Engineering Universe for Scientific Research and Management
ISSN (Online): 2319-3069
Vol. XVII Issue VII
July 2025

Handling Problem of Multicollinearity using Variance Inflation Factor (VIF) for Selecting Best Features

Swati Yadav[1], Nilesh Parmar [2] and Kamlesh Patidar[3]
M Tech 4th Sem Jawaharlal Institute of Technology Borawan Khargone India[1]
Asst. Prof. CSE Dept. Jawaharlal Institute of Technology Borawan Khargone India[2]
Asst Prof. CSE Dept. Jawaharlal Institute of Technology Borawan Khargone India[3]
swatiofficial2001@gmail.com[1], nparmarlife@jitechno.com[2], hodcse@jitechno.com[3]

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 your 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. In this paper our objective is to remove multicollinearity present in the dataset. First we identify Multicollinearity in the given data set using correlation matrix and then use VIF (Variable Inflation Factors) to determines the strength of the correlation between the independent Factors) identify multicollinearity using VIF (Variable Inflation Factors). Create a modal free from problem of multicollinearity.

Keywords: *Multicollinearity, Dependent, Independent, Correlation, Regression, Inflation, Factors*

1. 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. 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. For example, let's assume that in the following linear equation:

$$Y = W_0 + W_1X_1 + W_2X_2$$

Coefficient W1 is the increase in Y for a unit increase in X1 while keeping X2 constant. But since X1 and X2 are highly correlated, changes in X1 would also cause changes in X2 and we would not be able to see their individual effect on Y. This makes the effects of X1 on Y difficult to distinguish from the effects of X2 on Y. Multicollinearity may not affect the accuracy of the model as much. But we might lose reliability in determining the effects of individual features in your model and that can be a problem when it comes to interpretability[10,11].

Multicollinearity is the presence of high correlations between two or more independent variables (predictors). It is basically a phenomenon where independent variables are correlated. Let us first understand what the term correlation means. 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 we weigh (this is considered a general trend if we leave the exception case aside)[12].

2. Correlation vs. Collinearity vs. Multicollinearity

Correlation measures the strength and direction between two columns in your dataset. Correlation is often used to find the relationship between a feature and the target[13,14,15]:

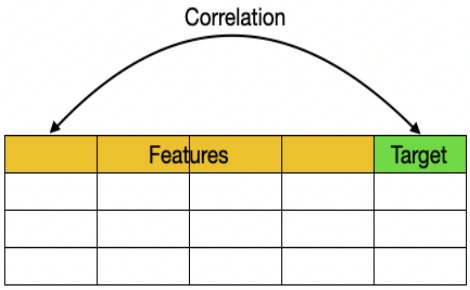

Figure 1 Correlation feature and the target

For example, if one of the features has a high correlation with the target, it tells you that this particular feature heavily influences the target and should be included when we are training the model.
Collinearity, on the other hand, is a situation where two features are linearly associated (high correlation), and they are used as predictors for the target.

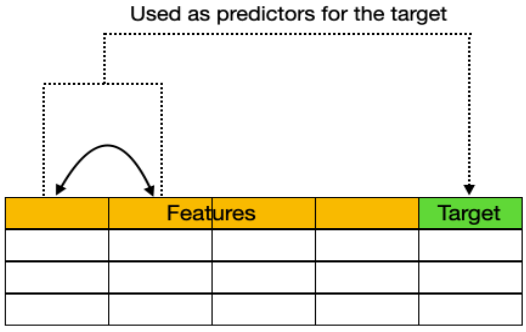

Figure 2 Feature predicted as target variable

Multicollinearity is a special case of collinearity where a feature exhibits a linear relationship with two or more features.

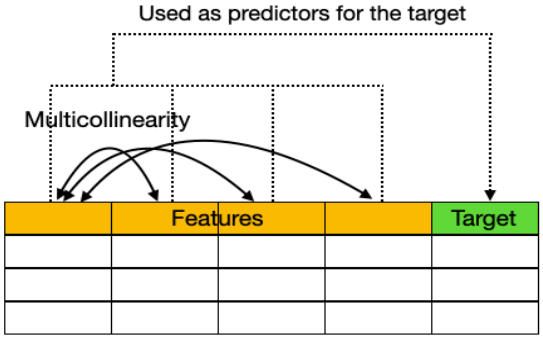

Figure 3 Multicollinearity between variables

3. Causes of Multicollinearity

Multicollinearity could occur due to the following problems [16,17]:

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.

4. How to Deal with Multicollinearity

If we only want to predict the value of a dependent variable, you may not have to worry about multicollinearity. Multiple regressions can produce a regression equation that will work for you, even when independent variables are highly correlated. The problem arises when we want to assess the relative importance of an independent variable with a high R2k (or, equivalently, a high VIFk). In this situation, try the following [15,18,19]:

1. Redesign the study to avoid multicollinearity. If you are working on a true experiment, the experimenter controls treatment levels. Choose treatment levels to minimize or eliminate correlations between independent variables.
2. Increase sample size. Other things being equal, a bigger sample means reduced sampling error. The increased precision may overcome potential problems from multicollinearity.

3. Remove one or more of the highly correlated independent variables. Then, define a new regression equation, based on the remaining variables. Because the removed variables were redundant, the new equation should be nearly as predictive as the old equation; and coefficients should be easier to interpret because multicollinearity is reduced.
4. Define a new variable equal to a linear combination of the highly correlated variables. Then, define a new regression equation, using the new variable in place of the old highly correlated variables.

5. Literature survey

In 2015 Ahmad A. Suleiman et al proposed "Analysis of Multicollinearity in Multiple Regressions". They concentrated on residuals analysis to check the assumptions for a multiple linear regression model by using graphical methods. Specifically, they plotted the residuals and standardized residuals given by model against predicted values of the dependent variables, normal probability plot, histogram of residuals and Quantile plot of residuals. They introduced the concept of multicollinearity to check whether one of the assumptions of the linear regression model that there is no multicollinearity among the explanatory variables is satisfied. They gave an example that indicated the presence of multicollinearity in the regression model using review (statistical software)[1].

In 2016 Kristina Vatcheva et al proposed "Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies". They simulated datasets and real-life data from the Cameron County Hispanic Cohort to demonstrate the adverse effects of multicollinearity in the regression analysis and encourage researchers to consider the diagnostic for multicollinearity as one of the steps in regression analysis. They recommend that along with the bivariate correlation coefficients between the predictors in the model and the VIFs, researchers should always examine the changes in the coefficient estimates along with the changes in their standard errors and even the changes in VIF. VIF less than 5 (VIF) does not always indicate low multicollinearity [2].

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. 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 R. They discovered collinearity in the regression output; we should reject the interpretation of the relationships as false until the issue is resolved. 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 [3].

In 2018 Yunus Kologlu et al proposed "A Multiple Linear Regression Approach For Estimating the Market Value of Football Players in Forward Position". They used values of the football players in the forward positions are estimated using multiple linear regressions by including the physical and performance factors in 2017-2018 seasons. They achieved to build a regression model in 0.10 significance level with 52 attributes, 20 MAPE..The only interesting thing in the study is the fact that card numbers didn't affect the market value of the players; which can be caused by the reason that valuable players act more cautiously to do not get any penalty. Overall, the study could be improved by a more reasonable collection of data [4].

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 carefully observes how those explanatory variables inter correlated with one another and to the response variable. Multicollinearity was discovered to work with a severe Multicollinearity form is not a problematic sometimes especially if the aims of the analysis is to use multiple regression for prediction purposes, it will be accurate as it is supposed to be despite the presence of multicollinearity, where the problem lies is if to check the contribution of each individual independent variables[5].

In 2020 Noora Shrestha et al proposed "Detecting Multicollinearity in Regression Analysis". They discussed the three primary techniques for detecting multicollinearity using the questionnaire survey data on customer satisfaction. They observed that product attractiveness is more rational cause for the customer satisfaction than other predictors. Furthermore, advanced regression procedures such as principal components regression, weighted regression, and ridge regression method can be used to determine the presence of multicollinearity the relationship between customer satisfaction with the major factors product quality, brand experience, product feature,

product attractiveness, and product price are significant with p [6].

In 2021 Alhassan Umar Ahmad et al proposed. "A Study of Multicollinearity Detection and Rectification under Missing Values". They discussed the consequences of missing observations on data-based multicollinearity that were analyzed. Different missing values have a different effect on multicollinearity in the system of multiple regression models. They found that tolerance and variance inflation factors fluctuate due to the missing of information from the sample analyzed at different percentages of the missing values. They observed that the more missing values available in the sample obtain from either population statistics or survey than multicollinearity will be found in the system of multiple regression, this is because as the number of Missing ness increase it shows a drastic decrease from the tolerance level on both monotone. [7].

In 2022 Katrina I. et al "Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to a fully annotated breast cancer dataset" . They 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. They applied the CRISP-DM extension for the medical domain methodology to design and construct the dataset. They experimented with the JBRCA dataset to solve many problems and issues related to the dataset's construction during a one-year journey.[8]

In 2023 Amin Otoni et al proposed "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. Descriptive statistical table of response variables and predictor variables, where the average results are rounded. They obtained more stable and accurate regression coefficient estimates and a more reliable linear regression model. Multicollinearity can cause problems in estimating accurate regression coefficients.[9]

6. Problem statement

There are so many problems can have with having with Multicollinearity some of them are

1. 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.
2. From a practical point of view, there is no point in keeping 2 very similar predictors in the predicted model.
3. Multicollinearity may not affect the accuracy of the model as much. But we might lose reliability in determining the effects of individual features in a model.

