Calculating Variance Inflation Factor (VIF) in Excel: A Guide to Detecting Multicollinearity

Authored by

Mohammed looti

November 8, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Calculating Variance Inflation Factor (VIF) in Excel: A Guide to Detecting Multicollinearity*. PSYCHOLOGICAL STATISTICS. Retrieved from https://statistics.arabpsychology.com/?p=13560

Detecting Multicollinearity with the Variance Inflation Factor (VIF)

In the realm of regression analysis, a significant challenge known as Multicollinearity can dramatically compromise the integrity of statistical models. This issue arises when two or more independent inputs, commonly referred to as **predictor variables** or explanatory variables, exhibit a high degree of linear correlation with one another. When this strong inter-correlation exists, the variables essentially cease to provide unique or independent information to the overall model, leading to redundancy.

If the correlation between predictors is sufficiently high, it can severely complicate the process of model fitting and lead to significant interpretation difficulties. The most critical symptoms include **unreliable coefficient estimates**, meaning the contribution attributed to each predictor becomes highly unstable, and severely inflated standard errors. These inflated standard errors make it difficult to determine which predictors are statistically significant, hindering sound data-driven decision-making.

Fortunately, statisticians rely on a robust diagnostic metric to quantify and detect the presence and severity of this issue: the **variance inflation factor (VIF)**. The VIF is specifically designed to measure how much the variance of an estimated regression coefficient is increased due to the presence of multicollinearity with the other variables in the model. Essentially, a higher VIF value serves as a clear signal of stronger correlation between that specific explanatory variable and the rest of the predictors utilized within the model.

The Underlying Mathematics of the VIF Calculation

The core principle behind calculating the variance inflation factor (VIF) is rooted in the concept of **auxiliary regression**. For any specific explanatory variable we wish to test, its VIF value is derived by running a separate, secondary regression. In this auxiliary model, the predictor being tested is temporarily treated as the response variable, while all other predictors from the original model serve as the new explanatory variables.

The result of this auxiliary regression that is crucial for VIF calculation is the coefficient of determination (R-squared). This R-squared value quantifies how well the variable in question can be linearly predicted by the other predictors in the model. The VIF is then calculated using the following elegantly simple formula:

VIF = 1 / (1 - R2)

In this formula, R2 represents the R-squared value obtained specifically from the auxiliary regression. A high R2 value in this context implies that the predictor is highly redundant, resulting in a small denominator (1 - R2) and, consequently, a **high VIF score**. Understanding this relationship is critical, as the subsequent steps in Microsoft Excel require us to systematically calculate a distinct R-squared value for every predictor variable in our primary model.

Preparing the Dataset in Microsoft Excel

To effectively demonstrate the VIF calculation process, we will use a sample dataset detailing attributes for 10 basketball players. Our objective is to fit a regression analysis model where the player **rating** is the **response variable**, and **points**, **assists**, and **rebounds** are the explanatory variables. The foundational requirement for any statistical analysis in Excel is the meticulous organization of raw data.

The structured dataset, complete with clear column headers, must be prepared within an Excel worksheet, as illustrated below. It is essential to ensure that this data is correctly formatted and ready for processing by Excel's statistical computation tools, particularly the Data Analysis Toolpak.

rating	points	assists	rebounds
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

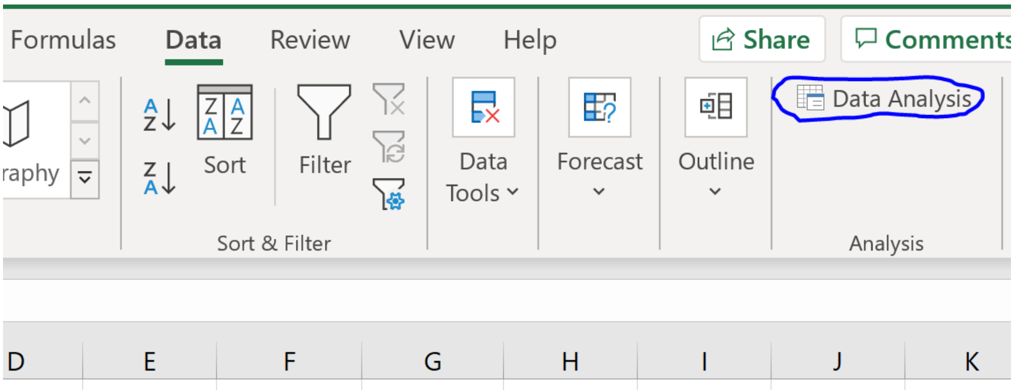
Once the data is accurately structured, we can proceed to the initial multiple linear regression. This step not only establishes the baseline model being investigated for multicollinearity but also confirms the functionality of the necessary Excel add-ins.

Step 1: Executing the Primary Regression Model

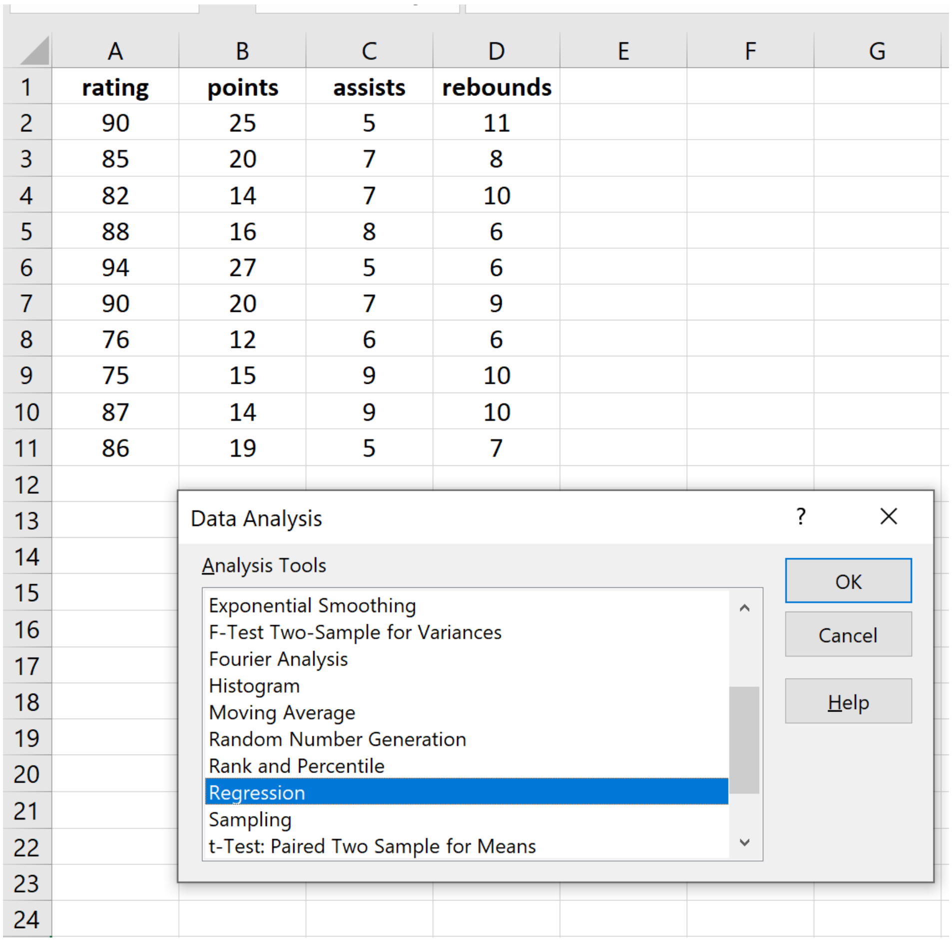
The first practical step in the VIF calculation workflow is to execute the primary **multiple linear**

regression model using the statistical capabilities built into Excel. This process is necessary to set up the context for the subsequent auxiliary regressions required for VIF.

To begin the analysis, navigate to the **Data** tab located on the top ribbon and select the **Data Analysis** option. If this option is unavailable, you must first enable the **Analysis Toolpak** add-in within your Excel settings, as it is indispensable for performing advanced statistical computations, including regression.

