How to Detect and Address Multicollinearity with VIF in Regression Analysis

Authored by

stats writer

March 2, 2026

Multicollinearity refers to the presence of high correlation among independent variables in a regression model. This means that the independent variables are highly related to each other, making it difficult for the model to accurately determine their individual effects on the dependent variable. In order to measure the extent of multicollinearity, the Variance Inflation Factor (VIF) is commonly used in regression analysis. VIF calculates the ratio of the variance of a coefficient in a model with multiple independent variables, to the variance of that same coefficient in a model with only one independent variable. A high VIF value indicates a strong correlation between the independent variables, and values above 5 or 10 are generally considered problematic. By identifying and addressing multicollinearity, the accuracy and reliability of a regression model can be improved.

A Guide to Multicollinearity & VIF in Regression

Multicollinearity in regression analysis occurs when two or more predictor variables are highly correlated to 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 variables, it can cause problems when fitting and interpreting the regression model.

For example, suppose you run a regression analysis using the response variable*max vertical jump* and the following predictor variables:

heightshoe sizehours spent practicing per day

In this case, *height* and *shoe size* are likely to be highly

correlated with each other since taller people tend to have larger shoe sizes. This means that multicollinearity is likely to be a problem in this regression.

This tutorial explains why multicollinearity is a problem, how to detect it, and how to resolve it.

Why Multicollinearity is a Problem

One of the main goals of regression analysis is to isolate the relationship between each predictor variable and the response variable.

In particular, when we run a regression analysis, we interpret each regression coefficient as the mean change in the response variable, *assuming all of the other predictor variables in the model are held constant.*

This means we assume that we're able to change the values of a given predictor variable without changing the values of the other predictor variables.

However, when two or more predictor variables are highly correlated, it becomes difficult to change one variable without changing another.

This makes it difficult for the regression model to estimate the relationship between each predictor variable and the response variable independently because the predictor variables tend to change in unison.

In general, multicollinearity causes two types of problems:

The coefficient estimates of the model (and even the signs of the coefficients) can fluctuate significantly based on which other predictor variables are included in the model.The precision of the coefficient estimates are reduced, which makes the p-values unreliable. This makes it difficult to determine which predictor variables are actually statistically significant.

How to Detect Multicollinearity

The most common way to detect multicollinearity is by using the variance inflation factor (VIF), which measures the correlation and strength of correlation between the predictor variables in a regression model.

Utilizing the Variance Inflation Factor (VIF)

A value of 1 indicates there is no correlation between a given predictor variable and any other predictor variables in the model.A value between 1 and 5 indicates moderate correlation between a given predictor variable and other predictor variables in the model, but this is often not severe enough to require attention.A value greater than 5 indicates potentially severe correlation between a given predictor variable and other predictor variables in the model. In this case, the coefficient estimates and p-values in the regression output are likely unreliable.

For example, suppose we run a regression analysis using predictor variables *height*, *shoe* size, and *hours spent practicing per day* to predict *max vertical jump* for basketball players and receive the following output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*VIF*
Intercept	-15.26179784	3.612395656	-4.2248	0.002897	
height	4.32876917	0.926682091	4.6713	0.0016	12.33
shoe size	-0.674982676	0.176166878	-3.8315	0.005007	22.34
practice hours per day	0.72611633	0.087247634	8.3225	0.000004	1.08

From the last column, we can see that the VIF values for *height* and *shoe size* are both greater than 5. This indicates that they're likely suffering from

multicollinearity and that their coefficient estimates and p-values are likely unreliable.

If we look at the coefficient estimate for shoe size, the model is telling us that for each additional one unit increase in shoe size, the average increase in *max vertical jump* is -0.67498 inches, assuming height and practice hours are held constant.

This doesn't seem to make sense, considering we would expect players with larger shoe sizes to be taller and thus have a higher max vertical jump.

This is a classic example of multicollinearity causing the coefficient estimates to appear a bit whacky and unintuitive.

How to Resolve Multicollinearity

If you detect multicollinearity, the next step is to decide if you need to resolve it in some way. Depending on the goal of your regression analysis, you might not actually need to resolve the multicollinearity.

Namely:

1. If there is only moderate multicollinearity, you likely don't need to resolve it in any way.

2. Multicollinearity only affects the predictor variables that are correlated with one another. If you are interested in a predictor variable in the model that doesn't suffer from multicollinearity, then multicollinearity isn't a concern.

3. Multicollinearity impacts the coefficient estimates and the p-values, but it doesn't impact predictions or goodness-of-fit statistics. This means if your main goal with the regression is to make predictions and you're not concerned with understanding the exact relationship between the predictor variables and response variable, then multicollinearity doesn't need to be resolved.

If you determine that you *do* need to fix multicollinearity, then some common solutions include:

1. Remove one or more of the highly correlated variables. This is the quickest fix in most cases and is often an acceptable solution because the variables you're removing are redundant anyway and add little

unique or independent information the model.

2. Linearly combine the predictor variables in some way, such as adding or subtracting them from one way. By doing so, you can create one new variables that encompasses the information from both variables and you no longer have an issue of multicollinearity.

3. Perform an analysis that is designed to account for highly correlated variables such as **principal component analysis** or **partial least squares (PLS) regression**. These techniques are specifically designed to handle highly correlated predictor variables.