Journal Home: www.biospub.com

Popular Article

Article: RT1005

Biotica Research Today Vol 4:6 2022

426/428

How to cite this article?

Chittora *et al.*, 2022. Multicollinearity: A Problem in Multiple Linear Regression. Biotica Research Today 4(6):426-428.

Multicollinearity: A Problem in Multiple Linear Regression

Vaibhav Chittora[1*], Heerendra Prasad[1], Prashant Vasishth[2] and Mohit Sharma[2]

[1]Dr. YSPUHF, Nauni, Solan, Himachal Pradesh (173 230), India

[2]ICAR-Indian Agricultural Research Institute, Pusa, New Delhi, Delhi (110 012), India

Open Access

Corresponding Author

Vaibhav Chittora

e-mail: vchittora97@gmail.com

Keywords

Correlation, Matrix, MLR, VIF

Abstract

In regression analysis it is obvious to have a relation between the response and regressor(s) variables, but having linear relation among regressor variables is an undesired thing. Multicollinearity refers to the linear relation among two or more variables. If this happens, the standard error of the coefficients will increase. It is a data problem that may cause serious difficulty with the reliability of the estimates of the model parameters. Multicollinearity makes some variables statistically insignificant when they should be significant. In this article, we focus on the multicollinearity, reasons, and consequences of the reliability of the regression model.

Introduction

In linear regression when the study variable has a functional relationship with more than one explanatory or independent variable, then the functional form is termed as multiple regression model (Montgomery *et al*., 2014).

Where the linear multiple models can be expressed as:

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p + \in$$

Where,

Y = Dependent variable

X_1 through X_p = Independent variable

b_1 through b_p = Regression coefficients

Multicollinearity

Multicollinearity (also collinearity) is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.

In this situation, the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data.

If the degree of correlation between the independent variables is high enough, it can cause problems when we fit the model and interpret the results.

The presence of multicollinearity indicates high pseudo predictive power. It will have non-significant individual predictors with high standard errors.

Detection of Multicollinearity

In the presence of multicollinearity, usually coefficient of regression will have high standard error and occasionally the sign of the regression coefficient will not be in the expected lines.

In the individual *t*-tests for each of the slopes, many are non-

Article History

Received on: 31st May 2022

Revised on: 12th June 2022

Accepted on: 13th June 2022

E-mail: *bioticapublications@gmail.com*

significant (P > 0.05), but the overall *F*-test for testing all of the slopes are simultaneously significant (P < 0.05).

The correlations among pairs of independent variables are large.

Some Numerical Techniques for Detecting Multicollinearity

i) Variance Inflation Factor (VIF)

- $VIF = \frac{1}{1-R_i^2}$, I = 1, 2, 3, ... n

- The variance inflation factor (VIF) indicates the correlation between independent variables.
- It is used to explain how much amount of multicollinearity exists in a regression analysis.
- VIFs range from 1 to infinity. A value of 1 indicates that there is no multicollinearity among the regressor.
- VIFs between 1 and 5 suggest that there is a moderate correlation.
- VIFs greater than 5 represent critical levels of multicollinearity hence the coefficients are poorly estimated, because of which it might render other significant variables redundant (Daoud, 2017).

ii) Condition Number

- Compute the condition number. That is the ratio of the largest to the smallest characteristic root of the matrix x'x.
- If the condition number is less than 100 it is considered as non-harm full multicollinearity.
- Condition number between 100-1000 is considered moderate multicollinearity.
- If it is greater than 1000 then it's considered as severe or strong multicollinearity.

Why is Multicollinearity a Potential Problem?

- A key goal of regression analysis is to isolate the relationship between each independent variable and the dependent variable.
- The interpretation of a regression coefficient is that it represents the mean change in the dependent variable for each 1 unit change in an independent variable when we hold all of the other independent variables constant. This is very crucial for discussion about multicollinearity.
- However, when independent variables are correlated, it indicates that changes in one variable are associated with shifts in another variable.
- The stronger the correlation, the more difficult it is to change one variable without changing another.
- It becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable because the independent variables tend to change in unison (Greene, 2003).

Causes of Multicollinearity

Multicollinearity causes the following types of problems:

- The coefficients become very sensitive to small changes in the model.
- Multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of the regression model.
- We might not be able to trust the p-values to identify independent variables that are statistically significant.
- It produces parameter estimates of the "incorrect sign" and of implausible magnitude.

Effects of Multicollinearity

- Multicollinearity practically inflates unnecessarily the standard errors of the coefficients.
- In other words, by over-inflating the standard errors, it makes some variables statistically insignificant when they should be significant.
- Without multicollinearity (that is, with lower standard errors), those coefficients might be significant.
- Estimates of standard errors and parameters tend to be sensitive to changes in the data and the specification of the model.

How to Deal with Multicollinearity

- Re specify the model by eliminating one or more of the independent variables that are highly correlated with the other independent variables.
- Linearly combine the independent variables, such as adding them together.
- Using step-wise regression to select variables, leads to a model that is not theoretically well-motivated.
- Ridge regression is yet another technique used to overcome the problem of collinearity. This technique biases the estimated regression coefficients but reduces the level of multicollinearity (Gujarati, 2005).

Conclusion

Multicollinearity is one of the major issues that ought to be settled before beginning the modelling of the data. It is extremely suggested that all the assumption of regression analysis must be met as they are contributing to truthful conclusion and supports making inference about the population.

References

Montgomery, D.C., Peck, E.A., Vining, G.G., 2014. Introduction to linear regression analysis. WILEY, Singapore, p. 325.

Greene, W.H., 2003. Econometric Analysis. Prentice Hall, New Jersey, p. 56.

Gujarati, D.N., 2005. Basic Econometrics. McGraw-Hill, New York, p. 341.

Daoud, J.I., 2017. Multicollinearity and Regression Analysis. *Journal of Physics: Conf. Ser.* 949, 012009.