Volume 8, Issue 12, December 2023 ISSN (online): 2456-0006
International Journal of Science Technology Management and Research
Available online at: www.ijstmr.com

Treatment Of Multicollinearity Problem To Identify Effect Of Independent Variables

Twinkal Manawat
M Tech 4th Sem
Computer Science and Engineering Department
LNCT(Bhopal) Indore Campus
Indore M.P. India

Mr Nilesh Avinash Joshi
Assistant professor
Computer Science and Engineering Department
LNCT(Bhopal) Indore Campus
Indore M.P. India

Abstract:- ***Multicollinearity can be a problem in a regression model because we would not be able to distinguish between the individual effects of the independent variables on the dependent variable. Multicollinearity may not affect the accuracy of the model as much. But we might lose reliability in determining the effects of individual features in model and that can be a problem when it comes to interpretability. Multicollinearity is the presence of high correlations between two or more independent variables (predictors). Correlation is the association between variables, and it tells us the measure of the extent to which two variables are related to each other. Two variables can have positive (change in one variable causes change in another variable in the same direction), negative (change in one variable causes change in another variable in the opposite direction), or no correlation. A simple example of positive correlation can be weight and height. A simple example of a negative correlation can be the altitude and oxygen level.***
Keywords: Multicollinearity, Correlation , Predictors, Response R squared, Variance Inflation Factor

I. INTRODUCTION

Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a regression model. For example, height and weight, household income and water consumption, mileage and price of a car, study time and leisure time, etc. For example, from our everyday life to explain this. Colin loves watching television while munching on chips. The more television he watches, the more chips he eats and the happier he gets! Now, if we could quantify happiness and measure Colin's happiness while he's busy doing his favorite activity, we think would have a greater impact on his happiness? Having chips or watching television? That's difficult to determine because the moment we try to measure Colin's happiness from eating chips, he starts watching television. And the moment we try to measure his happiness from watching television, he starts eating chips. Eating chips and watching television are highly correlated in the case of Colin and we cannot individually determine the impact of the individual activities on his happiness. This is the multicollinearity problem. Multicollinearity can be a problem in a regression model because we would not be able to distinguish between the individual effects of the independent variables on the dependent variable. Correlation is the association between variables and it tells us the measure of the extent to which two variables are related to each other. Two variables can have positive (change in one variable causes change in another variable in the same direction), negative (change in one variable causes change in another variable in the opposite direction), or no correlation. It is easy to remember these terms if we keep some examples in our minds. A simple example of positive correlation can be weight and height. The taller you are, the heavier you weigh (this is considered a general trend if we leave the exception case aside).

1. A simple example of a negative correlation can be the altitude and oxygen level. The higher you go, the lower the oxygen level is.
2. A simple example of no correlation can be the depth of the sea and the number of apples bought from the store. None of them is related to the other.

Simply put, we can say that multicollinearity occurs when two or more predictors in regression analysis are highly related to one another. For example, the level of education and annual income. It is generally considered that the more educated you are, the more you earn. Thus, one variable can be easily predicted using another variable. If we keep both these variables in our analysis, it can cause problems for our model.

TYPES OF MULTICOLLINEARITY

There are two basic kinds of multicollinearity:

1. ***Structural multicollinearity*:** This type of multicollinearity is caused by the researchers (people like us) who create new predictors using the given predictors in order to solve the problem. For example, the creation of variable x^2 from the predictor variable x. Thus, this type of multicollinearity is a byproduct of the model we specify and not present in the data itself.
2. ***Data multicollinearity*:** This type of multicollinearity is the result of poorly designed experiments that are purely observational. Thus, it is present in the data itself and has not been specified/designed by us.

In a few places, you might come across terms like perfect multicollinearity and high multicollinearity as the two different types of multicollinearities.

3. ***Perfect multicollinearity*** occurs when two or more independent predictors in a regression model exhibit a perfectly predictable (exact or no randomness) linear relationship. The correlation, in this case, is equal to +1 or -1. For example, weight in pounds and weight in kilograms. However, we rarely face issues of perfect multicollinearity in a dataset.
4. ***High/Imperfect/Near multicollinearity*** occurs when two or more independent predictors are approximately linearly related. This is a common type and is problematic to us. All our analyses are based on detecting and dealing with this type of multicollinearity.

CAUSES OF MULTICOLLINEARITY

Multicollinearity could occur due to the following problems:

1. Multicollinearity could exist because of the problems in the dataset at the time of creation. These problems could be because of poorly designed experiments, highly observational data, or the inability to manipulate the data. For example, determining the electricity consumption of a household from the household income and the number of electrical appliances. Here, we know that the number of electrical appliances in a household will increase with household income. However, this cannot be removed from the dataset
2. Multicollinearity could also occur when new variables are created which are dependent on other variables. For example, creating a variable for BMI from the height and weight variables would include redundant information in the model.
3. Including identical variables in the dataset. For example, including variables for temperature in Fahrenheit and temperature in Celsius.
4. Inaccurate use of dummy variables can also cause a multicollinearity problem. This is called the Dummy variable trap. For example, in a dataset containing the status of marriage variable with two unique values: 'married', 'single'. Creating dummy variables for both of them would include redundant information. We can make do with only one variable containing 0/1 for 'married'/'single' status.
5. Insufficient data in some cases can also cause multicollinearity problems.

Some more reasons of multicollinearity can occur when developing a regression model?

1. Inaccurate use of different types of variables
2. Poor selection of questions or null hypothesis
3. The selection of a dependent variable
4. Variable repetition in a linear regression model
5. A high correlation between variables – one variable could be developed through another *varia*
6. Poor usage and choice of dummy variables

II. LITERATURE SURVEY

In 2017 Jamal I. Daoud et al proposed "Multicollinearity and Regression Analysis". In regression analysis it is obvious to have a correlation between the response and predictor(s), but having correlation among predictors is something undesired. Increased standard errors mean that the coefficients for some or all independent variables may be found to be significantly different from 0. They focused on multicollinearity, reasons and consequences on the reliability of the regression model. 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 [1].

In 2018 Neeraj Tiwari et al proposed "Diagnostics of Multicollinearity in Multiple Regression Model for Small Area Estimation" They discussed the multicollinearity problem in regression models for small area estimation and propose Ridge Regression Model (RRM) to deal with the problem of multicollinearity. The proposed model has been empirically compared with the existing Multiple Linear Regression (MLR) Model. Analysis of data obtained from a survey carried out from Directorate of Economics and Statistics, Uttarakhand, India revealed that RR methodology performs better as compared to MLR model in terms of the criterion of MSE[2].

In 2019 Alhassan Umar et al proposed "Detection of Collinearity Effects on Explanatory Variables and Error Terms in Multiple Regressions". They investigated the effects and consequences of multicollinearity on both standard error and explanatory variables in multiple regression, the correlation between X1 to X6 (independent variables) measure their individual effect and performance on Y (Response variable) and it is carefully observes how those explanatory variables inter correlated with one another and to the response variable. There are many procedures available in literature for detecting presence, degree and severity of multicollinearity in multiple regression analysis here they used correlation analysis to discover it is presence; we use variance inflation factors, tolerance level, indices number, eigenvalues to access fluctuation and influence of multicollinearity present in the model[3].

In 2019 N. A. M. R. Senaviratna et al proposed" Diagnosing Multicollinearity of Logistic Regression Model". One of the key problems arises in binary logistic regression model is that explanatory variables being considered for the logistic regression model are highly correlated among themselves. Multicollinearity will cause unstable estimates and inaccurate variances that affect confidence intervals and hypothesis tests. They discussed some diagnostic measurements to detect multicollinearity namely tolerance, Variance Inflation Factor (VIF), condition index and variance proportions. The adapted diagnostics are illustrated with data based on a study of road accidents[4].

In 2020 Solly Matshonisa at al proposed "Dealing with Multicollinearity in Regression Analysis: A Case in Psychology". In regression analysis, the main interest is to predict the response variable using the exploratory variables by estimating parameters of the linear model. In reality, the exploratory variables may share similar characteristics. This interdependency between the exploratory variables is called multicollinearity and causes parameter estimation in regression analysis to be unreliable. Different approaches to address the multicollinearity problem in regression modelling include variable selection, principal component regression and ridge regression[5].

In 2021 Mariella Gregorich et al proposed "Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution". Regression models have been in use for decades to explore and quantify the association between a dependent response and several independent variables in environmental sciences, epidemiology and public health. However, researchers often encounter situations in which some independent variables exhibit high bivariate correlation, or may even be collinear. They demonstrated how diagnostic tools for collinearity or near-collinearity may fail in guiding the analyst. Instead, the most appropriate way of handling collinearity should be driven by the research question at hand and, in particular, by the distinction between predictive or explanatory aims[6].

In 2022 Katrina I. Sundus , Bassam "Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to a fully annotated breast cancer dataset" . In this study, we presented a novel, fully-annotated national breast cancer dataset built from the cancer database registry of King Hussein Cancer Center, a medical center in Amman, Jordan, to predict recurrent breast cancer cases. Initially, the dataset had 35 attributes and 7562 instances of patients diagnosed with breast cancer between 2006 and 2021. Although this is still an ongoing project, research on breast cancer by the international community may benefit from the JBRCA dataset. The dataset can be used in its current state to predict breast cancer and recurrent breast cancer cases[7].

In 2023 Amin Otoni Harefa , Yulisman Zega , Ratna Natalia The Application of the Least Squares Method to Multicollinear Data Regression analysis is an analysis that aims to determine whether there is a statistically dependent relationship between two variables, namely the predictor variable and the response variable. One of the methods for estimating multiple linear regression parameters is the Least Squares Method. Therefore, careful and meticulous analysis and selection of appropriate techniques are required to overcome the multicollinearity problem and ensure accurate and meaningful regression analysis results[8].

FIXING MULTICOLLINEARITY IN A REGRESSION MODEL

Once we have determined that there's an issue with multicollinearity in your model, there are several different ways that you can go about trying to fix it so that you can create an accurate regression model. Below are some of the ways to make it possible:

1. Obtain more data: The more data you obtain for your model, the more precise the measurements can be and the less variance there will be. This is one of the more obvious solutions to multicollinearity.
2. Removing a variable: Removing a variable can make your model less representative; however, it can sometimes be the only solution to removing and avoiding multicollinearity altogether.
3. Create a standard set of independent variables.
4. Utilize a ridge regression or partial squares regression in conjunction with your model.
5. If all else fails or you decide it's not worth it to do any additional work on the model, do nothing: Even by not changing a model where we know multicollinearity exists, it still may not affect the efficiency of taking data from the existing model

III. RESULT ANALYSIS

We evaluate the performance of proposed algorithm and compare it with terms only based approach. The experiments were performed on Intel Core i5processor 4GB main memory and RAM: 4GB Inbuilt HDD: 500GB OS: Windows7. The algorithms are implemented in using python language design user interface and to used CSV file format to store data set.

Description of Columns of data set

- blood pressure (y = *BP*, in mm Hg)
- age ($x1$ = *Age*, in years)
- weight ($x2$ = *Weight*, in kg)
- body surface area ($x3$ = *BSA*, in sq m)
- duration of hypertension ($x4$ = *Dur*, in years)
- basal pulse ($x5$ = *Pulse*, in beats per minute)
- stress index ($x6$ = *Stress*)

allow us to investigate the various marginal relationships between the response BP and the predictors. Blood pressure appears to be related fairly strongly to Weight and BSA, and hardly related at all to the Stress level.

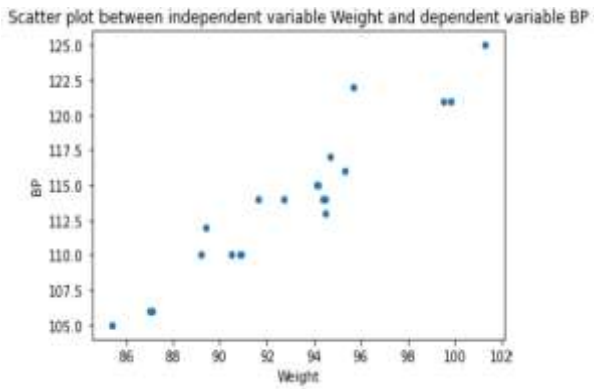

Fig 1 Scatter plot shows relationship between Weight and BSA

Identify Relationship between predictor weight and response BP

Table 1 Relationship between BSA and Weight

S No	Weight	BP
1	85.4	105
2	94.2	115
3	95.3	116
4	94.7	117
5	89.4	112
6	99.5	121
7	99.8	121
8	90.9	110
9	89.2	110
10	92.7	114
11	94.4	114
12	94.1	115
13	91.6	114
14	87.1	106
15	101.3	125
16	94.5	114
17	87	106
18	94.5	113
19	90.5	110
20	95.7	122

CONCLUSION

Multicollinearity can be a problem in a regression model because we would not be able to distinguish between the individual effects of the independent variables on the dependent variable. Multicollinearity may not affect the accuracy of the model as much. But we might lose reliability in determining the effects of individual features in model and that can be a problem when it comes to interpretability. Multicollinearity is the presence of high correlations between two or more independent variables (predictors). Correlation is the association between variables, and it tells us the measure of the extent to which two variables are related to each other. Two variables can have positive (change in one variable causes change in another variable in the same direction), negative (change in one variable causes change in another variable in the opposite direction), or no correlation. A simple example of positive correlation can be weight and height. A simple example of a negative correlation can be the altitude and oxygen level.

REFERENCE

1. In 2017 Jamal I. Daoud et al Multicollinearity and Regression Analysis December 2017 Journal of Physics Conference Series 949(1):012009 DOI:10.1088/1742-6596/949/1/012009 License CC BY 3.0
2. Neeraj Tiwari and Ankuri Agarwal Diagnostics of Multicollinearity in Multiple Regression Model for Small Area Estimation Statistics and Applications {ISSN 2454-7395 (online)} Volume 16 No. 2, 2018 (New Series), pp 37-47
3. Alhassan Umar Ahmad, U.V. Balakrishnan, Prem Shankar Jha Detection of Collinearity Effects on Explanatory Variables and Error Terms in Multiple Regressions International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue- 6S4, April 2019.
4. Senaviratna, N. A. M. R., & A. Cooray, T. M. J. (2019). Diagnosing Multicollinearity of Logistic Regression Model. Asian Journal of Probability and Statistics, 5(2), 1–9. https://doi.org/10.9734/ajpas/2019/v5i230132.
5. Solly Matshonisa Seeletse and 2 Motlalepula Grace Phalwane Dealing with Multicollinearity in Regression Analysis: A Case in Psychology Page No.: 2693-2703 Volume: 15, Issue 13, 2020 ISSN: 1816-949x Journal of Engineering and Applied Sciences Copy Right: Medwell Publication
6. Mariella Gregorich Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution Original submission received: 17 March 2021 / Revised: 14 April 2021 / Accepted: 15 April 2021 / Published: 17 April 2021
7. Katrina I. Sundus a, Bassam H. Hammo a b, Mohammad B. Al-Zoubi a, Amal Al-Omari Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to a fully annotated breast cancer dataset Informatics in MedicineUnlocked Volume 33, 2022, 101088
8. Amin Otoni Harefa1 , Yulisman Zega2 , Ratna Natalia Mendrofa3 The Application of the Least Squares Method to Multicollinear Data International Journal of Mathematics and Statistics Studies Vol.11, No.1, pp.30-39, 2023 Print ISSN: 2053-2229 (Print), Online ISSN: 2053-2210 (Online) Website: https://www.eajournals.org