How to Easily Perform Linear Regression Analysis with PROC REG in SAS

Authored by

stats writer

November 19, 2025

Introduction to PROC REG and Linear Modeling

The **PROC REG** procedure in SAS is the primary tool utilized by analysts and statisticians for performing comprehensive
linear regression analysis. This powerful statistical procedure is designed to model the relationship between one or more
independent variables (predictors) and a single
dependent variable (response). PROC REG generates a rich output that includes the estimated regression equation, the
Analysis of Variance (ANOVA) table, detailed
parameter estimates, and critical diagnostic statistics.

A classic application of PROC REG involves predicting a business metric, such as retail sales, based on influential factors like product price or advertising spend. By establishing the linear relationship between these variables, the resulting regression equation allows practitioners to quantify the impact of changes in the predictor variables and make reliable forecasts. Understanding the output--specifically the **regression coefficients** and model summary statistics--is essential for drawing valid conclusions from the data.

Essential Syntax for PROC REG

The structure of the **PROC REG** statement is highly intuitive, relying primarily on two main statements: the procedure call itself and the `MODEL` statement. The `MODEL` statement is crucial as it defines the precise relationship being tested, specifying the dependent variable (on the left side) and the independent variables (on the right side).

For fitting a **Simple Linear Regression** model, which involves only one predictor variable, the syntax is straightforward. This setup aims to estimate the equation y = b0 + b1x, where b0 is the intercept and b1 is the coefficient for the predictor X.

```
proc reg data = my_data;
model y = x;
run;
```

To conduct **Multiple Linear Regression**, where the response variable is modeled using two or more predictors (e.g., x1, x2, x3), you simply list all the predictor variables in the `MODEL` statement. This approach fits a model represented mathematically as y = b0 + b1x1 + b2x2 + b3x3, allowing for the assessment of multiple independent effects simultaneously.

```
proc reg data = my_data;
```

```
model y = x1 x2 x3;
run;
```

Detailed Example: Simple Linear Regression in SAS

To demonstrate the application and interpretation of **PROC REG**, we will utilize a dataset containing academic performance metrics. Imagine we have collected data from 15 students, recording the total **hours studied** and their corresponding **final exam score**. Our goal is to determine if study time significantly predicts exam performance.

We must first create and load this sample data into a SAS dataset named `exam_data` using the standard `DATA` and `DATALINES` steps. The `PROC PRINT` step is then used to verify that the data has been imported correctly, ensuring the integrity of the subsequent regression analysis.

```
/*Create the dataset: 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 the dataset content using PROC PRINT*/
proc print data=exam_data;
```

The output generated by `PROC PRINT` confirms the structure of our dataset, showing two variables, `hours` and `score`, for all 15 observations.

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

Fitting the Simple Linear Regression Model

With the data prepared, we proceed to fit the simple linear regression model. In this scenario, we hypothesize that the student's `score` is the **dependent variable** (Y) and the `hours` studied is the **independent variable** (X). The `MODEL` statement specifies this relationship: `score = hours`.

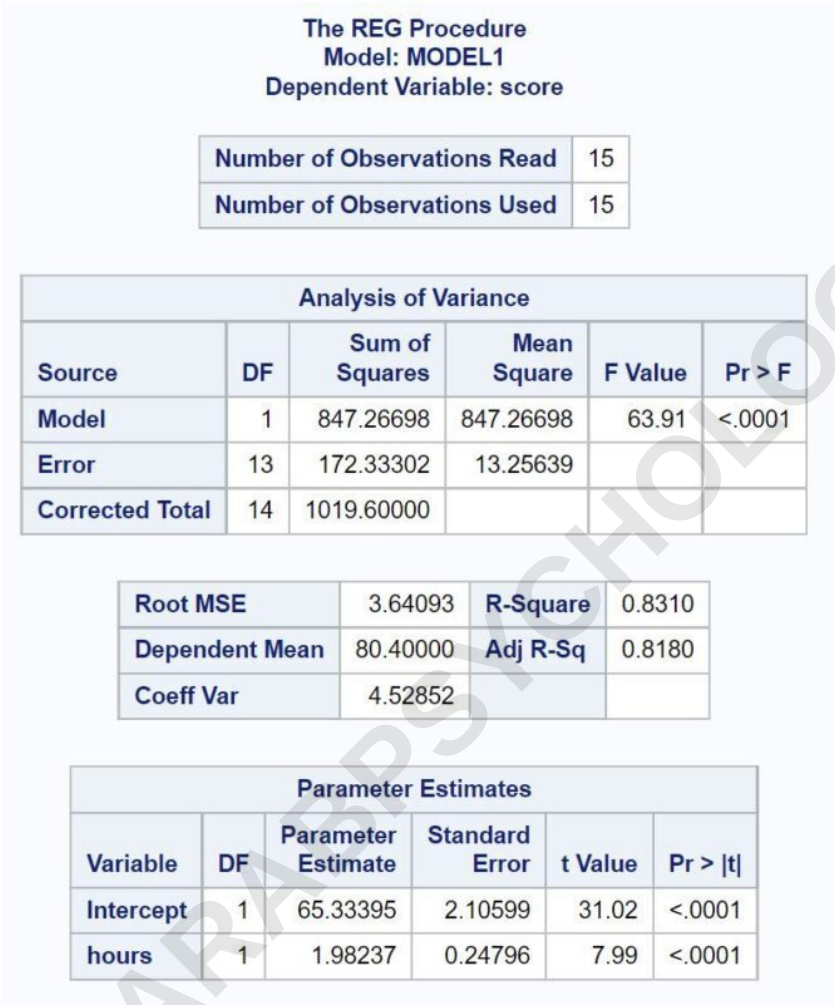
The following SAS code executes the regression analysis using **PROC REG**:

```
/*Fit the simple linear regression model using score as the response variable*/
proc reg data = exam_data;
model score = hours;
run;
```

Upon execution, SAS generates a comprehensive output composed of several tables crucial for statistical assessment, including the Model Fit Summary, the ANOVA table, and the Parameter Estimates table.

Interpreting Key Output Tables

The initial segments of the PROC REG output provide crucial indicators of how well the regression line fits the observed data. The Model Summary table includes metrics like R-squared, which quantifies the proportion of variance in the dependent variable explained by the independent variable(s). Following this is the Analysis of Variance (ANOVA) table, which tests the overall significance of the 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

The most critical information for constructing the predictive equation is found in the **Parameter Estimates** table. This table lists the estimated coefficients for the model's intercept and each predictor variable, along with their standard errors, t-statistics, and p-values, which determine the statistical significance of each predictor.

Deriving the Regression Equation

By extracting the values from the **Parameter Estimates** table, we can formally state the fitted

regression equation. The intercept (labeled as 'Intercept' or 'B0') represents the expected score when the hours studied (X) is zero. The coefficient for 'hours' (labeled as 'B1') represents the change in score for every one-unit increase in hours studied.

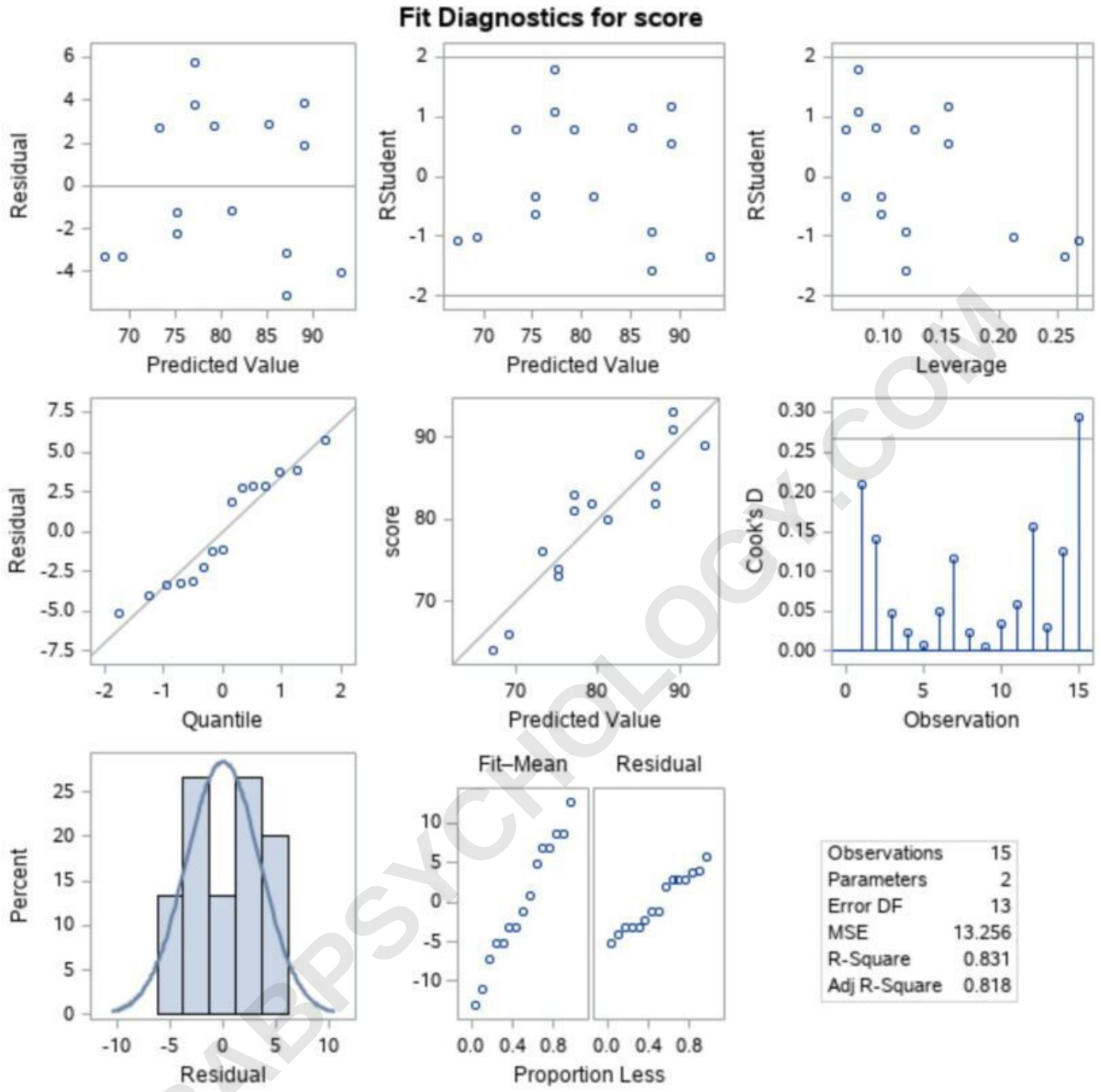
Based on the provided output, the fitted model equation is:
Score = 65.33 + 1.98 * (hours studied)

This equation suggests that a student who studies for zero hours is predicted to score 65.33, and for every additional hour studied, the predicted final score increases by approximately 1.98 points. This allows for prediction and quantification of the relationship between **study time** (independent variable) and **exam performance** (dependent variable).

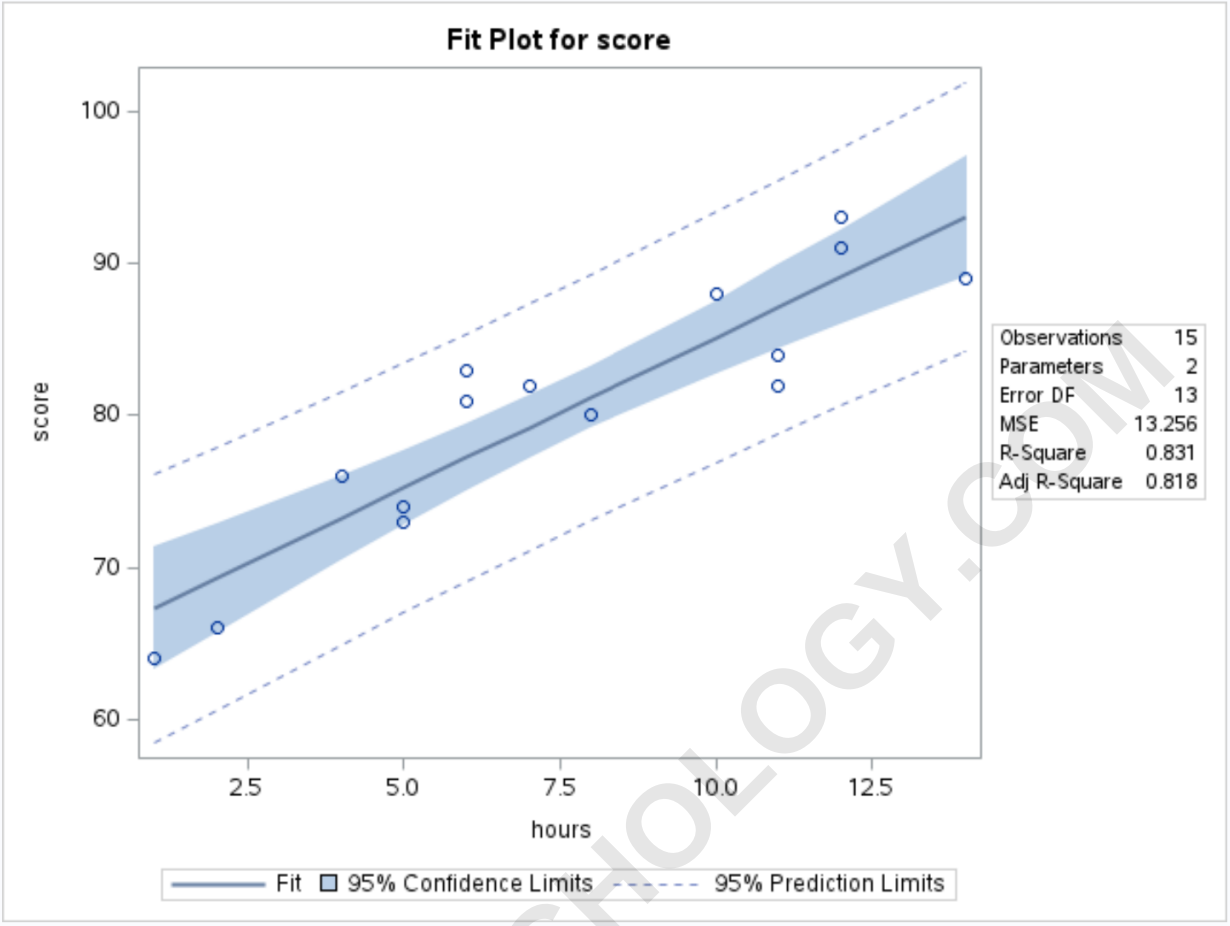
Analyzing Diagnostic and Scatter Plots

Beyond the statistical tables, **PROC REG** automatically generates several essential graphical outputs, which are vital for validating the underlying assumptions of the linear model. These diagnostic graphs, especially the
residual plots, help assess linearity, homoscedasticity (constant variance of errors), and the normality of residuals.

Checking the residual plots is a crucial step in regression analysis. For example, a plot of residuals versus predicted values should ideally show a random scatter of points around zero, without any discernible patterns (such as a cone shape or curve). Deviations from this pattern indicate potential violations of key regression assumptions, which may necessitate model adjustment or the use of alternative analytical methods.



Furthermore, PROC REG provides a visualization of the raw data points alongside the calculated regression line. This **scatterplot** offers an immediate visual confirmation of the relationship between the independent and dependent variables, allowing the analyst to intuitively assess the strength and direction of the correlation and how well the fitted line represents the data cloud.



If the regression line visibly passes close to the majority of data points, as demonstrated above, it suggests a strong fit, reinforcing the conclusions drawn from the statistical metrics presented in the Parameter Estimates and ANOVA tables.

Summary and Advanced Applications

The **PROC REG** procedure is foundational for statistical modeling in SAS, providing a streamlined process for conducting both simple and multiple linear regression analyses. By thoroughly examining the statistical output--including the ANOVA results and parameter estimates--and validating the model assumptions using the automatically generated residual plots, analysts can develop robust predictive models based on their data.

While this tutorial focused on basic usage, **PROC REG** supports advanced options such as subset selection methods (e.g., stepwise, best subsets), handling categorical predictors, and generating specialized influence statistics to detect outliers. Mastering these options allows for greater flexibility and sophistication in model building.

Note: For detailed command options, usage specifications, and comprehensive examples related

to advanced features of the procedure, consult the official SAS documentation for **PROC REG**.

Related SAS Tutorials

To further enhance your skills in SAS programming and statistical analysis, the following resources cover other common tasks:

How to perform Logistic Regression in SAS.
Understanding the use of PROC MEANS for descriptive statistics.
Methods for creating custom formats using PROC FORMAT.