Perform Simple Linear Regression in SAS

Authored by

Mohammed looti

November 1, 2025

Simple linear regression is a foundational statistical technique used extensively across data science and analytics. Its primary function is to quantify the relationship between two continuous variables: one **predictor variable** (independent) and one response variable (dependent). Mastery of this method is essential for tasks ranging from forecasting future trends to establishing potential causality in empirical studies.

The core objective of simple linear regression is to identify the straight line that best fits the observed data points. This "line of best fit" minimizes the sum of squared distances between the line and the actual data observations. This fundamental relationship is defined by the following mathematical linear equation:

? = b0 + b1x

Understanding the components of this widely used regression formula is key to interpretation:

?: This term represents the estimated or predicted value of the **response variable** (Y).
b0: This is the **Y-intercept** of the regression line. It defines the predicted value of Y when the predictor variable X is equal to zero.
b1: This coefficient is the **slope** of the regression line. It precisely quantifies the average change in Y expected for every one-unit increase in X.

By calculating and interpreting these coefficients (b0 and b1), analysts gain profound insight into both the direction and magnitude of the relationship between the two variables. The comprehensive tutorial that follows will demonstrate the precise steps for performing and interpreting a simple linear regression analysis using **SAS** (Statistical Analysis System), a powerful software environment trusted by statisticians worldwide.

Step 1: Preparing and Importing the Data into SAS

To illustrate the implementation process effectively, we will utilize a hypothetical dataset designed to explore the relationship between study effort and academic outcome. Specifically, our dataset tracks the total **hours studied** (the predictor variable) and the corresponding **final exam score** (the response variable) for a cohort of 15 students.

Our analytical objective is to fit a simple linear regression model where *hours* predicts *score*. This analysis will statistically determine whether the amount of time dedicated to studying is a significant factor in forecasting student performance on the final examination.

The following code snippet demonstrates how to efficiently create and load this sample dataset directly within the SAS programming environment using the fundamental `DATALINES` statement. This method is common for quickly initializing small datasets for statistical demonstration or testing purposes.

```
/*create dataset*/
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*/
proc print data=exam_data;
```

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

Step 2: Fitting the Simple Linear Regression Model using PROC REG

With the data successfully loaded and verified, the next critical step is to execute the regression analysis. In SAS, the standard and most robust procedure for fitting linear regression models is **PROC REG**. This procedure not only calculates the required regression coefficients but also generates a comprehensive suite of statistical output essential for model evaluation.

The model specification is achieved using the mandatory `MODEL` statement. Within this statement, we explicitly define the response variable (`score`) as a function of the predictor variable (`hours`). The structure is intuitive and follows the standard statistical notation.

The following concise syntax initiates and executes the requested simple linear regression analysis within the SAS environment:

```
/*fit simple linear regression model*/
proc reg data=exam_data;
model score = hours;
run;
```

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

Step 3: Interpreting the Regression Output Tables

The output generated by **PROC REG** is divided into several tables, each providing critical metrics required to assess the model's validity, significance, and predictive power. We will focus our interpretation on three primary components: the Analysis of Variance (ANOVA) table, the Model Fit Summary, and the Parameter Estimates table.

Analysis of Variance (ANOVA) Table

The ANOVA table is crucial for evaluating the overall statistical significance of the regression model. In this specific analysis, the overall **F-statistic** for the model is calculated as **63.91**, yielding a corresponding p-value of **<.0001**.

Since this p-value is significantly lower than the conventional threshold of 0.05, we confidently reject the null hypothesis and conclude that the regression model, as a whole, is **statistically significant**. This finding confirms that the variable "hours studied" provides meaningful predictive

power over the final exam score.

Model Fit Summary (R-Square)

The R-Square value, also known as the Coefficient of Determination, is an essential metric that quantifies the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It is expressed as a percentage.

Generally, a higher R-Square value indicates a better fit, meaning the predictor variables are more successful at explaining the variation in the response variable. In this case, the analysis reveals that **83.1%** of the total variation observed in the exam scores can be attributed to--or explained by--the number of hours studied. This high value strongly confirms the substantial utility of "hours studied" as a predictor.

Parameter Estimates Table

The Parameter Estimates table yields the specific coefficients (intercept and slope) required to construct the fitted regression equation. This equation allows for predictive calculations:

Score = 65.33 + 1.98*(hours)

The slope coefficient (1.98) is interpreted as follows: For every additional hour a student spends studying, their predicted exam score increases by an average of **1.98 points**, assuming all other factors remain constant.

The intercept value (65.33) provides the baseline prediction: the average expected exam score for a student who studies zero hours is **65.33**. This baseline provides context for the relationship.

We can use this derived equation for direct forecasting. For instance, a student who studies for 10 hours is expected to achieve an exam score calculated as:

Score = 65.33 + 1.98*(10) = 85.13

Furthermore, by examining the p-value associated with the *hours* coefficient (<.0001), which is far below 0.05, we confirm that *hours studied* is a **statistically significant predictor variable** in this model.

