**第十五课 MultiCollinearity In Linear Regression多重共线性**

Python零基础学习：[https://www.ixigua.com/home/77346806707?utm\_source=xiguastudio](https://www.ixigua.com/home/77346806707?utm_source=xiguastudio" \t "_blank)

源文件下载链接: <https://pan.baidu.com/s/1yuNlG6u9_C31fzhzbzqASA> 提取码: ebrh

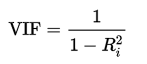
Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related. We have perfect multicollinearity if, for example as in the equation above, the correlation between two independent variables is equal to 1 or −1.

Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results

Variance Inflation Factor (VIF)

What is a Variance Inflation Factor?

A variance inflation factor(VIF) detects [multicollinearity](https://www.statisticshowto.com/multicollinearity/)in [regression analysis](https://www.statisticshowto.com/probability-and-statistics/regression-analysis/). Multicollinearity is when there’s [correlation](https://www.statisticshowto.com/probability-and-statistics/correlation-analysis/)between predictors (i.e. [independent variables](https://www.statisticshowto.com/independent-variable-definition/)) in a model; it’s presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

VIFs are usually calculated by software, as part of regression analysis. You’ll see a VIF column as part of the output. VIFs are calculated by taking a predictor, and regressing it against every other predictor in the model. This gives you the [R-squared](https://www.statisticshowto.com/adjusted-r2/) values, which can then be plugged into the VIF formula. “i” is the predictor you’re looking at (e.g. x1 or x2):  
[](https://www.statisticshowto.com/wp-content/uploads/2015/09/variance-inflation-factor.png)

**Question 1:** What classifies as too much correlation? For example, is a Pearson correlation of 0.5 too much?

There is no 'bright line' between not too much collinearity and too much collinearity (except in the trivial sense that r=1.0r=1.0 is definitely too much). Analysts would not typically think of r=.50r=.50 as too much collinearity between two variables. A rule of thumb regarding multicollinearity is that you have too much when the VIF is greater than 10 (this is probably because we have 10 fingers, so take such rules of thumb for what they're worth). The implication would be that you have too much collinearity between two variables if r≥.95

Don't use correlation coefficients. use VIFs of the model with all predictors and no interactions. VIFs of 5-10 are indicating too much correlation, your specific cutoff depends on what you need to do with the model.

**Question 2:** Can we fully determine whether there is collinearity between two variables based on the correlation coefficient or does it depend on other factors?

This depends on what you mean by "fully determine". If the correlation between two variables were r≥.95r≥.95, then most data analysts would say you had problematic collinearity. However, you can have multiple variables where no two variables have a pairwise correlation that high, and still have problematic collinearity hidden amongst the whole set of variables. This is where other metrics, such as the VIFs and condition numbers come in handy.