Perform Multiple Linear Regression in SAS

Authored by

Mohammed looti

November 1, 2025

Statistical modeling serves as the fundamental bedrock of modern data analysis, enabling researchers and analysts to rigorously quantify and understand the complex relationships that exist between various measured factors. Within this analytical framework, Multiple Linear Regression (MLR) stands out as one of the most powerful and frequently utilized methods. MLR is a robust statistical procedure employed to meticulously examine how two or more independent factors, commonly referred to as **predictor variables**, simultaneously influence a single continuous outcome, known as the response variable.

This comprehensive and expert tutorial is specifically designed to guide you through the entire process of performing and accurately interpreting a Multiple Linear Regression analysis within the industry-standard **SAS** software environment. Mastery of this procedure is crucial for anyone engaging in serious quantitative research. We will systematically navigate the essential phases of this analysis, beginning with the necessary data preparation steps, moving through the precise execution of the modeling procedure using SAS commands, and concluding with a critical, in-depth interpretation of the resulting statistical output tables. Our aim is to provide clarity and technical precision at every step.

Understanding the Foundation of Multiple Linear Regression

A key distinction of MLR from its simpler counterpart, Simple Linear Regression, is its capacity to incorporate the simultaneous influence of multiple predictors. By accounting for several variables at once, MLR significantly enhances the predictive power and the overall realism of the resulting model. This is especially vital when attempting to model complex phenomena in fields like economics, public health, or educational research, where outcomes are rarely determined by a single factor. The primary analytical objective of MLR is twofold: first, to definitively identify which **predictor variables** contribute significantly to the variation observed in the response variable, and second, to precisely quantify the direction (positive or negative) and the magnitude of these relationships.

To provide a concrete, practical demonstration, we will analyze factors hypothesized to affect student academic performance. Our central hypothesis posits that a student's final **exam score** (serving as the continuous response variable) is systematically influenced by two distinct predictor variables: the **number of hours spent studying** and the **number of preparatory exams taken** throughout the semester. This scenario allows us to test the unique contribution of each factor while holding the other constant, a capability only afforded by multiple regression techniques.

The underlying theoretical linear model that we aim to estimate using the **SAS** software forms the mathematical basis of our entire analysis. This equation formalizes the relationship between the outcome and the predictors, including an essential term for unaccounted variation. The model is mathematically expressed as follows:

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

In this formulation, β0 represents the estimated intercept (the expected score when both predictors are zero), β1 and β2 are the respective regression coefficients quantifying the effect of a one-unit change in each predictor, and ε rigorously accounts for the residual, random error term inherent in all statistical models.

Step 1: Defining and Preparing the Data in SAS

The initial and perhaps most critical phase in any successful statistical analysis within the **SAS** environment is the proper definition and loading of the analytical dataset. It is imperative to ensure that the data is meticulously structured, with clearly defined variables corresponding to both the response variable and all included predictor variables. For this instructional demonstration, we will manually construct a sample dataset, which we will formally name `exam_data`. This dataset will comprise observations for 20 hypothetical students, accurately capturing their reported study hours, the count of preparatory exams they completed, and their final examination score.

We leverage the foundational `data` step in conjunction with the precise `input` and `datalines` statements. This combination allows us to structure and input our raw data directly into the program editor, bypassing the need for external file imports. While external imports are standard for large-scale production data, this direct input method is highly efficient for smaller datasets and serves as an excellent pedagogical tool for demonstrating the core principles of data handling in SAS.

The following self-contained code block meticulously defines the variable structure and inputs the specific data points required, thereby establishing the necessary analytical foundation for our subsequent Multiple Linear Regression analysis:

```
/*create dataset containing study hours, prep exams, and score*/
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 2: Executing the Regression Analysis using PROC REG

Once the data is accurately prepared and loaded into the **SAS** session, the next logical step is to execute the statistical procedure. To fit the Multiple Linear Regression model, we utilize the specialized **PROC REG** procedure. PROC REG is the standard, highly robust tool specifically designed within the SAS Statistical Analysis System for fitting and evaluating general linear models, and it generates a comprehensive suite of diagnostic output essential for thorough statistical evaluation.

The successful execution of this procedure necessitates two primary SAS statements. First, the procedure call itself, `proc reg`, must clearly specify the name of the input dataset (`exam_data`). Second, and most critically, the `model` statement must define the exact structure of the linear relationship being tested. Within the `model` statement, the variable listed immediately after the equals sign (in our case, `score`) is designated as the continuous response variable, while all subsequent variables (`hours` and `prep_exams`) are explicitly defined as the independent predictor variables.

The concise yet powerful SAS code required to fit our specific multiple regression model is presented below. This command instructs the software to calculate the parameter estimates that best minimize the sum of squared errors between the predicted and observed exam scores:

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

Following the execution of these commands, SAS generates a series of detailed output tables. These tables must be systematically analyzed to assess both the overall quality of the model fit and to interpret the specific coefficient values. The subsequent sections walk through the essential components of this output, which collectively contain all the necessary information for a complete statistical interpretation.

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 Output: Overall Model Significance (ANOVA)

The first crucial component of the **PROC REG** output that requires meticulous scrutiny is the Analysis of Variance (ANOVA) table. This table provides an indispensable assessment of the overall statistical significance of the entire regression model. It executes a global F-test, which is designed to test the stringent null hypothesis: that all regression coefficients for the predictor variables included in the model are simultaneously equal to zero. Rejecting this null hypothesis

signifies that the model, as a whole, possesses significant predictive capability.

Our focus within the ANOVA table must be directed toward the calculated **F-statistic** and its corresponding p-value, as these metrics collectively determine the model's overall efficacy. Reviewing the provided output, we observe that the global F-statistic for our regression model is calculated to be **23.46**. More critically, the associated p-value is reported as **<.0001**. This outcome is exceptionally strong.

Given that the calculated p-value (which is practically zero) is dramatically smaller than the universally accepted statistical significance threshold of $\alpha = 0.05$, we possess strong empirical evidence to confidently reject the null hypothesis. This powerful result unequivocally confirms that the combined regression model, which utilizes both study hours and preparatory exams as predictors, is **statistically significant** in its ability to predict the final exam score. This initial step validates that at least one of the predictor variables is contributing meaningfully to the explanation of the variance in the response variable.

Assessing Model Fit and Predictive Strength

Following the validation of overall significance through the ANOVA F-test, attention shifts to the model fit statistics. These quantifiable metrics are essential for understanding how effectively the fitted linear equation explains the observed dispersion, or variation, in the response variable, thus providing robust insight into the model's practical predictive strength. These statistics help transition the analysis from simply confirming significance to evaluating utility.

The **R-Square** value, formally known as the Coefficient of Determination, is the premier indicator of model fit in regression analysis. This statistic specifically represents the proportion of the total variance present in the dependent variable (exam score) that is successfully accounted for or explained by the set of independent variables (hours and prep exams). In the context of this specific analysis, the R-Square value is reported as **0.734**, which is conventionally interpreted as 73.4%.

This high percentage yields a clear practical interpretation: **73.4%** of the total variability observed in the students' final exam scores can be accurately attributed to the combined linear effects of the number of hours studied and the number of preparatory exams taken. This robust finding strongly suggests that the model provides an excellent, reliable fit to the underlying relationship within the data, leaving only 26.6% of the variance unexplained by these two factors. Furthermore, we must also consider the **Root MSE** (Root Mean Square Error). Root MSE serves as a standardized measure of the standard deviation of the error term, essentially quantifying the average distance that the observed data points deviate from the mathematically fitted regression line. This measure is reported in the same units as the response variable, which, for this model, is **5.3657** score units. A lower Root MSE is desirable, as it indicates higher predictive accuracy--meaning the model's

predicted scores are, on average, very close to the students' actual scores.

Interpreting Individual Parameter Estimates and Coefficients

The Parameter Estimates table is arguably the most fundamental component of the **PROC REG** output, as it furnishes the specific numerical coefficients required to mathematically construct the final, empirical fitted regression equation. These precise coefficients quantify the expected mean change in the response variable (exam score) for every one-unit increase in a specific predictor variable, critically assuming that the values of all other predictors in the model are rigorously held constant (the "ceteris paribus" condition). This is where the unique power of Multiple Linear Regression truly manifests.

By extracting the coefficients provided in the output (the Intercept, Hours coefficient, and Prep Exams coefficient), we can formally write the estimated empirical regression equation derived from our sample data:

Estimated Exam Score = 67.674 + 5.556 × (hours) – 0.602 × (prep_exams)

This equation permits immediate practical application and forecasting. For instance, if a student reports studying for 3 hours and taking 2 preparatory exams, the model predicts their expected exam score to be approximately **83.1**. This is calculated directly through substitution: 67.674 + 5.556 × (3) – 0.602 × (2) = **83.1**. Furthermore, the coefficient of 5.556 for `hours` means that, holding the number of prep exams constant, every extra hour studied is associated with an increase of 5.556 points in the final score.

Finally, we must assess the individual statistical contribution of each predictor by carefully examining its corresponding p-value (t-test significance):

For the `hours` variable, the associated p-value is **<.0001**. Since this value is far below the 0.05 significance level, we conclude that the number of hours studied has a strong, **statistically significant** positive relationship with the final exam score, independent of the number of prep exams taken.
Conversely, for the `prep_exams` variable, the p-value is **.5193**. As this value substantially exceeds the 0.05 cutoff, we conclude that the number of prep exams taken does **not** have a statistically significant unique association or incremental contribution to predicting the exam score once the powerful effect of study hours has already been accounted for in the model.

Conclusion and Strategies for Model Refinement

Our comprehensive analysis of the parameter estimates and overall model fit leads to a nuanced conclusion regarding the factors influencing student performance. We confirmed that the overall

model is highly significant (validated by the strong ANOVA F-test), demonstrating that the predictors collectively explain a large portion of the variance (R-Square = 73.4%). However, a deeper look revealed that one of the chosen predictors, `prep_exams`, fails to provide a statistically significant incremental contribution to the prediction of the score when `hours` studied is already included in the equation.

In standard statistical practice, analysts are strongly encouraged to refine their models based on empirical evidence to achieve greater parsimony. Parsimony refers to the principle of achieving the best possible fit with the fewest possible variables. Model simplification, often involving the removal of non-significant variables (like `prep_exams` in this case), typically results in a model that is both more robust against noise and far easier to interpret and communicate to stakeholders. The inclusion of non-significant predictors can sometimes obscure the true effects of the important variables.

Given the weight of the statistical evidence derived from the SAS output, a compelling case exists to streamline the analysis. The next logical step in this iterative process would be to remove the `prep_exams` variable and proceed with a Simple Linear Regression using only `hours` studied as the solitary predictor variable. This ensures the final model is not only statistically sound but also optimally efficient, focused solely on the factors demonstrated to have a statistically meaningful impact on the response variable.

Expanding Your Proficiency in SAS Programming

To further solidify your expertise in statistical modeling, data management, and quantitative analysis within the robust **SAS** software environment, we highly recommend continuing your exploration of related analytical procedures. Mastering these fundamental tools is absolutely essential for progressing toward advanced data analysis techniques and handling more complex research questions:

Performing Correlation Analysis in SAS for measuring bivariate relationships.
Executing T-Tests and Z-Tests in SAS for hypothesis testing concerning population means.
Advanced Data Manipulation techniques using the **SAS DATA step** for cleaning and transforming raw data inputs.