How to Calculate R-Squared in SAS Using PROC REG

Authored by

stats writer

November 19, 2025

Introduction to R-Squared Calculation in SAS

The question of whether SAS offers a streamlined method for calculating the coefficient of determination, known as R-Squared, is easily answered with a resounding "Yes." The statistical power of the SAS system is largely driven by its comprehensive set of procedures, chief among them being the PROC REG procedure. This procedure is specifically designed for regression analysis and automatically computes essential diagnostic statistics, including the crucial R-Squared value, as a standard component of its output.

The utility of PROC REG extends far beyond simple calculations. Analysts rely on it to fit a wide array of statistical models, encompassing basic linear regression, multiple regression, and models that can be linearized through transformation. By handling these complex modeling tasks efficiently, PROC REG ensures that data scientists can focus on interpreting their results rather than tedious manual calculations. Every time a model is successfully fitted using this procedure, the R-Squared statistic is presented alongside parameter estimates, confidence intervals, and detailed ANOVA tables, providing a complete statistical picture of the fitted model.

Understanding how to leverage the capabilities of PROC REG is fundamental for anyone performing quantitative analysis in the SAS environment. The following sections will not only define R-Squared rigorously but also provide a meticulous, step-by-step walkthrough detailing how to implement a linear regression model and extract this vital measure of goodness-of-fit within the SAS programming language.

Understanding the Significance of R-Squared

R-squared, formally known as the coefficient of determination and often written as R2, is a fundamental statistical measure used universally to assess the quality of a statistical model's fit to a given dataset. It quantifies the proportion of the variance in the dependent variable that is predictable from the independent variable(s) in a regression model. A higher R2 value generally indicates that the model is better at explaining the observed data, although context and adjusted R2 must also be considered for a complete evaluation.

This value effectively represents the proportion of the total variation, or variance, in the response variable--the value we are trying to predict--that can be successfully accounted for by the inclusion of the predictor variable or variables within the model. A strong model implies that the predictors provide substantial explanatory power over the response. Conversely, a low R2 suggests that a large portion of the response variability remains unexplained by the chosen predictors, indicating potential issues with the model specification or the inherent randomness of the data.

The value for R2 is mathematically constrained to range from 0 to 1, offering a clear, bounded metric for assessment:

A value of 0 indicates that the predictor variable or variables cannot explain the response variable at all. In this scenario, the model is no better than simply using the mean of the response variable for prediction, and all observed variability remains unexplained by the model.
A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable(s). This signifies a perfect fit where the data points fall exactly on the regression line or plane, meaning the model accounts for 100% of the total variance. In reality, achieving R2 = 1 is extremely rare, particularly with real-world, complex data.

While a high R2 is often desirable, it is important to exercise caution, especially when dealing with multiple regression, as adding more predictors will always increase R2, even if those predictors are not statistically significant. For this reason, practitioners often rely on the **Adjusted R-Squared**, which penalizes the model for including excessive, non-contributing variables, providing a more reliable measure of model fit, although standard PROC REG output includes both measures automatically.

The Versatility of the PROC REG Procedure

The PROC REG procedure stands as the cornerstone for modeling relationships between variables in SAS. Its primary function is to fit various forms of linear regression models, including simple linear models (one predictor), multiple linear models (several predictors), and polynomial regression models. The procedure's syntax is highly intuitive, requiring only a dataset specification and a model statement that defines the relationship between the response and predictor variables.

Beyond simply computing the coefficients, PROC REG generates a comprehensive suite of output tables critical for robust analysis. These outputs include the analysis of variance (ANOVA) table, which tests the overall significance of the model; parameter estimates tables, which provide the intercept and slope coefficients along with their standard errors and p-values; and, most pertinent to this discussion, the summary statistics table containing the R-Squared and Adjusted R-Squared values.

Furthermore, PROC REG supports advanced diagnostic features that allow users to check model assumptions and identify potential problems such as outliers, multicollinearity, and non-constant variance. Options are available for requesting residual plots, influence statistics, and prediction statistics, ensuring that the model is not only fit efficiently but also rigorously validated against the underlying data distribution and structure. This holistic approach makes PROC REG the preferred choice for detailed regression studies.

The following detailed example provides a step-by-step demonstration showing precisely how to define a model, execute the regression, and immediately obtain the R-Squared value for a simple linear regression model in SAS.

Step 1: Preparing Data for Simple Linear Regression

For the purpose of this practical illustration, we will begin by creating a simple dataset suitable for linear regression analysis. This dataset will track the relationship between the total hours a student dedicated to studying and their corresponding final exam score. We are working with a sample of 15 hypothetical students.

Our objective is to fit a simple linear regression model where the variable representing *hours studied* will serve as the predictor (independent) variable, and the variable representing *final exam score* will function as the response (dependent) variable. This model seeks to determine how well the study time predicts the academic outcome.

The code below shows the necessary steps to create and then display this dataset in the SAS environment. We use the **DATA step** to define the variables and input the values, followed by **PROC PRINT** to visualize the data table and verify its successful creation before proceeding to the regression analysis.

```
/*create dataset named exam_data*/
data exam_data;
input hours score;
datalines;
1 64
2 66
4 76
5 73
5 74
6 81
6 83
7 82
8 80
10 88
11 84
11 82
12 91
12 93
14 89
;
run;

/*view dataset to confirm data integrity*/
proc print data=exam_data;
```

run;

Obs	hours	score
1	1	64
2	2	66
3	4	76
4	5	73
5	5	74
6	6	81
7	6	83
8	7	82
9	8	80
10	10	88
11	11	84
12	11	82
13	12	91
14	12	93
15	14	89

The resulting output from the **PROC PRINT** command confirms that our dataset, **exam_data**, containing the two variables **hours** and **score** for 15 observations, has been correctly loaded into the SAS work library. We are now ready to utilize PROC REG to model the relationship between these variables.

