How to Calculate Variance Inflation Factor (VIF) in SAS

Authored by

stats writer

November 19, 2025

The Variance Inflation Factor (VIF) is a critical diagnostic tool used in regression analysis to detect and quantify the severity of multicollinearity. Multicollinearity arises when two or more predictor variables (independent variables) in a model are highly linearly correlated with each other. This high correlation means the variables are essentially providing redundant information, which severely complicates the interpretation and stability of the model's coefficients.

Specifically, the VIF measures how much the variance of an estimated regression coefficient is inflated due to the correlation among the predictor variables. A high VIF value indicates that the corresponding predictor variable can be largely explained by the other predictors in the model, leading to unstable coefficient estimates, larger standard errors, and reduced statistical power. Without tools like VIF, researchers might mistakenly interpret insignificant coefficients, even if the underlying model structure is strong. In SAS, VIF is efficiently calculated using the REG procedure with the VIF option.

1. Understanding Variance Inflation Factor (VIF)

In regression analysis, multicollinearity occurs when two or more predictor variables are highly correlated with each other, such that they do not provide unique or independent information regarding the response variable. This phenomenon is problematic because it inflates the variance of the estimated regression coefficients, making them highly sensitive to small changes in the data.

If the degree of correlation is high enough between variables, it can cause significant problems when fitting and interpreting the model. High correlation leads to wide confidence intervals and large p-values for the predictor variables, even if the overall model (as measured by the F-statistic or R-squared) is statistically significant. This obscures the true individual relationship between each predictor and the outcome.

One way to detect and quantify multicollinearity is by using the metric known as the **variance inflation factor (VIF)**. The VIF measures the strength of the linear relationship between a specific explanatory variable and all other explanatory variables in the model. This tutorial focuses on how to precisely calculate and interpret the VIF metric using SAS statistical software.

2. The Theoretical Basis of VIF

The VIF for a specific predictor variable is mathematically derived from the coefficient of determination (R-squared) obtained by regressing that variable against all the remaining predictor variables in the model. The formal calculation for the VIF of the j-th predictor variable is given by the formula: $VIFj = 1 / (1 - R^2j)$. Here, R^2j represents the R-squared value resulting from the auxiliary regression where the j-th predictor is the dependent variable.

When the j-th predictor variable is entirely independent of all other predictors, R^2j approaches zero, resulting in a VIF of exactly 1.0. As the linear relationship between the predictor variable and the others strengthens, R^2j increases towards 1.0, and consequently, the VIF value increases dramatically, potentially approaching infinity. This relationship highlights why even moderate increases in correlation can lead to significant inflation of the coefficient variance.

Understanding this theoretical basis confirms that VIF directly quantifies the extent of redundancy in the model's predictor space. A high VIF value signals that the corresponding variable's unique contribution to explaining the response variable is minimal because its variance is largely explained by the other inputs. Therefore, the VIF is an essential metric for diagnosing the structural integrity of the predictor set within any regression model.

3. Step-by-Step Example: Creating Sample Data in SAS

To illustrate the process of VIF calculation, we will first establish a working dataset in SAS. This example utilizes fictional data describing the performance attributes of ten basketball players. Our objective is to determine if multicollinearity exists when modeling player **rating** based on **points**, **assists**, and **rebounds**.

The dataset, named `my_data`, is created using the `DATA` and `DATALINES` steps, which is the standard method for inputting small datasets directly into the SAS environment. Following the data definition, we use `PROC PRINT` to visually confirm that the data has been loaded and structured correctly, ensuring readiness for the regression analysis.

```
/*create dataset*/
data my_data;
input rating points assists rebounds;
datalines;
90 25 5 11
85 20 7 8
82 14 7 10
88 16 8 6
94 27 5 6
90 20 7 9
76 12 6 6
75 15 9 10
87 14 9 10
86 19 5 7
;
run;
```

```
/*view dataset*/
proc print data=my_data;
```

The resulting table displayed by `PROC PRINT` confirms the successful creation and loading of the dataset, with `rating`, `points`, `assists`, and `rebounds` clearly defined for subsequent statistical modeling.

Obs	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

4. Executing the REG Procedure with the VIF Option

Suppose we would like to fit a multiple linear regression model using **rating** as the response variable and **points**, **assists**, and **rebounds** as the predictor variables. To fit this model and simultaneously assess collinearity, we utilize the REG procedure along with the essential **VIF** option.

The REG procedure is the standard SAS tool for least squares fitting. By adding the `/ VIF` statement immediately after defining the model equation, we instruct SAS to perform the internal diagnostics necessary to calculate the variance inflation factor for each predictor variable in the model.

```
/*fit regression model and calculate VIF values*/
proc reg data=my_data;
model rating = points assists rebounds / vif;
run;
```

Executing this code yields detailed output, including the standard ANOVA and Parameter

Estimates tables, supplemented by the crucial VIF column. This integrated approach minimizes coding effort and ensures that the diagnostics are calculated using the exact model specification applied during the fitting process.

The REG Procedure
Model: MODEL1
Dependent Variable: rating

Number of Observations Read	10
Number of Observations Used	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	207.99697	69.33232	3.30	0.0995
Error	6	126.10303	21.01717		
Corrected Total	9	334.10000			

Root MSE	4.58445	R-Square	0.6226
Dependent Mean	85.30000	Adj R-Sq	0.4338
Coeff Var	5.37450		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	62.47163	14.58822	4.28	0.0052	0
points	1	1.11933	0.41088	2.72	0.0345	1.76398
assists	1	0.88340	1.38067	0.64	0.5459	1.95910
rebounds	1	-0.42777	0.85101	-0.50	0.6331	1.17503

5. Interpreting the VIF Results from SAS Output

From the detailed **Parameter Estimates** table generated by `PROC REG`, we can observe the calculated VIF values for each of the predictor variables. These values are listed in the far right column of the table, confirming the presence and magnitude of potential multicollinearity.

For our specific model predicting player rating, the VIF values are:

points: 1.76398
assists: 1.96591

rebounds: 1.17503

It is important to acknowledge that the output also includes a VIF value for the "Intercept" term. However, the VIF for the Intercept is generally ignored in the context of diagnosing multicollinearity among independent variables, as it primarily reflects data centering rather than structural correlation issues between predictors.

These numerical results must now be compared against established statistical guidelines to determine if the level of correlation found is severe enough to compromise the statistical reliability of the coefficient estimates. Since all observed VIFs are close to 1, this suggests a robust model structure for our current example.

6. Established Guidelines for VIF Interpretation

The value for VIF starts at 1 and has no definitive upper limit. The interpretation relies on commonly accepted rules of thumb to classify the severity of the correlation. Understanding these thresholds is essential for determining the necessity of model modification.

A value of **1**: This ideal value indicates that there is absolutely no correlation between the given predictor variable and any other predictor variables in the model. The variable provides entirely unique information.
A value between **1** and **5**: This range indicates a moderate level of correlation. While variance inflation is present, it is often not severe enough to require attention or corrective action. Models with VIFs in this range are typically considered acceptable.
A value greater than **5**: This suggests potentially severe correlation between a given predictor variable and other predictor variables in the model. This threshold serves as a warning sign that the coefficient estimates and p-values in the regression model output may be becoming unreliable.
A value greater than **10**: In most statistical fields, this value signifies extremely high multicollinearity. If a variable exhibits a VIF above 10, remedial measures are strongly recommended to stabilize the parameter estimates.

7. Strategies for Addressing Severe Multicollinearity

If the REG procedure output confirms that multicollinearity is a significant problem (VIF > 5 or > 10) in your regression model, several common analytical techniques can be employed to deal with the issue effectively:

Remove one or more of the highly correlated variables.

This is frequently the quickest and most practical fix. Because the variables causing the inflation are largely redundant, removing one or more of them often does not significantly harm the overall

predictive power of the model but substantially improves the stability and interpretability of the remaining coefficients. Analysts should use theoretical knowledge or relative VIF scores to decide which redundant variable to exclude.

Linearly combine the predictor variables.

If the highly correlated variables measure aspects of the same underlying construct, they can be combined mathematically, such as through addition, subtraction, or averaging, to create a single, meaningful composite score. By doing so, you create one new variable that encompasses the information from both, eliminating the issue of high correlation between the original separate variables.

Utilize advanced statistical analyses.

If retaining all original variables is necessary due to study design or theoretical imperatives, consider employing specialized techniques designed to account for highly correlated variables. Methods such as Principal Component Regression (PCR) or Partial Least Squares (PLS) Regression are specifically structured to handle correlated predictor sets by transforming them into orthogonal components before estimation.