How to calculate R-Squared in Excel (With Examples)

Authored by

stats writer

December 24, 2025

The measurement of model fit is fundamental in statistics and data analysis. Among the most critical metrics for evaluating the strength of a linear relationship between variables is R-Squared, also known as the Coefficient of Determination. This powerful statistical measure quantifies the degree to which the variance in a dependent variable can be predicted from the independent variable(s) within a regression model. When performing quantitative analysis, especially in business, academic research, or economics, understanding the exact percentage of variation explained by your model is crucial for drawing accurate conclusions.

While R-Squared is conceptually straightforward--it is calculated by simply squaring the correlation coefficient--its application and interpretation require careful attention to detail. Fortunately, Microsoft Excel provides specialized functions and tools that make calculating and verifying this metric highly efficient. This guide is designed to transform complex statistical theory into clear, actionable steps, showing you precisely how to leverage Excel's capabilities, particularly the dedicated RSQ function, to calculate R-Squared accurately.

This tutorial provides a comprehensive walkthrough, moving beyond simple formula entry to offer context on interpretation and comparison with alternative regression analysis methods available within the software. By the end of this expert guide, you will be equipped to confidently assess the goodness-of-fit of your own data models using **Microsoft Excel**.

Defining the Coefficient of Determination (R-Squared)

R-squared, often formally denoted as R2, is fundamentally a goodness-of-fit measure for linear regression models. It helps analysts determine how well the regression line approximates the real data points. In the simplest terms, R-Squared provides the percentage of the response variable variation that is explained by a linear model built using one or more predictor variables. If you are modeling the relationship between two variables, the R-Squared indicates how much of the variation in the dependent variable (Y) is shared with or explained by changes in the independent variable (X).

In technical terms, R-Squared is the proportion of the total variance in the dependent variable that is predictable from the independent variable. Understanding this concept is crucial because a high R-Squared value suggests that the model is a strong predictor of the outcome, meaning that fluctuations in the predictor variable directly correspond to the observed fluctuations in the response variable. Conversely, a low R-Squared value implies that other factors, not included in the model, are responsible for the majority of the variance observed in the dependent variable.

It is important to note the statistical distinction: in a simple linear regression involving only one predictor variable, R-Squared is mathematically equivalent to the square of the Pearson correlation coefficient (r). This means that calculating R-Squared essentially measures the linear correlation

strength between your dataset's two variables. However, for multiple linear regression models, the calculation relies on comparing the sum of squares of the residuals (error) against the total sum of squares (variance).

Interpreting the R-Squared Value Range

The value for R2 is standardized, meaning it is always constrained between 0 and 1, or 0% and 100%. These boundaries provide a clear framework for interpretation, allowing analysts to quickly gauge the explanatory power of their statistical model. While context is always vital, the proximity of the R-Squared value to 1 determines the level of confidence one can place in the model's ability to represent the underlying data relationship.

We can break down the significance of these boundary values to provide clarity on model fit:

A value approaching 0 indicates that the predictor variable provides virtually no explanatory power for the response variable. In this scenario, the regression line does not fit the data any better than a simple horizontal line drawn through the mean of the response variable. This suggests a weak or non-existent linear relationship.
A value approaching 1 indicates that the response variable can be almost perfectly explained by the predictor variable(s). This signifies an excellent fit where nearly all data points fall exactly on the regression line, meaning the model can account for almost all the observed variability without error.

While a higher R-Squared is generally desirable, it is critical to remember that context matters. In fields dealing with human behavior or complex natural systems (like ecology or finance), an R-Squared of 0.50 might be considered very strong due to the inherent noise and variability in the data. Conversely, in physics or engineering, where relationships are often deterministic, anything less than 0.90 might indicate a fundamental flaw in the experimental setup or data collection process. Analysts must use domain knowledge alongside statistical rigor when interpreting this metric.

Calculating R-Squared Directly Using the RSQ() Function in Excel

For users analyzing simple linear relationships in Excel, the most efficient method for determining R-Squared is by utilizing the built-in **RSQ()** function. This dedicated function is designed specifically to compute the square of the Pearson product-moment correlation coefficient based on known data points, saving the user the step of manually calculating the correlation (r) and then squaring the result. The function is robust and provides immediate feedback on the strength of the linear relationship being analyzed.

The structure of the RSQ function is concise and requires only two arguments: the range of the

dependent variable values and the range of the independent variable values. The syntax must be followed precisely for Excel to return a valid statistical result.

The required syntax for the function is:

=RSQ(known_ys, known_xs)

Where:

known_ys: This is the array or range of data points representing the dependent variable (Y). This is the variable whose variance you are trying to explain.
known_xs: This is the array or range of data points representing the independent variable (X). This is the predictor variable used to explain the variance in Y.

It is absolutely critical that both the known_ys and known_xs arrays have the same number of data points. If the arrays are of unequal length or contain non-numeric data, the **RSQ()** function will return an error, typically the #N/A or #VALUE! error, signaling an invalid input structure for the statistical calculation.

Step-by-Step Example: Predicting Student Performance

To illustrate the practical application of the **RSQ()** function, consider a hypothetical dataset compiled for 20 students. We are attempting to model the relationship between the time a student spends studying (the predictor variable, X) and the final exam score they receive (the response variable, Y). Our goal is to determine how much of the variation in scores can be explained solely by the variation in study hours.

Suppose we have the following sample data organized in two columns in an Excel sheet:

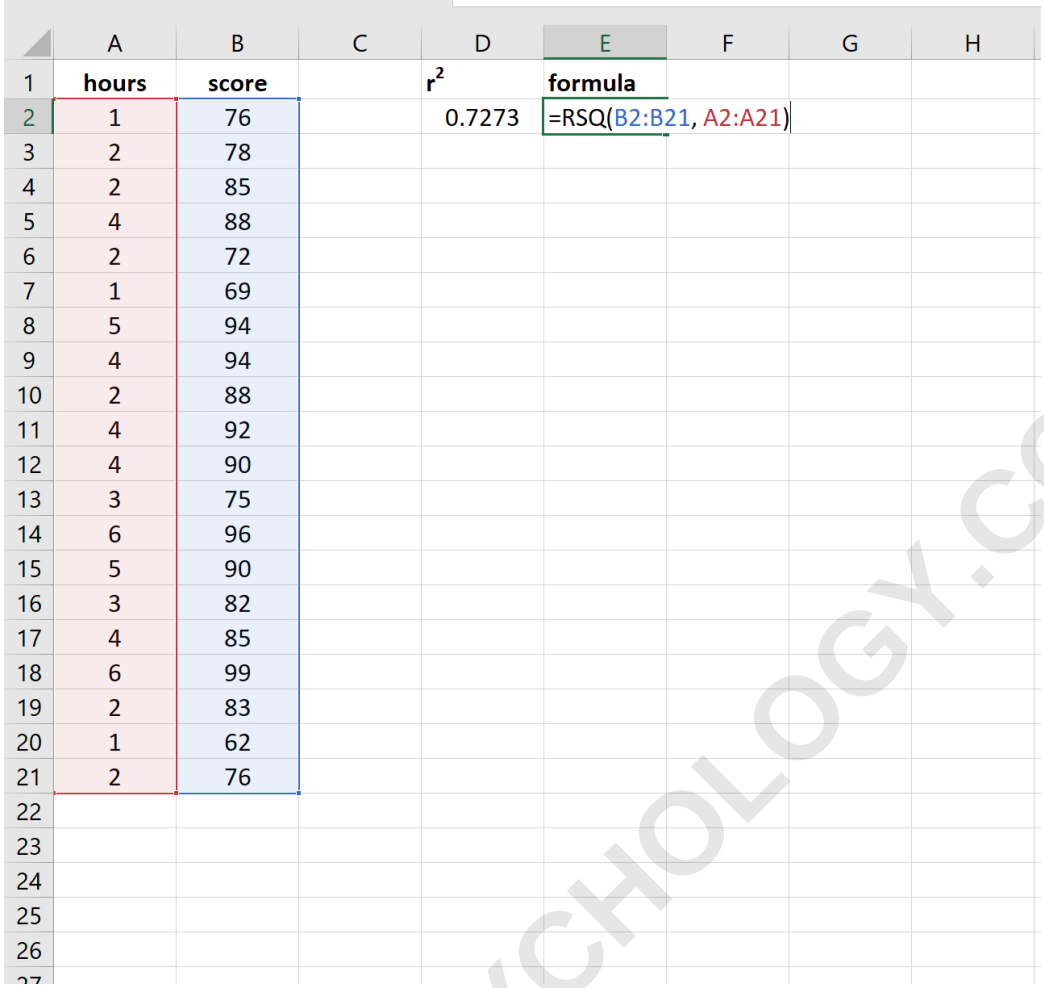
	A	B	C	D	E	F	G
1	**hours**	**score**					
2	1	76					
3	2	78					
4	2	85					
5	4	88					
6	2	72					
7	1	69					
8	5	94					
9	4	94					
10	2	88					
11	4	92					
12	4	90					
13	3	75					
14	6	96					
15	5	90					
16	3	82					
17	4	85					
18	6	99					
19	2	83					
20	1	62					
21	2	76					
22							
23							
24							
25							

The first step involves identifying the appropriate cell ranges for the formula. Assuming the 'Hours Studied' data is in column A (A2:A21) and the 'Exam Score' data is in column B (B2:B21), we will set up the RSQ function in an empty cell, such as C2, to perform the calculation. This simple linear regression model uses "hours" as the predictor variable (X) and "score" as the response variable (Y).

The formula entered into the chosen cell will be:

=RSQ(B2:B21, A2:A21)

Notice the arrangement: the known Ys (Scores) are listed first, followed by the known Xs (Hours). Executing this formula provides the R-Squared value directly, quantifying the explanatory power of the model we have constructed.



	A	B	C	D	E	F	G	H
1	hours	score		r^2	formula			
2	1	76		0.7273	=RSQ(B2:B21, A2:A21)			
3	2	78						
4	2	85						
5	4	88						
6	2	72						
7	1	69						
8	5	94						
9	4	94						
10	2	88						
11	4	92						
12	4	90						
13	3	75						
14	6	96						
15	5	90						
16	3	82						
17	4	85						
18	6	99						
19	2	83						
20	1	62						
21	2	76						
22								
23								
24								
25								
26								

Upon calculation, the result obtained for this specific dataset is **0.7273** (or 72.73% when expressed as a percentage). This figure holds significant interpretive power regarding our student performance model. Specifically, it means that **72.73%** of the total variation observed in the students' exam scores can be statistically accounted for or explained by the number of hours they dedicated to studying. The remaining variation (approximately 27.27%) is attributed to residual error, which includes other factors not captured by this simple model, such as prior knowledge, test-taking anxiety, or learning efficiency.

Using the Excel Data Analysis Toolpak for Comprehensive Regression

While the **RSQ()** function provides a fast and accurate R-Squared value, often analysts require a full regression output which includes coefficients, standard errors, P-values, and the ANOVA summary. For these more advanced needs, Excel's Data Analysis Toolpak is the preferred utility. If the Toolpak is not visible under the 'Data' tab, it must first be activated through the 'File' > 'Options' > 'Add-ins' menu.

Using the Toolpak allows for the calculation of the R-Squared value as part of a much larger

statistical summary, validating the result obtained via the simple function while providing the necessary components to assess the model's overall significance. To run a regression analysis using the Toolpak, you must input the Y Range and X Range similarly to the **RSQ()** function, and specify an output range for the results table.

When the regression analysis is executed, Excel generates several summary tables. The first table, titled "Regression Statistics," contains the key goodness-of-fit metrics, including **R Square**. Reviewing this output confirms the accuracy of the manual function calculation and places the R-Squared value into a broader context of the model's performance metrics.

D	E	F	G	H	I	J	K	L
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.8528							
R Square	0.7273							
Adjusted R Square	0.7121							
Standard Error	5.2805							
Observations	20							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	1338.2906	1338.2906	47.9952	0.0000			
Residual	18	501.9094	27.8839					
Total	19	1840.2000						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	67.1617	2.6633	25.2178	0.0000	61.5664	72.7570	61.5664	72.7570
hours	5.2503	0.7578	6.9279	0.0000	3.6581	6.8424	3.6581	6.8424

As shown in the "Regression Statistics" output generated by the Data Analysis Toolpak, the R Square value is explicitly listed as **0.7273**. This result perfectly matches the value derived from our previous use of the standalone **RSQ()** function, thereby confirming the consistency and accuracy of both Excel methods for calculating this critical metric. For professional reports or academic work, utilizing the Toolpak is often mandatory as it provides the full statistical context required for peer review and rigorous analysis.

Limitations and Caveats of R-Squared

While R-Squared is an invaluable tool for assessing model fit, it is not without its limitations, and relying solely on a high R-Squared value can sometimes lead to misleading conclusions. A critical caveat is that R-Squared is guaranteed to increase or at least remain constant whenever a new independent variable is added to the model, even if that new variable is statistically insignificant or

irrelevant to the overall prediction. This inherent tendency encourages model overfitting, where the model becomes too tailored to the sample data and loses its generalizability to new, unseen data.

To counteract the problem of automatic R-Squared inflation, particularly in multiple regression scenarios, statisticians often turn to the **Adjusted R-Squared**. This modified version penalizes the model for the inclusion of unnecessary or poorly contributing predictor variables. The Adjusted R-Squared is a more honest reflection of the model's explanatory power because it only increases if the newly added term improves the model more than would be expected by chance. For serious regression analysis, especially when comparing models with different numbers of predictors, the Adjusted R-Squared should always be prioritized over the standard R-Squared.

Furthermore, a high R-Squared does not guarantee that the regression model is appropriate for the data. For instance, a model might have a high R-Squared but still violate key assumptions of linear regression, such as linearity, independence of errors, or homoscedasticity. Therefore, R-Squared should always be examined alongside other diagnostic tools, such as residual plots, F-tests, and t-statistics, to ensure the model is statistically sound and structurally valid for the underlying data relationship. Ultimately, R-Squared is a measure of correlation, not causation, and its proper interpretation requires statistical scrutiny beyond just the percentage value itself.