How to Easily Perform Multiple Linear Regression in SAS

Authored by

stats writer

December 1, 2025

Multiple linear regression (MLR) is a foundational statistical technique used extensively in fields ranging from economics to social science, providing a powerful framework for understanding how a dependent variable is influenced by two or more independent, or predictor, variables. In the SAS statistical software environment, performing MLR is efficiently handled using the dedicated regression procedure, known as PROC REG.

The general syntax for the PROC REG statement is straightforward yet highly effective: `PROC REG ; MODEL response-variable = predictor-variable1 predictor-variable2 ... ; RUN;`. This structured command instructs SAS to fit the specified linear model, calculating essential statistical outputs such as coefficients, residuals, and model fit statistics.

Once executed, this procedure generates comprehensive output detailing the estimated coefficients, standardized residuals, various model fit statistics, and crucial hypothesis tests. Analyzing these results allows researchers and analysts to make robust inferences regarding the complex relationships between the response variable and its chosen predictors, ensuring data-driven conclusions are drawn.

Introduction to Multiple Linear Regression in SAS

At its core, Multiple linear regression serves as a crucial methodology for modeling the relationship between a single continuous response variable and several explanatory variables. Unlike simple linear regression, which utilizes only one predictor, MLR acknowledges the reality that most outcomes are influenced by multiple interacting factors simultaneously, thereby offering a more nuanced view of causality or association.

The mathematical representation of a multiple regression model adheres to the linear form: $Y = beta_0 + beta_1 X_1 + beta_2 X_2 + ... + beta_k X_k + epsilon$. Here, Y is the response variable, X_i are the predictor variables, $beta_0$ is the intercept, $beta_i$ are the respective coefficients or parameter estimates representing the effect of X_i on Y (holding other predictors constant), and $epsilon$ represents the irreducible error term.

This tutorial will guide you step-by-step through the process of setting up and running a complete Multiple linear regression analysis using SAS, focusing specifically on interpreting the wealth of information provided in the output tables to derive meaningful statistical insights regarding the significance and magnitude of predictor effects.

Understanding the PROC REG Statement

The `PROC REG` procedure is the standard tool in SAS for fitting ordinary least squares regression models, which is the underlying method for MLR. It is part of the STAT procedures and offers extensive capabilities for model specification, including options for subset selection, diagnostic

plotting, and hypothesis testing beyond the basic output.

A fundamental understanding of the structure required by `PROC REG` is vital for successful implementation. It begins with the procedure call, optionally specifying the input dataset using the `DATA=` option. The core of the analysis, however, lies within the `MODEL` statement, where the relationship between the dependent and independent variables is explicitly defined.

The `MODEL` statement syntax strictly places the response variable on the left side of the equals sign, followed by the predictor variables listed on the right side, separated by spaces. For example, if we are predicting 'Score' based on 'Hours' and 'Prep_Exams', the command is `MODEL score = hours prep_exams;`. This structure clearly articulates the functional relationship the procedure must estimate, seeking to minimize the sum of squared residuals.

Step 1: Defining the Research Goal and Data Structure

Before executing any code, defining the research objective is paramount. We hypothesize that student performance is a function of effort and preparation. Therefore, our goal is to fit a Multiple linear regression model that uses the number of hours spent studying and the number of preparatory exams taken to predict the final exam score of students.

The hypothesized linear model we aim to estimate can be mathematically expressed as:

Exam Score = β0 + β1(hours) + β2(prep exams) + Error Term

In this context, `Exam Score` is the response variable, while `hours` and `prep exams` are the predictor variables. We are attempting to estimate the unknown regression parameters (β coefficients) that quantify the influence of each predictor, controlling for the effect of the other. Proper identification of these variables is the first step in ensuring a valid statistical analysis.

Step 2: Implementing the Multiple Linear Regression Dataset in SAS

To perform the analysis, we first need to structure our raw data within a SAS dataset. For demonstration purposes, we will create a dataset named `exam_data` containing information for 20 hypothetical students, encompassing their study hours, prep exams taken, and corresponding final scores. This process simulates having gathered and organized empirical data for analysis.

The data creation process utilizes the `DATA` and `DATALINES` statements, standard procedures for inputting small datasets directly into the SAS environment. The `INPUT` statement defines the variables in the order they appear in the data block (hours, prep_exams, score). It is critical to ensure data integrity and variable type (all are numeric continuous variables in this scenario) before proceeding to the regression analysis.

The following code block demonstrates how to structure and load this raw data directly into the SAS environment, concluding with the necessary `RUN;` statement to finalize the dataset creation:

```
/*create dataset*/
data exam_data;
input hours prep_exams score;
datalines;
1 1 76
2 3 78
2 3 85
4 5 88
2 2 72
1 2 69
5 1 94
4 1 94
2 0 88
4 3 92
4 4 90
3 3 75
6 2 96
5 4 90
3 4 82
4 4 85
6 5 99
2 1 83
1 0 62
2 1 76
;
run;
```


Step 3: Executing the PROC REG Procedure

With the dataset successfully created and stored as `exam_data`, the next logical step is to execute the PROC REG procedure to estimate the Multiple linear regression model. This step requires specifying the input dataset and defining the precise model structure using the `MODEL` statement.

The `proc reg data=exam_data;` statement initializes the regression procedure, directing it to use our prepared student data. The subsequent `model score = hours prep_exams;` command defines the statistical relationship, specifying that 'score' is the outcome variable and 'hours' and

'prep_exams' are the independent variables whose effects we wish to quantify.

The complete SAS code required to fit the model is as follows, demonstrating the concise yet powerful nature of the PROC REG syntax:

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

The REG Procedure
Model: MODEL1
Dependent Variable: score

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1350.75688	675.37844	23.46	<.0001
Error	17	489.44312	28.79077		
Corrected Total	19	1840.20000			

Root MSE	5.36570	R-Square	0.7340
Dependent Mean	83.70000	Adj R-Sq	0.7027
Coeff Var	6.41064		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	67.67353	2.81580	24.03	<.0001
hours	1	5.55575	0.89919	6.18	<.0001
prep_exams	1	-0.60169	0.91439	-0.66	0.5193

Interpreting the Analysis of Variance (ANOVA) Table

The first critical output generated by PROC REG is the Analysis of Variance (ANOVA) table. This

table is essential for assessing the overall statistical significance of the regression model. It determines whether the collective influence of the independent variables significantly predicts the dependent variable.

The ANOVA table partitions the total sum of squares into the Sum of Squares due to the Model (SS Model) and the Sum of Squares due to Error (SS Error). The F-statistic is derived from these sums of squares and tests the global null hypothesis (H_0: $beta_1 = beta_2 = 0$), meaning none of the predictors have a linear relationship with the response.

In our output, the overall F-statistic for the regression model is reported as **23.46**, and the corresponding p-value is extremely small, indicated as **<.0001**. Since this p-value is substantially less than the standard significance level ($alpha = 0.05$), we reject the null hypothesis. This key finding confirms that the regression model, incorporating 'hours studied' and 'prep exams taken', is statistically significant and provides explanatory power for the variation in exam scores.

Evaluating Model Performance: R-Square and Root MSE

Beyond overall significance, model fit statistics provide quantitative measures of how well the regression line or plane approximates the actual data points. The Model Fit table focuses on key metrics such as R-Square and Root MSE (Root Mean Square Error).

The **R-Square** value, or the coefficient of determination, measures the proportion of the total variability in the response variable that is explained by the model. It is a critical metric for understanding the explanatory power of the predictors. A value closer to 1 (or 100%) indicates a strong fit, while a value near 0 suggests poor predictive capability.

For our student exam score model, the R-Square value is **0.734**. This high value signifies that **73.4%** of the variation observed in the final exam scores is successfully explained by the combined effects of study hours and prep exams. Additionally, the **Root MSE**, reported as **5.3657**, provides a tangible measure of model error. It represents the standard deviation of the residuals, meaning that, on average, the observed exam scores deviate by approximately 5.37 units from the values predicted by the regression equation.

Analyzing the Parameter Estimates and Formulating the Regression Equation

The Parameter Estimates table is the heart of the regression output, as it quantifies the specific contribution of each predictor variable. This table provides the estimated coefficients (the $beta$ values) necessary to construct the fitted regression equation, which is used for prediction and effect size interpretation.

The regression equation is constructed using the estimate for the intercept (the expected score

when both predictors are zero) and the estimates for the slopes of each predictor variable. Based on the SAS output, we identify the following estimated parameters:

Intercept: 67.674
Hours Studied: 5.556
Prep Exams: -0.602

Using these estimates, we formulate the final fitted regression equation:

Exam score = 67.674 + 5.556*(hours) - 0.602*(prep_exams)

Each coefficient provides a ceteris paribus interpretation: holding the number of prep exams constant, an increase of one hour of study is associated with a 5.556 point increase in the predicted exam score. Conversely, holding study hours constant, each additional prep exam is associated with a predicted decrease of 0.602 points.

Practical Application and Statistical Significance of Predictors

The fitted regression equation is a powerful tool for empirical prediction. We can use it to forecast the expected performance of any student given specific inputs for the predictor variables, provided those inputs fall within the range of our observed data.

For example, consider a student who studies for **3 hours** and takes **2 prep exams**. Substituting these values into our derived equation yields the estimated exam score:

Estimated exam score = 67.674 + 5.556*(3) - 0.602*(2) = 67.674 + 16.668 - 1.204 = **83.138**

Thus, the expected exam score for this student is approximately **83.1**. Beyond prediction, it is crucial to review the statistical significance of each predictor, indicated by the p-value associated with its respective T-test, which determines if the predictor is necessary in the model.

The p-value for 'hours' is **<.0001**. Since this is far less than 0.05, we conclude that study hours have a statistically significant, positive association with the exam score. However, the p-value for 'prep_exams' is **.5193**. As this value is substantially greater than the 0.05 threshold, we conclude that the number of prep exams taken does not demonstrate a statistically significant independent association with the exam score in this model.

Further Exploration in Regression Analysis using SAS

The lack of statistical significance for the 'prep_exams' variable warrants careful consideration. While the model is globally significant (F-test), the individual contribution of one predictor may not be meaningful. A common practice in statistical modeling is to simplify the model by removing non-

significant predictors, a process known as model selection or reduction, thereby potentially increasing model parsimony and reducing variance.

We might decide to perform a simpler analysis, known as simple linear regression, using only 'hours studied' as the single predictor variable, given its strong significance. This iterative process of model refinement is standard in regression analysis and often leads to the most robust and interpretable model.

To continue honing your statistical skills, consider exploring other advanced analytical tasks available in SAS, which are crucial for model validation:

Performing diagnostic checks on residuals to assess assumptions like normality and homoscedasticity.
Testing for multicollinearity among predictors, which can destabilize parameter estimates.
Running analysis of covariance (ANCOVA) procedures to incorporate categorical variables.

Mastering these methods allows for a deeper understanding and validation of the fitted model, ensuring that the assumptions underlying Multiple linear regression are met and that the results are reliable for inferential purposes.