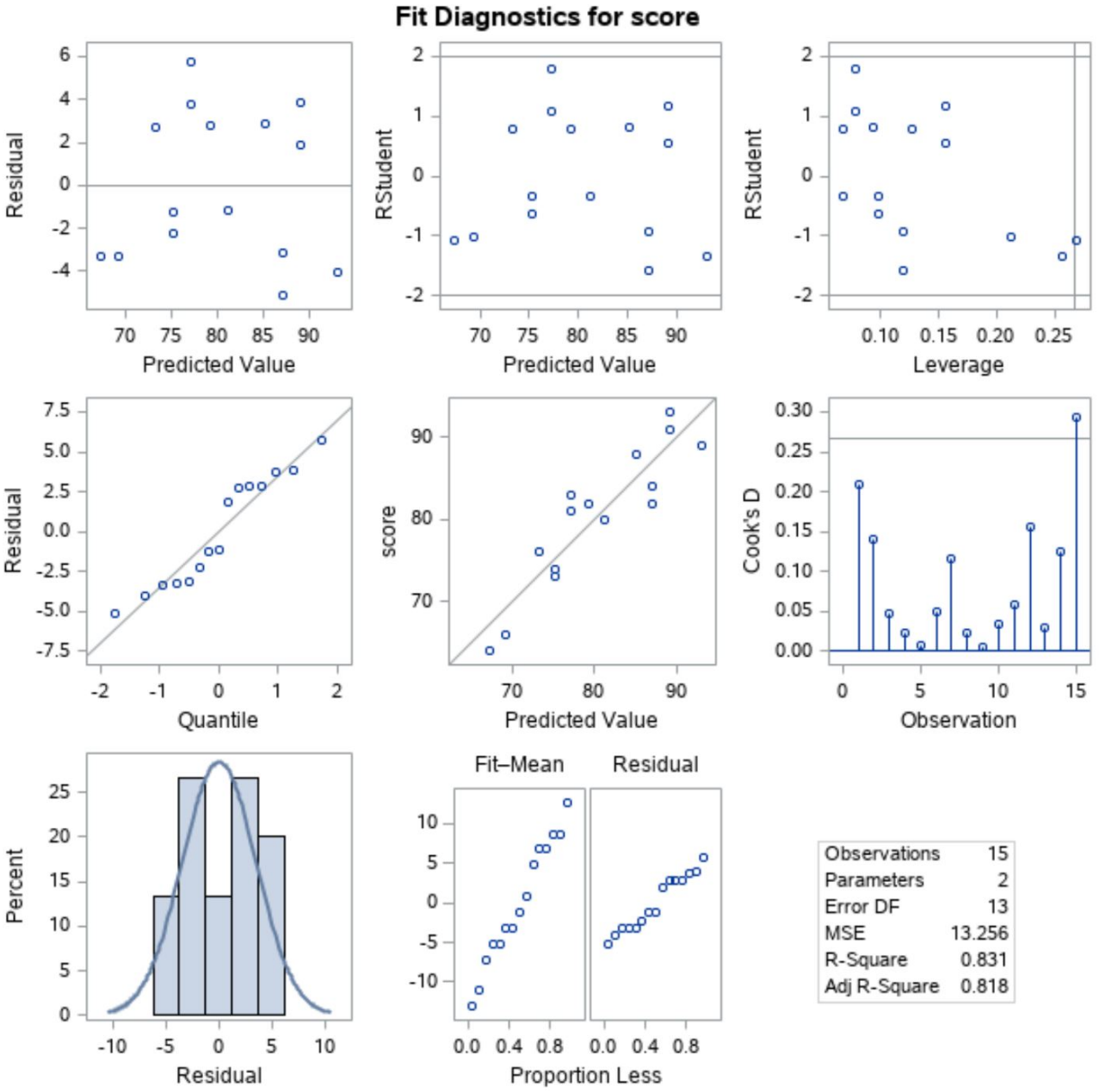
Step 4: Validating Model Assumptions with Residual Plots

The reliability of any statistical inference drawn from a simple linear regression model hinges entirely upon meeting several critical assumptions concerning the model's residuals. Residuals are defined as the vertical distances between the observed data points and the values predicted by the regression line.

The two primary statistical assumptions that must be rigorously verified are:

The residuals must be **normally distributed** around the mean of zero.
The residuals must exhibit **equal variance** (a condition known as homoscedasticity) across all levels of the predictor variable.

Violations of these assumptions can lead to biased coefficient estimates and unreliable p-values. Fortunately, SAS automates the generation of diagnostic plots, allowing us to visually verify these assumptions. The combined diagnostic plot output is displayed below:



Checking for Normality of Residuals

To determine if the residuals are **normally distributed**, we analyze the **Q-Q plot** (Quantile-Quantile plot) provided in the output. This plot compares the standardized residuals against the theoretical values expected under a perfect normal distribution.

A strong indication of normality is achieved when the data points align closely along the straight diagonal line in the Q-Q plot. Observing the generated plot, we see that the points generally track this line, leading us to conclude that the normality assumption for the residuals is reasonably met.

Checking for Homoscedasticity

To verify that the variance of the residuals is constant (i.e., they are homoscedastic), we inspect the plot labeled "Residual" vs. "Predicted Value" (typically located in the top left).

The key indicator of homoscedasticity is the random scattering of points around the horizontal zero line, without any discernible pattern such as a fan or funnel shape (which would indicate increasing or decreasing variance, known as heteroscedasticity).

In our plot, the points are scattered randomly about the zero line with roughly equal vertical spread throughout the range of predicted scores. Since both the normality and homoscedasticity assumptions are satisfied through visual inspection of these diagnostic plots, we can proceed with high confidence in the statistical results derived from this simple linear regression model.

Additional Resources

For users interested in expanding their SAS capabilities, the following tutorials explain how to perform other common statistical and data manipulation tasks: