How to Easily Perform Simple Linear Regression in SAS

Authored by

stats writer

December 1, 2025

The goal of Simple linear regression is to model the relationship between two continuous variables: a single response variable (also known as the dependent variable) and a single predictor variable (or independent variable). In the SAS statistical software environment, this powerful technique is implemented using the **PROC REG** procedure. This procedure systematically fits a straight line to the observed data points.

The core mechanism employed by PROC REG to determine this line is the least squares method. This method minimizes the sum of the squared differences between the observed response values and the values predicted by the model, ensuring the determination of the optimal coefficients for the intercept and the slope. The resulting output provides a comprehensive statistical summary, including model fit statistics, parameter estimates, and the final regression equation, allowing analysts to interpret the strength and direction of the relationship.

Mathematically, the simple linear regression model takes a canonical form, which represents the linear relationship established between the variables. Understanding this formula is fundamental to interpreting the regression output and translating statistical findings into practical insights. The following section details this formula and its essential components before we dive into the practical SAS implementation.

The Mathematical Foundation of Simple Linear Regression

The statistical model determines the line that minimizes prediction error, often referred to as the line of best fit. This line is expressed through a simple linear equation, which encapsulates the predicted relationship between the input and output variables. Recognizing the role of each component in this equation is vital for accurate model interpretation.

The equation is formally written as:

? = b0 + b1x

Where each term contributes specifically to the prediction:

?: Represents the **estimated response value**, which is the predicted value of the dependent variable for a given input *x*.
b0: Denotes the **intercept** of the regression line. This is the predicted value of *?* when the predictor

variable *x* is equal to zero.
b1: Represents the **slope** coefficient. This value quantifies the expected change in *?* for every one-unit increase in the predictor variable *x*.

By solving for the optimal values of the coefficients (b? and b?), the equation provides a quantifiable measure of the association between the predictor and response variables, forming the basis for statistical inference and prediction in our analysis.

Step 1: Preparing the Dataset for SAS Analysis

To demonstrate the implementation of simple linear regression in SAS, we must first define the dataset we will analyze. Our example focuses on a common educational scenario: examining the potential linear relationship between the time spent studying and the resulting academic performance. We will construct a dataset capturing the total hours studied and the final exam score for a cohort of 15 students.

In this particular model, the variable **hours** will serve as our predictor variable (X), as we hypothesize that study time influences the outcome. Conversely, the variable **score** will be designated as the response variable (Y), representing the outcome we wish to predict based on the input variable. Establishing these roles is crucial before executing the regression procedure.

The following SAS code utilizes the `DATA` step to create the necessary input table, named `exam_data`, followed by a `PROC PRINT` statement to display the initial observations and verify data integrity before modeling commences.

```
/*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;
```

The resulting table confirms that the 15 observations are correctly loaded into the SAS environment, ready for the regression analysis step.

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: Executing and Interpreting the PROC REG Output

The next logical step is to execute the simple linear regression analysis using the specialized PROC REG procedure in SAS. This command is the standard mechanism for fitting least squares regression models. The `MODEL` statement specifies the structure of the regression equation,

positioning the response variable (`score`) on the left and the predictor variable (`hours`) on the right.

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

Upon running this procedure, SAS generates extensive output structured into several tables. We must systematically examine these tables--specifically the Analysis of Variance (ANOVA), Model Fit Statistics, and Parameter Estimates--to fully understand the validity and predictive power of our regression model.

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

Analysis of Variance and Model Significance

The **Analysis of Variance (ANOVA) Table** assesses the overall statistical significance of the

regression model. It partitions the total variation in the response variable into the portion explained by the model (Regression) and the unexplained portion (Error). In our example, the F-statistic for the overall regression model is found to be **63.91**.

Crucially, the associated P-value is reported as **<.0001**. Since this P-value is significantly less than the conventional significance level of 0.05, we have strong statistical evidence to reject the null hypothesis. This leads us to the conclusion that the linear regression model, using hours studied as a predictor, is statistically significant in explaining the variation in exam scores.

Assessing Model Fit: The R-Square Metric

The **Model Fit Table** provides key metrics for evaluating how well the fitted line adheres to the observed data points. The most widely used metric here is the coefficient of determination, or R-Square. R-Square quantifies the proportion of the total variation in the response variable that is explained by the predictor variable in the model.

A higher R-Square value generally indicates a stronger fit, meaning the predictor variables are highly effective in predicting the response variable's value. For this analysis, the R-Square value is 0.831, meaning **83.1%** of the variation observed in the final exam scores can be successfully accounted for by the number of hours studied. This high percentage suggests that *hours studied* is an exceptionally useful and powerful variable for predicting academic performance in this dataset.

Interpreting the Regression Coefficients

The **Parameter Estimates Table** is perhaps the most informative section, as it delivers the specific coefficients (b? and b?) required to construct the final fitted regression equation. These estimates allow us to quantitatively describe the relationship between the variables:

The derived equation is: **Score = 65.33 + 1.98 * (Hours)**.

The **Intercept (b?)** is **65.33**. This suggests that a student who studied for zero hours is predicted to achieve an average exam score of 65.33 points.
The **Slope (b?)** for *hours* is **1.98**. This positive slope indicates that for every additional hour a student dedicates to studying, their expected exam score increases by an average of **1.98 points**.

Furthermore, we can use this explicit equation for prediction. For instance, if a student studies for

10 hours, their expected score is calculated as: **Score = 65.33 + 1.98*(10) = 85.13**. Since the P-value (<.0001) associated with the *hours* coefficient is highly significant (less than 0.05), we confirm that hours studied is a statistically significant predictor in this model.

Step 3: Validating Model Assumptions through Residual Analysis

While the statistical tables confirm the significance of the model, a crucial part of reliable regression analysis involves validating the underlying statistical assumptions. Violations of these assumptions can render the model results, including the P-values and confidence intervals, inaccurate or misleading. Simple linear regression relies fundamentally on two core assumptions concerning the distribution of the residuals (the differences between observed and predicted values):

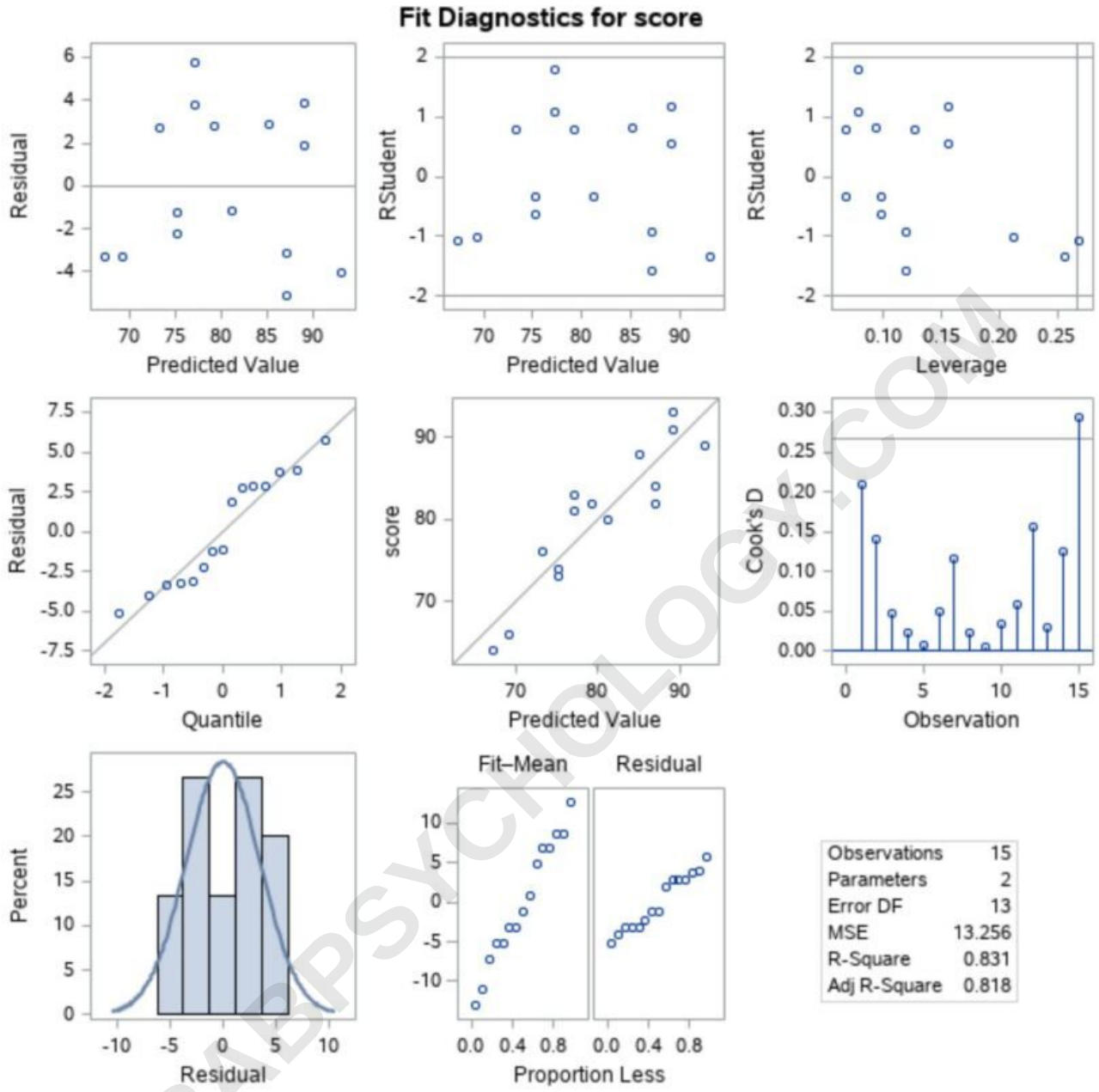
The **Normality of Residuals**: The error terms (residuals) must follow a normal distribution.
The **Homoscedasticity** (Constant Variance) of Residuals: The variance of the residuals must be equal across all levels of the predictor variable.

SAS automatically generates several diagnostic plots to assist in this visual inspection, allowing analysts to quickly assess whether these crucial assumptions have been met for the current dataset.

Checking for Normal Distribution (Q-Q Plot)

To verify the assumption of normally distributed residuals, we examine the Quantile-Quantile (Q-Q) Plot, typically located in the left position of the middle row in the SAS graphical output. This plot compares the observed residual quantiles against the theoretical quantiles expected from a normal distribution.

If the residuals are indeed normally distributed, the data points plotted in the Q-Q graph should align closely along a straight diagonal line. Observing the provided plot, the points generally cluster along this diagonal reference line, providing sufficient visual confirmation that we can reasonably assume the residuals in our `exam_data` model are normally distributed.



Checking for Constant Variance (Homoscedasticity)

The second critical assumption is homoscedasticity, which means the variability of the residuals must remain constant regardless of the predicted value. We assess this assumption by examining the plot of Residual vs. Predicted Value, usually found in the top left of the graphical output.

In a model demonstrating homoscedasticity, the residual points should be scattered randomly around the horizontal line at zero, exhibiting no funnel shape, cone shape, or discernible pattern. Critically, the spread (variance) of the points should appear roughly uniform across the entire range

of predicted values.

Our residual plot confirms that the points are scattered about zero randomly with roughly equal variance across the x-axis. Since both the normality and homoscedasticity assumptions are satisfied through this visual inspection, we can confidently assert that the results derived from our simple linear regression model are statistically robust and reliable for interpretation and prediction.

Conclusion and Further Resources

This comprehensive example demonstrates the complete process of executing and validating a simple linear regression model using **PROC REG** in SAS. By systematically following the steps of data preparation, model fitting, coefficient interpretation, and residual diagnostics, we successfully established a highly significant and reliable linear relationship between hours studied and exam scores.

The skills demonstrated here--understanding the output tables (ANOVA, R-Square, Parameter Estimates) and visually inspecting diagnostic plots--are essential for any advanced statistical modeling endeavor.

The following external resources provide additional guidance on performing other common statistical tasks and modeling techniques within the SAS environment: