27 December 2017
21:34

Multicollinearity

One of the assumptions of Classical Linear Regression Model is that there is no exact collinearity between the explanatory variables. If the explanatory variables are perfectly correlated, you will face with these problems:

- Parameters of the model become indeterminate
- Standard errors of the estimates become infinitely large

However, the case of perfect collinearity is very rare in practical cases. Imperfect or less than perfect multicollinearity is the more common problem and it arises when in multiple regression modelling two or more of the explanatory variables are approximately linearly related.

The consequences are:

- OLS estimators are still unbiased, but they have large variances and covariance's, making precise estimation difficult
- Multicollinearity tends to produce coefficients that are too far from population coefficients.
- As a result, the confidence intervals tend to be wider. Therefore, we may not reject the “zero null hypothesis” (i.e. the true population coefficient is zero).
- Model coefficient values with Negative sign will appear when we expect the Positive sign. Example: Sales, TV, Radio, Newspaper.
- The OLS estimators and their standard errors can be sensitive to small changes in the data.
- If we get multicollinearity then definitely you will see different results in model selection procedures like (Forward, Backward, Stepwise).

So, it is must to detect the collinearity as well as to remove them. The collinearity can be detected in the following ways:

- The easiest way for the detection of multicollinearity is to examine the correlation between each pair of explanatory variables. If two of the variables are highly correlated, then this may the possible source of multicollinearity. However, pair-wise correlation between the explanatory variables may be considered as the sufficient, but not the necessary condition for the multicollinearity.

- The second easy way for detecting the multicollinearity is to estimate the multiple regression and then examine the output carefully. The rule of thumb to doubt about the presence of multicollinearity is very high R2 but most of the coefficients are not significant according to their p-values. However, this cannot be considered as an acid test for detecting multicollinearity. It will provide an apparent idea for the presence of multicollinearity.

- As, the coefficient of determination in the regression of regressor Xj on the remaining regressors in the model, increases toward unity, that is, as the collinearity of Xj with the other regressors increases, VIF also increases and in the limit it can be infinite. Therefore, we can use the VIF as an indicator of multicollinearity. The larger the value of VIFj, the more “troublesome” or collinear the variable Xj. As a rule of thumb, if the VIF of a variable exceeds 10, which will happen if multiple correlation coefficient for j-th variable R^2_j exceeds 0.90, that variable is said to be highly collinear.
- $$VIF = \frac{1}{1-R_i^2} < 5 \text{ is ok,}$$

  if it is lies between 5-10 then it is caution. If it >10 then it is serious problem.

- The Farrar-Glauber test (F-G test) for multicollinearity is the best way to deal with the problem of multicollinearity.

  The ‘mctest’ package in R provides the Farrar-Glauber test and other relevant tests for multicollinearity. There are two functions viz. ‘omcdiag’ and ‘imcdiag’ under ‘mctest’ package in R which will provide the overall and individual diagnostic checking for multicollinearity respectively.

Correcting Multicollinearity

- Remove one of highly correlated independent variable from the model.
  If you have two or more factors with a high VIF, remove one from the model.
- Principle Component Analysis (PCA) - It cut the number of interdependent variables to a smaller set of uncorrelated components. Instead of using highly correlated variables, use components in the model that have eigenvalue greater than 1.
- Ridge Regression - It is a technique for analyzing multiple regression data that suffer from multicollinearity.
- If you include an interaction term (the product of two independent variables), you can also reduce multicollinearity by "centering" the variables. By "centering", it means subtracting the mean from the independent variables values before creating the products.