Ana Clara Tupinamba Freitas

Metro College of Technology, Data Science and Application

[anaclarat@gmail.com](mailto:anaclarat@gmail.com)

Abstract

The occurrence of Multicollinearity impacts the assumption in which the features in a model are independent of each other affecting estimates and compromise the power of a model, verifying and addressing its existence is simply done using softwares package like statsmodel and considering dropping variables.

Keywords: multicollinearity, VIF, python, bias, correlation.

Introduction

The occurrence of Multicollinearity impacts the assumption in which the features in a model are independent of each other affecting estimates and compromise the power of a model. Verifying its occurrence and addressing its existence is vital to create good and reliable models.

Methods

Consulting of papers and blogs about on python and Statistics.

Discussion

Multicollinearity refers to the occurrence of high correlation among predictive variables of a model. Its existence results in less reliable inferences (biased coefficient estimators and loss of power), becoming difficult to disentangle individual effects of the features in the model. It can occur especially when encoding categorical features.

Warning signs of multicollinearity

* A regression coefficient is not significant even though, theoretically, that variable should be highly correlated with Y.
* When you add or delete an X variable, the regression coefficients change dramatically.
* You see a negative regression coefficient when your response should increase along with X.
* You see a positive regression coefficient when the response should decrease as X increases.
* Your X variables have high pairwise correlations.

[[Enough Is Enough! Handling Multicollinearity in Regression Analysis, Minitab Blog Editor](https://blog.minitab.com/en/understanding-statistics/handling-multicollinearity-in-regression-analysis)]

How to identify

Chart, bar chart

Description automatically generatedIdentification of Multicollinearity can be done by:

* Plotting correlation plot with heatmap of all dependent features to verify the degree of correlation as figure 1 – Heatmap correlation. Although, this method don’t consider when multicollinearity arises from interaction between 3 or more features.

We can see that most of the features are not linearly correlated, although there is a high correlation value of 0.83 between Month (mnth) and Season.

Figure 1 – Heatmap correlation

* Another method is to calculate and evaluate the Variance Inflation Factor (VIF):

Step one

First we run an ordinary least square regression that has Xi as a function of all the other explanatory variables in the first equation.

If i = 1, for example, equation would be



Where is a constant and e is the error term.

Step two

Then, calculate the VIF factor for with the following formula :

Diagram

Description automatically generated with medium confidence

where R2 is the coefficient of determination of the regression equation in step one, with Xi on the left hand side, and all other predictor variables (all the other X variables) on the right hand side.

Step three

Analyze the magnitude of multicollinearity by considering the size of the VIF( ). A rule of thumb is that if VIF( ) > 10 then multicollinearity is high (a cutoff of 5 is also commonly used).

Some software instead calculates the tolerance which is just the reciprocal of the VIF. The choice of which to use is a matter of personal preference.

[[Variance inflation factor, Wikipedia](https://en.wikipedia.org/wiki/Variance_inflation_factor)]

Python, for example, provides via statsmodels package a function to easily calculate and return VIF:

from statsmodels.stats.outliers\_influence import variance\_inflation\_factor

#[…]

# For each X, calculate VIF and save in dataframe

vif = pd.DataFrame()

vif["VIF Factor"] = [variance\_inflation\_factor(X.values, i) for i in range(X.shape[1])]

vif["features"] = X.columns

[[Variance Inflation Factor (VIF) Explained, Ernest Tavares III](https://etav.github.io/python/vif_factor_python.html)]

Addressing Multicollinearity

Consider dropping a variable having in mind that there are exceptions as explained by Paul Allison ([When Can You Safely Ignore Multicollinearity?)](https://statisticalhorizons.com/multicollinearity):

* The variables with high VIFs are control variables, and the variables of interest do not have high VIFs.
* The high VIFs are caused by the inclusion of powers or products of other variables.
* The variables with high VIFs are indicator (dummy) variables that represent a categorical variable with three or more categories.

Conclusion

Consequences of multicollinearity can be preventable by evaluating your data and consider your features and goal and eventually dropping a variable.

References

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4318006/>

<https://www.investopedia.com/terms/m/multicollinearity.asp>

<https://blog.minitab.com/en/understanding-statistics/handling-multicollinearity-in-regression-analysis>

<https://en.wikipedia.org/wiki/Variance_inflation_factor>

<https://etav.github.io/python/vif_factor_python.html>

<https://statisticalhorizons.com/multicollinearity>