How can I perform multiple linear regression in SAS?

Authored by

stats writer

July 1, 2024

Multiple linear regression is a statistical method used to analyze the relationship between multiple independent variables and a dependent variable. In SAS, this can be performed by using the PROC REG procedure. This procedure allows the user to specify the variables to be included in the regression model and provides various diagnostic measures to assess the model's goodness of fit. Additionally, SAS also allows for the inclusion of categorical variables and interaction terms in the regression model. By using PROC REG in SAS, users can efficiently conduct multiple linear regression analysis and make informed decisions based on the results.

Perform Multiple Linear Regression in SAS

is a method we can use to understand the relationship between two or more predictor variables and a .

This tutorial explains how to perform multiple linear regression in SAS.

Step 1: Create the Data

Suppose we want to fit a multiple linear regression model that uses number of hours spent studying and number of prep exams taken to predict the final exam score of students:

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

First, we'll use the following code to create a dataset that contains this information for 20 students:

/*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 2: Perform Multiple Linear Regression

Next, we'll use proc reg to fit a multiple linear regression model to the data:

```
/*fit multiple linear regression model*/proc regdata=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

Here is how to interpret the most relevant numbers in each table:

Analysis of Variance Table:

The overall of the regression model is 23.46 and the corresponding p-value is <.0001.

Since this p-value is less than .05, we conclude that the regression model as a whole is statistically significant.

Model Fit Table:

The R-Square value tells us the percentage of variation in the exam scores that can be explained by the number of hours studied and the number of prep exams taken.

In this case, 73.4% of the variation in exam scores can be explained by the number of hours studied and number of prep exams taken.

The Root MSE value is also useful to know. This represents the average distance that the observed values fall from the regression line.

In this regression model, the observed values fall an average of 5.3657 units from the regression line.

Parameter Estimates Table:

We can use the parameter estimate values in this table to write the fitted regression equation:

Exam score = 67.674 + 5.556*(hours) -

.602*(prep_exams)

We can use this equation to find the estimated exam score for a student, based on the number of hours they studied and the number of prep exams they took.

For example, a student that studies for 3 hours and takes 2 prep exams is expected to receive an exam score of 83.1:

Estimated exam score = 67.674 + 5.556*(3) - .602*(2) = 83.1

The p-value for hours (<.0001) is less than .05, which means that it has a statistically significant association with exam score.

However, the p-value for prep exams (.5193) is not less than .05, which means it does not have a statistically significant association with exam score.

We may decide to remove prep exams from the model since it isn't statistically significant and instead perform using hours studied as the only predictor variable.

Additional Resources

The following tutorials explain how to perform other common tasks in SAS: