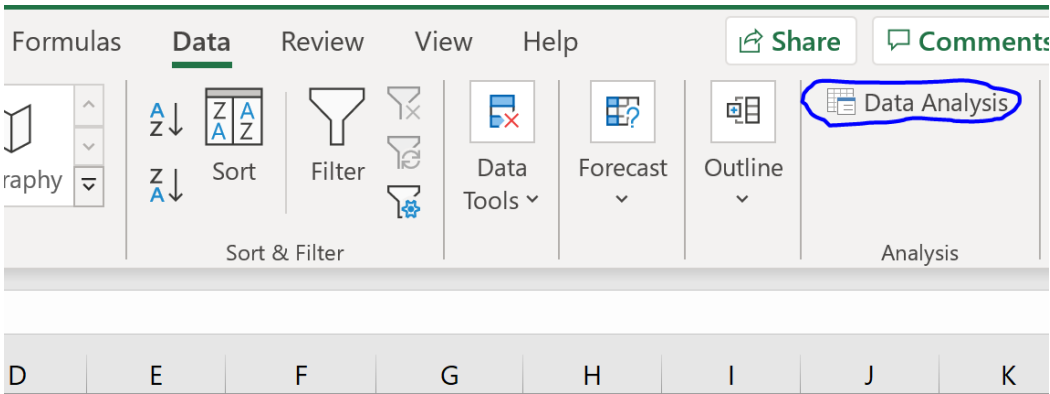
	A	B	C	D	E	F	G
1	Hours	Score					
2	1	68					
3	1	76					
4	1	74					
5	2	80					
6	2	76					
7	2	78					
8	2	81					
9	2	84					
10	3	86					
11	3	83					
12	3	88					
13	4	85					
14	4	89					
15	4	94					
16	5	93					
17	5	94					
18	6	96					
19	7	89					
20	7	92					
21	8	97					
22							
23							
24							
25							
26							
27							
28							

Once the data is accurately entered and reviewed, we can move on to executing the regression procedure, which will automatically handle the complex calculations of means, deviations, and squared sums necessary to derive SST, SSR, and SSE.

Step 2: Executing Regression Analysis Using Data Analysis ToolPak

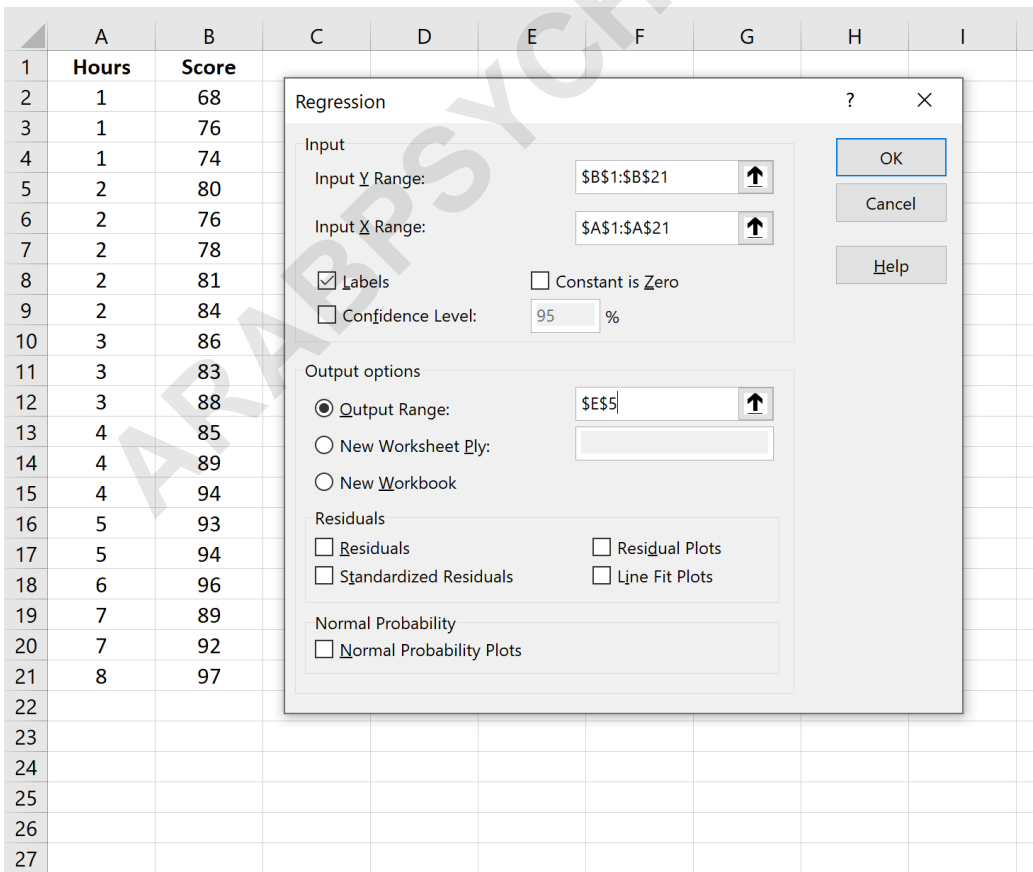
To fit a regression model in Excel and generate the required sums of squares, you must utilize the Data Analysis ToolPak. If this option is not visible, it must first be enabled via Excel's Add-Ins menu (File > Options > Add-ins > Manage Excel Add-ins > Go > Check Analysis ToolPak).

Navigate to the **Data** tab on the Excel ribbon, and then click on the **Data Analysis** button, typically located on the far right. This action opens a dialog box listing various statistical procedures. Scroll down, select **Regression**, and click OK.



The subsequent Regression dialog box requires specific input ranges. The **Input Y Range** should contain the exam scores (the dependent variable), and the **Input X Range** should contain the hours studied (the independent variable). It is generally advisable to check the **Labels** box if your selected range includes column headings, ensuring that Excel correctly labels the output.

Furthermore, define an **Output Range** where the results of the analysis will be displayed. This range should ideally be in a clear, empty section of the worksheet to avoid overwriting existing data. After specifying these parameters, click OK to run the regression analysis.



After clicking OK, Excel processes the data and generates a comprehensive regression report, which includes multiple tables detailing the coefficients, statistics, and, most importantly for our current purpose, the Analysis of Variance (ANOVA) table.

E	F	G	H	I	J
SUMMARY OUTPUT					
<i>Regression Statistics</i>					
Multiple R	0.857224				
R Square	0.734832				
Adjusted R Square	0.720101				
Standard Error	4.288712				
Observations	20				
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	917.4751	917.4751	49.88162	1.37823E-06
Residual	18	331.0749	18.39305		
Total	19	1248.55			
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>
Intercept	73.44585	1.914653	38.35988	1.03E-18	69.42331605
Hours	3.251152	0.460328	7.062692	1.38E-06	2.284039699

Step 3: Interpreting the ANOVA Table Output

The ANOVA table is the section of the regression output that contains the calculated Sums of Squares. This table systematically breaks down the total variability, allowing for immediate identification of SST, SSR, and SSE. The three sums of squares metrics are found specifically in the column labeled **SS**.

In the ANOVA table, the row labeled "Regression" corresponds to the Sum of Squares Regression (SSR), representing the explained variation. The row labeled "Residual" corresponds to the Sum of Squares Error (SSE), representing the unexplained variation. Finally, the row labeled "Total" corresponds to the Sum of Squares Total (SST), which is the sum of the other two components.

E	F	G	H	I	J
SUMMARY OUTPUT					
<i>Regression Statistics</i>					
Multiple R	0.857224				
R Square	0.734832				
Adjusted R Square	0.720101				
Standard Error	4.288712				
Observations	20				
<i>ANOVA</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	917.4751	917.4751	49.88162	1.37823E-06
Residual	18	331.0749	18.39305		
Total	19	1248.55			
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>
Intercept	73.44585	1.914653	38.35988	1.03E-18	69.42331605
Hours	3.251152	0.460328	7.062692	1.38E-06	2.284039699

For the example analysis conducted, the metrics are determined to be:

Sum of Squares Total (SST): 1248.55

Sum of Squares Regression (SSR): 917.4751

Sum of Squares Error (SSE): 331.0749

We can easily verify the fundamental identity by summing SSR and SSE:

$$SST = SSR + SSE$$

$$1248.55 = 917.4751 + 331.0749$$

$$1248.55 = 1248.55 \text{ (Verification successful)}$$

Calculating and Understanding R-squared

While SST, SSR, and SSE provide the raw components of variability, the most common metric derived from them is the coefficient of determination, or R-squared (R^2). This statistic provides a readily interpretable measure of model fit, indicating the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

The calculation of R-squared relies directly on the relationship between the explained variation (SSR) and the total variation (SST). The formula is defined as:

$$R\text{-squared} = SSR / SST$$

Using the results from our Excel regression output:

$$R\text{-squared} = 917.4751 / 1248.55$$

$$R\text{-squared} \approx 0.7348$$

This result implies that **73.48%** of the total variation observed in the students' exam scores can be successfully explained by the variation in the number of hours they studied. The remaining 26.52% is attributed to residual variation or error, suggesting that while the relationship is strong, other unmeasured factors (like study quality, prior knowledge, or test anxiety) contribute to the overall score variability.

Mastering the calculation and interpretation of SST, SSR, and SSE is crucial for moving beyond basic model fitting and into advanced model evaluation, ensuring that any derived conclusions are statistically sound and reliable.