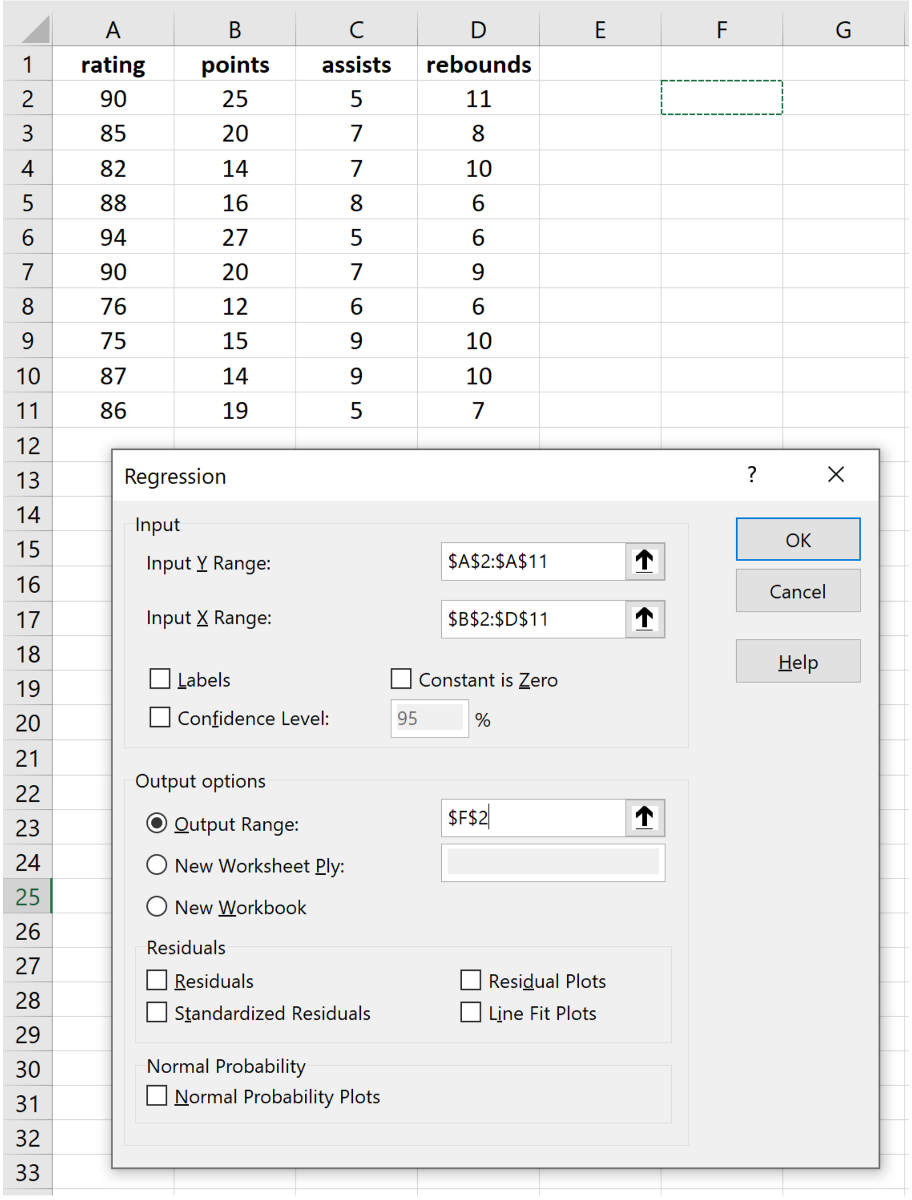


After launching the Data Analysis dialog box, select the *Regression* option from the list and click **OK**. This will prompt the Regression input window, where you must precisely define the dependent and independent variables for your model.



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

Carefully specify the **Y Input Range** (the response variable: Rating) and the **X Input Range** (the explanatory variables: Points, Assists, and Rebounds). Ensure the **Labels** box is checked if you included header rows in your selections, and designate an appropriate **Output Range** to display the results. Clicking **OK** will generate the summary output containing various summary statistics and coefficient estimates for the primary model.



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

While this initial output provides crucial information about the overall model performance, it is important to understand that Excel's native regression analysis feature does not directly calculate or display the VIF values. This initial step merely sets the stage for the iterative process of calculating the VIF for each predictor individually.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	rating	points	assists	rebounds										
2	90	25	5	11		SUMMARY OUTPUT								
3	85	20	7	8										
4	82	14	7	10		*Regression Statistics*								
5	88	16	8	6		Multiple R	0.789024							
6	94	27	5	6		R Square	0.622559							
7	90	20	7	9		Adjusted R	0.433839							
8	76	12	6	6		Standard E	4.584449							
9	75	15	9	10		Observatio	10							
10	87	14	9	10										
11	86	19	5	7		ANOVA								
12							*df*	*SS*	*MS*	*F*	*Significance F*			
13						Regression	3	207.997	69.33232228	3.298841616	0.099468483			
14						Residual	6	126.103	21.01717219					
15						Total	9	334.1						
16														
17							*Coefficients*	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
18						Intercept	62.47163	14.58822	4.28233534	0.005192295	26.77555401	98.16771	26.77555	98.16771046
19						X Variable	1.119326	0.410882	2.724201924	0.03445006	0.113933593	2.124719	0.113934	2.124719302
20						X Variable	0.883401	1.380667	0.639836482	0.545919049	-2.494969629	4.261772	-2.49497	4.261772073
21						X Variable	-0.42777	0.851009	-0.502663446	0.633113984	-2.510116271	1.654574	-2.51012	1.654573647
22														

Step 2: Conducting Auxiliary Regressions and Calculating VIF

To determine the **variance inflation factor (VIF)** for each predictor, we must run a separate auxiliary regression for every explanatory variable present in our original model. Since our basketball model includes three explanatory variables (Points, Assists, Rebounds), we are required to perform three distinct regressions. In each instance, one predictor from the original model is designated as the response variable (Y), and the remaining predictors serve as the independent variables (X).

The required auxiliary regressions are structured as follows:

To calculate the VIF for **Points**: Run a regression where Points is the response variable (Y) and Assists and Rebounds are the explanatory variables (X).
To calculate the VIF for **Assists**: Run a regression where Assists is the response variable (Y) and Points and Rebounds are the explanatory variables (X).
To calculate the VIF for **Rebounds**: Run a regression where Rebounds is the response variable (Y) and Points and Assists are the explanatory variables (X).

After running the first auxiliary regression (e.g., Points as Y), we must extract the critical piece of information from the output summary: the **R Square** value (or Coefficient of Determination), which quantifies the degree to which Points is predicted by the other variables. We then apply the VIF formula: VIF = 1 / (1 - R2).

For the auxiliary regression where *Points* is the response variable, the resultant output summary provides the necessary R Square value:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.658103							
R Square	0.433099							
Adjusted R Square	0.271128							
Standard Error	4.217166							
Observations	10							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	95.1086	47.55430124	2.67392057	0.137174859			
Residual	7	124.4914	17.78448536					
Total	9	219.6						
	Coefficients	*andard Errc*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	29.23513	7.614685	3.839309516	0.006380251	11.22926443	47.241	11.22926	47.24100421
X Variable 1	-2.21008	0.956692	-2.310128637	0.054176527	-4.472298511	0.052136	-4.4723	0.052135516
X Variable 2	0.481135	0.761416	0.63189522	0.54752935	-1.319327038	2.281597	-1.31933	2.281596884

Based on this output, the R Square value is 0.433099. We can then calculate the VIF for *Points*: VIF (Points) = 1 / (1 - R Square) = 1 / (1 - 0.433099) = 1 / 0.566901 ≈ **1.76**. This exact process must be repeated for both *Assists* and *Rebounds*. Upon completion of all three auxiliary regressions, the final VIF results for all explanatory variables are aggregated:

VIF for Points: **1.76**
VIF for Assists: **1.96**
VIF for Rebounds: **1.18**

Interpreting and Applying VIF Thresholds

The **VIF** value serves as a quantitative measure of the severity of multicollinearity in a model. VIF values begin at 1 and have no upper theoretical limit. Interpreting these results is crucial for diagnosing the overall statistical health of the regression analysis model and deciding whether corrective action is necessary. Statisticians generally rely on accepted rules of thumb for interpreting VIF scores:

A VIF value of **1** indicates a perfect scenario where the specific explanatory variable is entirely uncorrelated with all other predictors in the model. This signifies the absence of multicollinearity.
A VIF value between **1 and 5** suggests a moderate degree of correlation among the explanatory variables. While some correlation exists, it is typically not considered severe enough to significantly bias coefficient estimates or dangerously inflate **standard errors**. Models in this range are generally considered reliable.
A VIF value **greater than 5** (or sometimes 10, depending on the field of study) indicates potentially severe multicollinearity. When scores exceed this **threshold**, the coefficient estimates and

associated p-values derived from the primary model are likely **unreliable, unstable**, and highly sensitive to minor modifications in the dataset. Serious corrective measures, such as removing the highly correlated predictor or utilizing advanced statistical techniques like ridge regression, must be considered.

Based on the results obtained from our basketball player rating example, the calculated VIF values--Points (1.76), Assists (1.96), and Rebounds (1.18)--are all close to 1 and fall well below the commonly accepted conservative threshold of 5. Therefore, we can confidently conclude that multicollinearity is **not** a significant statistical concern for this specific model, and the coefficient estimates are likely robust.