Multicollinearity and Regression Analysis

Multicollinearity and Regression Analysis

Jamal I. Daoud
Department of Science in Engineering, IIUM, 53100, Jalan Gombak,
Selangor Darul Ehsan, Malaysia

E-mail: jamal58@iium.edu.my

Abstract: In regression analysis it is obvious to have a correlation between the response and predictor(s), but having correlation among predictors is something undesired. The number of predictors included in the regression model depends on many factors among which, historical data, experience, etc. At the end selection of most important predictors is something objective due to the researcher. Multicollinearity is a phenomena when two or more predictors are correlated, if this happens, the standard error of the coefficients will increase [8]. Increased standard errors means that the coefficients for some or all independent variables may be found to be significantly different from 0. In other words, by overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they should be significant. In this paper we focus on the multicollinearity, reasons and consequences on the reliability of the regression model.

1. Introduction

In regression analysis there are many assumptions about the model, namely, multicollinearity, nonconsistant variance (non-homogeneity), linearity, and autocorrelation [6]. If one or more assumption is violated, then the model in hand is no more reliable and also is not acceptable in estimating the population parameters.

In this study we focus on multicollinearity as a violation of one of basic assumption for successful regression model assumptions of successful regression model. Multicollinearity appears when two or more independent variables in the regression model are correlated. a little bit of multicollinearity sometimes will cause big problem but when it is moderate of high then it will be a problem to be solved.

Multicollinearity, or near-linear dependence, is a statistical phenomenon in which two or more predictors variables in a multiple regression model are highly correlated. If there is no linear relationship between predictor variables, they are said to be orthogonal [2].

In most applications of regression, the predictor's variables are usually not orthogonal. Multicollinearity can be observed in the following cases

i) Large changes in the estimated coefficients when a variable is added or deleted.
ii) Large changes in the coefficients when a data point is altered or dropped.

Multicollinearity may be present if:

i) The algebraic signs of the estimated coefficients do not conform to the prior expectation; or
ii) Coefficients of variables that are expected to be important have large standard errors (small t-values).

In fact, the researcher has no tools to know the multicollinearity unless the data has been collected. There are two types of multicollinearity:

i) Data-based multicollinearity which occurs because of the researcher, when the experiment is poorly designed, or the collected data are purely observational
ii) Structural multicollinearity: it occurs when the researcher generates new independent variable from one or more existing variables, for example creating x^3 from x, it is in fact mathematical artifact which leads to multicollinearity.

Therefore, in this research we will focus on the impact of multicollinearity existence among predictor variables on hypotheses testing decision taken.

2. Correlation of predictors and the impact on regression model

What impact does the correlation between predictors have on the regression model and subsequent conclusions? Correlation can be high or otherwise, to illustrate the impact of correlation among predictors on the reliability of the model obtained we use two sets of data one set with low correlation among predictors and other set with high correlation between predictors. The analysis of regression for the first set of data yielded the following regression information.

We start by fitting simple models with one predictor variable each time, then by fitting multiple model containing both predictor variables. The multiple regression model found include both variables the correlation coefficients between two predictors was very low (-0.038). Results are shown in Table (1)

Table-1: Coefficients of models for first data set

Term	Coef	SECoef	t-value	P-value	VIF
X1	2.20	3.30	0.67	0.524	1.00
X2	0.354	0.638	0.55	0.594	1.00
X1	2.27	3.46	0.66	0.523	1.00
X2	0.371	0.662	0.56	0.593	1.00

From the Table (1), we can see that t-value for x_1 when included alone in the model is not far from the t-value when both predictors were included, same for x_2 the t-value if not different from its value when both predictors are included. The decision of the parameter test will be same. On the other hand the standard error of coefficients have not been changed dramatically, for x_1 from 3.3 to 3.46 for simple and multiple model, and for x_2 from 0.638 to 0.662. All this can be viewed as a result of low correlation between predictors

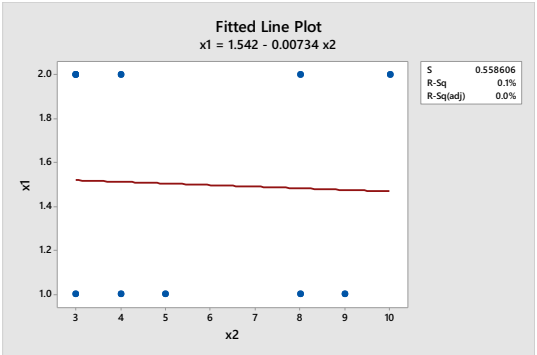
Now for the second set of data which is showing a very high correlation between predictors (0.996) the summary of the analysis are shown in Table (2).

Table-2: Coefficients of models for first data set

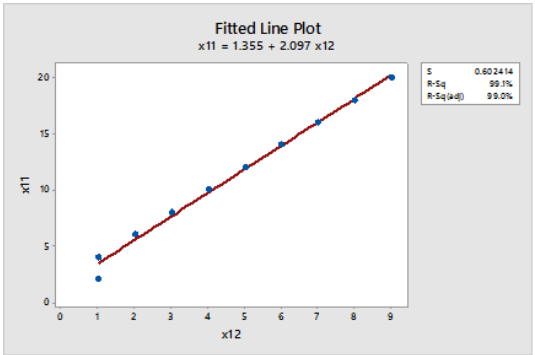
Term	Coef	SECoef	t-value	P-value	VIF
X11	-0.309	0.279	-1.11	0.299	1.00
X21	-0.704	0.579	-1.22	0.258	1.00
X11	2.71	2.96	0.92	0.390	113.67
X21	-6.39	6.24	-1.02	0.340	113.67

Also, from Table 2 we can see that, there is a huge change in the coefficients values for x_{11} from (-0.309) to (2.71) in simple and multiple models. In addition to that, we can notice that the standard errors for coefficients have been increased hugely as well, for x_1 from (0.279) to (2.96) and for x_{21} from (0.579) to (6.24) when compare simple and multiple models.

The relation between predictor variables for first and second sets of data are shown below.



Graph-1: Predictors with low correlation



Graph-2 Predictors with high correlation

This comparison evident on the impact of correlation on the standard error of coefficients was huge in changing the standard errors of the coefficients for the second set of data, which will lead to wrong conclusions on the model. Some or all predictors will become insignificant when they should be significant because of inflation in standard error for predictors coefficients. As a summary having the high correlation among predictors will prevent the researcher capturing the most influential predictors for inclusion in the model.

3. Diagnostic of Multicollinearity

There are many signs in the analysis for the multicollinearity among which

- The correlation among predictors is large
- In case if the correlation is not calculated the following are signs of having the multicollinearity:
  i) When the predictor's coefficients vary from one to another model.
  ii) When applying t-test, the coefficient are not significant but put all together ($F-test$) for the whole model it is significant.

Relying only on correlation between pairs of predictors has limitation, the small or large value of correlation is something subjective depends on individual and also on the field of research that is why most of the time to detect the multicollinearity we use some indicator called variance inflation factors (VIF).

4. Variance Inflation Factors (VIF)

When correlation exists among predictor's the standard error of predictors coefficients will increase and consequently the variance of predictor's coefficients are inflated. The VIF is a tool to measure and quantify how much the variance is inflated. $VIFs$ are usually calculated by the software as part of regression analysis and will appear in VIF column as part of the output. To interpret the value of VIF the following rule is used in the table below:

Table-3: *VIF* interpretation

VIF -value	conclusion
$VIF = 1$	Not correlated
$1 < VIF \leq 5$	Moderately correlated
$VIF > 5$	Highly correlated

In addition to the meaning of VIF itself in showing whether the predictors are correlated, the square roof of VIF indicates how much larger the standard error is, for example if $VIF = 9$ this means that the standard error for the coefficient of that predictor is 3 times as large as it would be if that predictor is uncorrelated with other predictors. VIF can be calculated using the formula:

$$VIF = \frac{1}{1-R_i^2}$$

This VIF can be calculated for each predictor in the model, and the way is to regress the variable assume it is i^{th} variable against all other predictors. We obtain R_i^2 which can be used to find VIF , same thing can be applied to all other predictors.

Back to results of our data analysis, from Table-1, we can see that the value of variance inflation factor for x_1 was 1 for both simple and multiple models, same for x_2 unchanged and it was 1, this is due to a very

low correlation for first set of data, while for second set of data, *VIF* for both variables were changed from 1 for simple model to 113.67 for multiple model. In the latter case we cannot commence with regression unless this problem is solved [3].

5. Problem solving

When two or more predictors are highly correlated, the relationship between the independent variables and the dependent variables is distorted by the very strong relationship between the independent variables, leading to the likelihood that our interpretation of relationships will be incorrect. In the worst case, if the variables are perfectly correlated, the regression cannot be computed [4].

Multicollinearity is detected by examining the tolerance for each independent variable. Tolerance is the amount of variability in one independent variable that is no explained by the other independent variables, and it is in fact $1-R^2$ .Tolerance values less than 0.10 indicate collinearity.

If we discover collinearity in the regression output, we should reject the interpretation of the relationships as false until the issue is resolved [3].

Multicollinearity can be resolved by combining the highly correlated variables through principal component analysis, or omitting a variable from the analysis that associated with other variable(s) highly.

6. Conclusions

1. Multicollinearity is one of serious problems that should be resolved before starting the process of modeling the data**.**
2. It is highly recommended that all regression analysis assumption should be met as they are contributing to accurate conclusion and helps to make inference on the population.
3. Ignore and dismiss the model if the multicollinearity discovered after finding the model specially with high correlation as the model cannot be interpreted.

References

[1] Carl F. Mela* and Praveen K. Kopalle. The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations. J. of Applied Economics, 2002, 43,667-677.
[2] D.R.Jensen and D.E. Ramirez. Variance Inflation in Regression, Advances in Decision Sciences, 2012, 2013, 1-15.
[3] Debbie J. Dupuis1 and Maria-Pia Victoria-Feser. Robust VIF regression with application to variable Selection in large data sets, The Annals of Applied Statistics, 2013, 7,319-341.
[4] George A. Milliken, Dallas E. Johnson (2002). Analysis of Messy data, Vol.3, Chapman & Hall/CRC
[5] Golberg, M. Introduction to regression analysis, Billerica, MA: Computational Mechanics Publications, 2004, 436.
[6] Jason W. Osborne and Elaine Waters (2002). Four Assumptions of Multiple Regressions that Researchers should always Test. J. of Practical Assessment, Research, and Evaluation. Vol.8, No.2, PP1-5.
[7] Kleinbaum, David G. Applied regression analysis and other multivariable methods, Australia;

Belmont, CA: Brooks/Cole, 2008,906.
[8] McClendon, McKee J. Multiple regression and causal analysis, Prospects Heights, Ill.: Waveland Press 2002,358.
[9] Seber, G. A. F. (George Arthur Frederick). Linear regression analysis, Hoboken, N.J.: Wiley-Interscience, 2003,557.