Step 2: Implementing and Running the Regression Model

Once the data is prepared, the next step involves calling the **PROC REG** procedure to fit the simple linear regression model. This is achieved using a concise set of statements within the SAS code block. The primary statement required is the **MODEL** statement, where we define the relationship between the response and predictor variables.

We specify **score** as the response variable and **hours** as the predictor variable. The syntax is structured as `MODEL response = predictors;`. The procedure automatically calculates all necessary statistics, including the intercept, slope coefficient, and, most importantly for this discussion, the R-Squared value, without needing any special options.

The following code demonstrates the fitting of this simple linear regression model:

/*fit simple linear regression model using PROC REG*/

```
proc reg data=exam_data;
model score = hours;
run;
quit;
```

Interpreting the Regression Output and R-Squared

Upon execution of the **PROC REG** code, SAS generates a comprehensive set of tables detailing the model fit. The core table for model assessment typically appears early in the output, containing the goodness-of-fit statistics. This table includes the Root MSE, the Dependent Mean, the Coefficient of Variance (CoVar), and the critical R-Squared values.

Examining the output generated by the regression analysis for our exam data, we find the following results prominently displayed:

The REG Procedure
Model: MODEL1
Dependent Variable: score

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	847.26698	847.26698	63.91	<.0001
Error	13	172.33302	13.25639		
Corrected Total	14	1019.60000			

Root MSE	3.64093	R-Square	0.8310
Dependent Mean	80.40000	Adj R-Sq	0.8180
Coeff Var	4.52852		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	65.33395	2.10599	31.02	<.0001
hours	1	1.98237	0.24796	7.99	<.0001

By reviewing the output table labeled **Root MSE, R-Square, and Adjusted R-Square**, we can

immediately observe the calculated R-Squared value. In this specific output, the R-squared value is reported as **0.8310**. This result is highly significant, indicating that approximately 83.10% of the total variability observed in the final exam scores (the response variable) can be explained by the number of hours studied (the predictor variable). This suggests a strong predictive relationship between study time and exam performance.

The remaining 16.90% of the variance is attributed to other factors not included in the model, such as prior knowledge, test anxiety, or inherent ability. The proximity of 0.8310 to 1.0 confirms that the linear model provides a very good fit to this particular dataset. Additionally, the adjacent Adjusted R-Squared value, which is particularly useful in multiple regression, is also provided, offering a slightly more conservative estimate of the model's explanatory power.

Step 3: Extracting Only the R-Squared Value

In many analytical scenarios, especially when processing numerous models or integrating results into automated reports, an analyst may only require the R-Squared value itself, without the verbose output of the full regression analysis (ANOVA table, parameter estimates, etc.). PROC REG offers powerful options to suppress standard output and redirect specific results to a new dataset.

To achieve this efficient extraction, we utilize two key options within the PROC REG statement: **NOPRINT** and **OUTEST=**. The **NOPRINT** option suppresses the printing of all standard statistical output. The **OUTEST=** option creates an output dataset (here named **outest**) which contains all model estimates, including the R-Squared value, stored under the variable name **_RSQ_**. We must also explicitly request the R-Squared value using the **/RSQUARE** option within the **MODEL** statement.

The subsequent **PROC PRINT** statement is then used to access this newly created **outest** dataset and display only the variable **_RSQ_**, isolating the desired R-Squared result:

```
/*fit simple linear regression model and suppress standard output*/
proc reg data=exam_data outest=outest noprint;
model score = hours / rsquare;
run;
quit;

/*print only the R-squared value contained in the output dataset*/
proc print data=outest;
var _RSQ_;
run;
```

Obs	_RSQ_
1	0.83098

As observed in this final, streamlined output, only the R-squared value, displayed as **0.83098**, is presented. This confirms that the R-Squared value calculated by the model is successfully extracted and isolated. The use of **noprint** in proc reg is a powerful optimization tool, ensuring that SAS only processes and outputs the absolutely essential statistics required for the subsequent steps in the analytic workflow.

Advanced Diagnostic Features of PROC REG

While the focus here has been on calculating and extracting R-Squared, it is critical to remember that PROC REG offers many additional powerful features indispensable for thorough regression diagnostics. Analysts should rarely rely solely on R-Squared; the validation of model assumptions is equally important. PROC REG simplifies this process by allowing users to request various plots and statistics related to residuals.

For example, using options like **PLOTS=RESIDUALS(UNPACK)** in the **PROC REG** statement allows the user to generate detailed graphical assessments of the residuals, helping to verify assumptions of linearity, homoscedasticity (constant variance), and normality of errors. The procedure also provides leverage plots, Cook's D statistics, and DFFITS statistics, which are essential for identifying influential observations or outliers that might unduly bias the regression coefficients and the R-Squared value itself.

In summary, PROC REG is the definitive procedure in SAS for calculating R-Squared and performing all facets of linear regression. Whether you need a quick R-Squared value or a full diagnostic report, this procedure provides the flexibility and robustness required for high-quality statistical modeling.

The following tutorials explain how to perform other common tasks in SAS: