How do you perform Simple Linear Regression in SAS?

Authored by

stats writer

July 1, 2024

RECOMMENDED CITATION

stats writer (2024). *How do you perform Simple Linear Regression in SAS?*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=165033

ARABPSYCHOLOGY.COM

Simple Linear Regression is a statistical technique used to model the relationship between a single dependent variable and one or more independent variables. In SAS, this can be achieved by using the PROC REG procedure. The procedure allows the user to specify the dependent and independent variables, and provides various options to fit the regression model, including least squares, maximum likelihood, and robust methods. It also allows for the examination of the model's assumptions and diagnostics for checking the validity of the model. The output from PROC REG includes important model statistics such as the coefficient estimates, standard errors, and p-values, which can be used to interpret the relationship between the variables. By following the steps outlined in the PROC REG procedure, one can easily and accurately perform Simple Linear Regression in SAS.

Perform Simple Linear Regression in SAS

Simple linear regression is a technique that we can use to understand the relationship between one predictor variable and a response variable.

This technique finds a line that best "fits" the data and takes on the following form:

? = b0 + b1x

where:

?: The estimated response valueb0: The intercept of the regression lineb1: The slope of the regression line

This equation helps us understand the relationship between the predictor variable and the response variable.

ARABPSYCHOLOGY.COM

The following step-by-step example shows how to perform simple linear regression in SAS.

Step 1: Create the Data

For this example, we'll create a dataset that contains the total hours studied and final exam score for 15 students.

We'll to fit a simple linear regression model using *hours* as the predictor variable and *score* as the response variable.

The following code shows how to create this dataset in SAS:

```
/*create dataset*/
data exam_data;
input hours score;
datalines;
1 64
2 66
4 76
5 73
5 74
6 81
```

ARABPSYCHOLOGY.COM

```
6 83
7 82
8 80
10 88
11 84
11 82
12 91
12 93
14 89
;
run;

/*view dataset*/
proc printdata=exam_data;
```

ARABPSYCHOLOGY.COM

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: Fit the Simple Linear Regression Model

Next, we'll use proc reg to fit the simple linear regression model:

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

ARABPSYCHOLOGY.COM

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

ARABPSYCHOLOGY.COM

Analysis of Variance Table:

The overall of the regression model is 63.91 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. In other words, hours is a useful variable for predicting exam score.

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.

In general, the larger the of a regression model the better the predictor variables are able to predict the value of the response variable.

In this case, 83.1% of the variation in exam scores can be explained by the number of hours studied. This value is quite high, which indicates that hours studied is a highly useful variable for predicting exam score.

Parameter Estimates Table:

From this table we can see the fitted regression equation:

Score = 65.33 + 1.98*(hours)

We interpret this to mean that each additional hour studied is associated with an average increase of 1.98 points in exam score.

ARABPSYCHOLOGY.COM

The intercept value tells us that the average exam score for a student who studies zero hours is 65.33.

We can also use this equation to find the expected exam score based on the number of hours that a student studies.

For example, a student who studies for 10 hours is expected to receive an exam score of 85.13:

Score = 65.33 + 1.98*(10) = 85.13

Since the p-value (<.0001) for *hours* is less than .05 in this table, we conclude that it's a statistically significant predictor variable.

Step 3: Analyze the Residual Plots

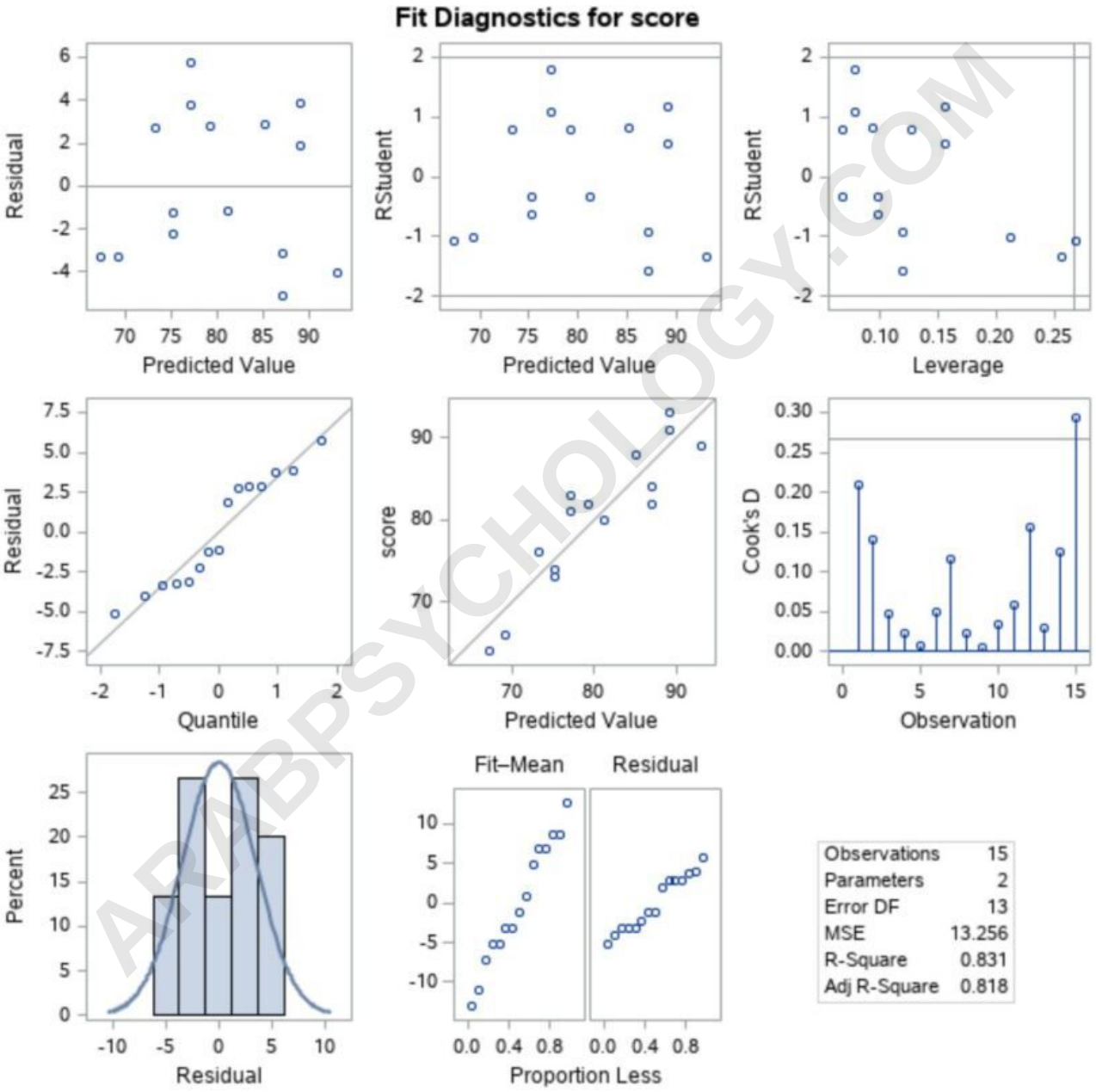
Simple linear regression makes two important about the of the model:

The residuals are normally distributed.The residuals have equal variance ("") at each level of the predictor variable.

If these assumptions are violated, then the results of our regression model can be unreliable.

ARABPSYCHOLOGY.COM

To verify that these assumptions are met, we can analyze the residual plots that SAS automatically in the output:



ARABPSYCHOLOGY.COM

To verify that the residuals are normally distributed, we can analyze the plot in the left position of the middle

row with "Quantile" along the x-axis and "Residual" along the y-axis.

This plot is called a , short for "quantile-quantile" plot, and is used to determine whether or not data is normally distributed. If the data is normally distributed, the points in a Q-Q plot will lie on a straight diagonal line.

From the plot we can see that the points fall roughly along a straight diagonal line, so we can assume that the residuals are normally distributed.

Next, to verify that the residuals are homoscedastic we can look at the plot in the left position of the first row with "Predicted Value" along the x-axis and "Residual" along the y-axis.

If the points in the plot are scattered randomly about zero with no clear pattern then we can assume that the residuals are homoscedastic.

From the plot we can see that the points are scattered about zero randomly with roughly equal variance at each level throughout the plot so we can assume that

ARABPSYCHOLOGY.COM

the residuals are homoscedastic.

Since both assumptions are met, we can assume that the results of the simple linear regression model are reliable.

Additional Resources

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

ARABPSYCHOLOGY.COM