How can I calculate Cook’s Distance in SAS?

Authored by

stats writer

June 23, 2024

Cook's Distance is a statistical measure used to identify influential data points in a regression analysis. It is calculated by measuring the change in the regression coefficients when a particular observation is removed from the data set. In SAS, Cook's Distance can be calculated using the COOKD function, which takes the fitted regression model and the data set as inputs. The resulting values can then be used to identify any outliers or influential data points that may have a significant impact on the overall regression analysis. This measure is important in identifying any potential issues with the data and ensuring the accuracy and validity of the regression results.

Calculate Cook's Distance in SAS

Cook's distance is used to identify influential in a regression model.

The formula for Cook's distance is:

Di = (ri2 / p*MSE) * (hii / (1-hii)2)

where:

ri is the ith residualp is the number of coefficients in the regression modelMSE is the mean squared errorhii is the ith leverage value

Essentially Cook's distance measures how much all of the fitted values in the model change when the ith observation is deleted.

The larger the value for Cook's distance, the more influential a given observation.

A rule of thumb is that any observation with a Cook's distance greater than 4/n (where *n* = total observations) is considered to be highly influential.

The following example shows how to calculate Cook's distance for each observation in a regression model in SAS.

Example: Calculating Cook's Distance in SAS

Suppose we have the following dataset in SAS:

```
/*create dataset*/
data my_data;
input x y;
datalines;
8 41
12 42
12 39
13 37
14 35
16 39
17 45
22 46
24 39
```

```
26 49
29 55
30 57
;
run;

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

Obs	x	y
1	8	41
2	12	42
3	12	39
4	13	37
5	14	35
6	16	39
7	17	45
8	22	46
9	24	39
10	26	49
11	29	55
12	30	57

We can use PROC REG to fit a to this dataset and then use the OUTPUT statement along with the COOKD statement to calculate Cook's distance for each observation in the regression model:

```
/*fit simple linear regression model and calculate Cook's distance for each obs*/
proc regdata=my_data;
model y=x;
output out=cooksData cookd=cookd;
run;

/*print Cook's distance values for each observation*/
proc printdata=cooksData;
```

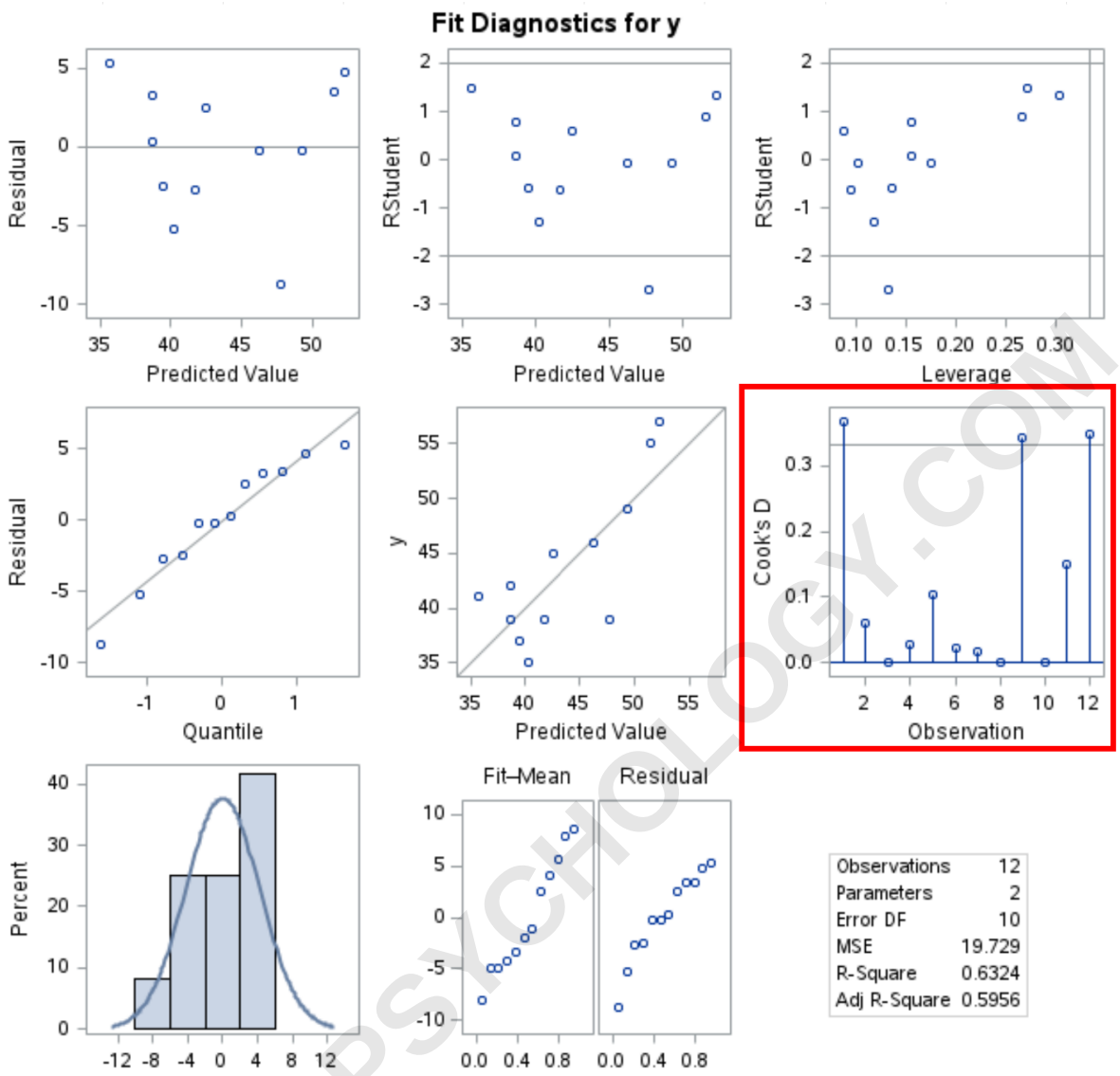
The final table in the output displays the original dataset along with Cook's distance for each observation:

Obs	x	y	cookd
1	8	41	0.36813
2	12	42	0.06075
3	12	39	0.00052
4	13	37	0.02764
5	14	35	0.10487
6	16	39	0.02155
7	17	45	0.01705
8	22	46	0.00020
9	24	39	0.34275
10	26	49	0.00047
11	29	55	0.15003
12	30	57	0.34948

Cook's distance for the first observation is 0.36813.Cook's distance for the second observation is 0.06075.Cook's distance for the third observation is 0.00052.

And so on.

The PROC REG procedure also produces several diagnostic plots in the output and the chart for Cook's distance can be seen in this output:



The x-axis shows the observation number and the y-axis shows Cook's distance for each observation.

Note that a cutoff line is placed at 4/n (in this case n = 12, thus the cutoff is at 0.33) and we can see that three observations in the dataset are greater than this line.

This indicates that these observations could be highly

influential to the regression model and should perhaps be examined more closely before interpreting the output of the model.

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