**Is Multicollinearity Necessarily Bad?**

Tushar Raj

Clark University  MASTERS IN DATA ANALYTICS

**Write a short report on why multicollinearity might not be always bad.**

If we are doing regression analysis for the sole purpose to forecast, then multicolanirity is not a big problem as the bigger the R square value, the better the prediction, but it is only possible when the explanatory variable has the same linear dependency the original data. The most widely used diagnostic for multicolanirity is the variance inflation factor (VIF), and below are the three scenarios in which the high VIF is not a problem and can be ignored.

1. Control variables have high VIF, and variables of interest do not have high VIF

Multicolanirity causes the problem to the variable which are colinear as it increases the coefficient and also makes them unstable, but if the colinear variable is just used as a control variable as the performance of the contrail variable as control is not impacted.

1. High VIF caused by the inclusion of product and other variables, In regression model for a and a square will high correlated, and similarly, we have a variable a and b and the product of a and b will be ab will be highly correlated with the variables. Thus we can significantly reduce the correlation of variables by centring the variable before creating the power.
2. The variables with high VIF are dummy variables in more the three category variables; consider a small category in that the cases will be high the indicator variable will have high VIF. Even though the category variable is not erlated to the other variables in the regression model.

**Support your report with an example**

A cardiac patient status variable has three categories: currently, a patient, never a patient, and was a patient earlier. If we choose was a patient earlier as the reference category, with indicator variables for the other two. Then the correlation between those two indicators gets more negative as the fraction of people in the reference category gets smaller. Consider this scenario, if 45 percent of people are never a patient, 45 percent are patient, and 10 percent are formerly were a ptient. Is this scenario? Well, it does mean that *p*-values for the indicator variables may be high. However, the comprehensive test that *all* indicators have coefficients of zero is unaffected by the high VIFs.

Moreover, nothing else in the regression is affected. If we want to avoid the high VIFs, we have to choose a reference category with a more significant fraction of the cases and not the cases which are insignificant. It may be desirable to avoid situations where none of the individual indicators is statistically significant even though the broad set of indicators is significant. – by paul allison

**References:**

##### When Can You Safely Ignore Multicollinearity? ----- by **PAUL ALLISON**

<https://statisticalhorizons.com/multicollinearity>

* Basic Economatrics --- by DamodarN. Gujarati
* [www.google.com](http://www.google.com)