How can we test for multicollinearity in R?

Authored by

stats writer

June 26, 2024

Multicollinearity is a phenomenon that occurs when two or more independent variables in a regression model are highly correlated, making it difficult to accurately estimate the effect of each variable on the dependent variable. To test for multicollinearity in R, there are several methods that can be used. One method is to calculate the variance inflation factor (VIF) for each independent variable, which measures how much the variance of a particular variable is inflated due to collinearity with other variables. A high VIF value (typically above 5) indicates a strong correlation and potential multicollinearity. Another method is to perform a correlation matrix, which displays the correlation coefficients between all pairs of variables. A high correlation coefficient (typically above 0.7) suggests a potential issue with multicollinearity. Additionally, conducting a principal component analysis (PCA) can help identify the presence of multicollinearity by identifying groups of variables that are highly correlated. Overall, testing for multicollinearity in R helps ensure the accuracy and reliability of regression models.

Test for Multicollinearity in R

In regression analysis, occurs when two or more predictor variables are highly correlated with each other, such that they do not provide unique or independent information in the regression model.

If the degree of correlation is high enough between predictor variables, it can cause problems when fitting and interpreting the regression model.

The most straightforward way to detect multicollinearity in a regression model is by calculating a metric known as the variance inflation factor, often abbreviated VIF.

VIF measures the strength of correlation between predictor variables in a model. It takes on a value

between 1 and positive infinity.

We use the following rules of thumb for interpreting VIF values:

VIF = 1: There is no correlation between a given predictor variable and any other predictor variables in the model.VIF between 1 and 5: There is moderate correlation between a given predictor variable and other predictor variables in the model.VIF > 5: There is severe correlation between a given predictor variable and other predictor variables in the model.

The following example shows how to detect multicollinearity in a regression model in R by calculating VIF values for each predictor variable in the model.

Example: Testing for Multicollinearity in R

Suppose we have the following data frame that contains information about various basketball players:

```
#create data frame
df = data.frame(rating = c(90, 85, 82, 88, 94, 90, 76, 75, 87, 86),
```

```
points=c(25, 20, 14, 16, 27, 20, 12, 15, 14, 19),
assists=c(5, 7, 7, 8, 5, 7, 6, 9, 9, 5),
rebounds=c(11, 8, 10, 6, 6, 9, 6, 10, 10, 7))

#view data frame
df
```

rating	points	assists	rebounds
1	90	25	5	11
2	85	20	7	8
3	82	14	7	10
4	88	16	8	6
5	94	27	5	6
6	90	20	7	9
7	76	12	6	6
8	75	15	9	10
9	87	14	9	10
10	86	19	5	7

Suppose we would like to fit a using rating as the response variable and points, assists, and rebounds as the predictor variables.

To calculate the VIF for each predictor variable in the model, we can use the vif() function from the car

package:

library(car)

#define multiple linear regression model
model <- lm(rating ~ points + assists + rebounds, data=df)

#calculate the VIF for each predictor variable in the model
vif(model)

points assists rebounds
1.763977 1.959104 1.175030

We can see the VIF values for each of the predictor variables:

points: 1.76assists: 1.96rebounds: 1.18

Since each of the VIF values for the predictor variables in the model are close to 1, multicollinearity is not a problem in the model.

Note: If multicollinearity does turn out to be a problem in your model, the quickest fix in most cases is to

remove one or more of the highly correlated variables.