7. Proposed algorithm

Proposed approach has following steps

Step 1 calculate residual by using formula

$$r_i = y_i - \hat{y}_i$$

Step2 Calculate sum of squared regression (SSR)

$$SSR = \sum (y_i - \hat{y}_i)^2 = \sum r_i$$

Step 3 Calculate sum of squared total (SST)

$$SSR = \sum (y_i - \bar{y}_i)^2$$

Step 4 Calculate coefficient of determination

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Step 5 Calculate tolerance value

Tolerance = 1-R^2

Step 6 Calculate Variance Inflation Factor (VIF)

$$VIF = \frac{1}{\text{Tolerance}}$$

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

8. Implementation and dataset

We evaluate the performance of proposed algorithms 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 using python language design user interface and to used CSV file format to store data set

Description about the Dataset

Blood pressure data in which researchers observed the following data on 20 individuals with high blood pressure:

Table 1: BP dataset

S N	BP	Age	Weight	BSA	Dur	Pulse	Stress
0	105	47	85.4	1.75	5.1	63	33
1	115	49	94.2	2.10	3.8	70	14
2	116	49	95.3	1.98	8.2	72	10
3	117	50	94.7	2.01	5.8	73	99
4	112	51	89.4	1.89	7.0	72	95
5	121	48	99.5	2.25	9.3	71	10
6	121	49	99.8	2.25	2.5	69	42
7	110	47	90.9	1.90	6.2	66	8

Description of Columns of data set

- blood pressure (*y* = *BP*, in mm Hg)
- age (*x*1 = *Age*, in years)
- weight (*x*2 = *Weight*, in kg)
- body surface area (*x*3 = *BSA*, in sq m)
- duration of hypertension (*x*4 = *Dur*, in years)
- basal pulse (*x*5 = *Pulse*, in beats per minute)
- stress index (*x*6 = *Stress*)

9. Results and analysis

After implementation we can see that features BSA and Weight have high correlated 0.88. They play same role in predicting the target variable. These two features degrade accuracy of the model. We must remove any one of them.

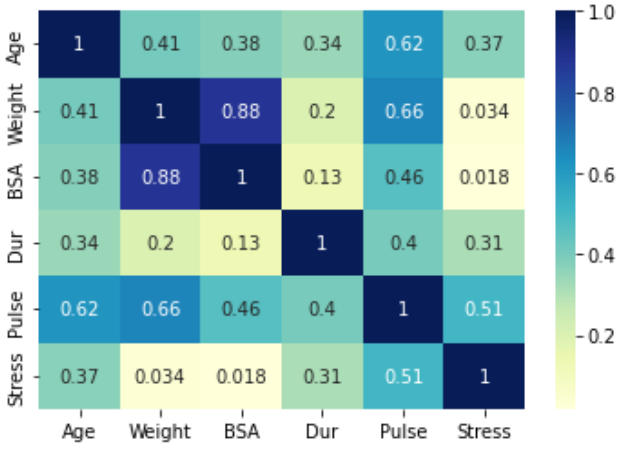

Figure 4 Value of correlation

10. Conclusions

In this paper our objective is to detect highly correlated independent variables. After detecting Multicollinearity, we remove one or more of the highly correlated. We used real life data set of heat patient data set which contain 6 features, 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) and response variable blood pressure (y = BP, in mm Hg). Based on the features we need to predict that a person has high belongs blood pressure. We Implement the proposed system using python language. We found that BSA and Weight two predictor are highly correlated. Due to this we cannot identify actual effect of other variables. We delete BSA and then check the effect of the other variables. We also calculate the R squared value for each predictor. We also used scatter plot to check the relation between the BSA, Weight and BP.

References

[1]. Ahmad A. Suleiman Analysis of Multicollinearity in Multiple Regressions International Journal of Advanced Technology in Engineering and Science www.ijates.com Volume No 03, Special Issue No. 01, April 2015 ISSN (online): 2348 – 7550

[2]. Kristina Vatcheva Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies Vatcheva KP, Lee M, McCormick JB, Rahbar MH (2016) Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. Epidemiol 6: 227. doi:10.4172/2161-1165.1000227.

[3]. Jamal I. Daoud Multicollinearity and Regression Analysis ICMAE'17 IOP Publishing IOP Conf. Series: Journal of Physics: Conf. Series 949 (2017) 012009 doi :10.1088/1742-6596/949/1/012009.

[4]. Yunus Kologlu, Hasan Birinci, Sevde Ilgaz Kanalmaz, Burhan Özyılmaz A Multiple Linear Regression Approach For Estimating the Market Value of Football Players in Forward Position Abdullah Gül University Industrial Engineering Department 2018.

[5]. Alhassan Umar Ahmad 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.

[6]. Noora Shrestha Detecting Multicollinearity in Regression Analysis American Journal of Applied Mathematics and Statistics, 2020, Vol. 8, No. 2, 39-42 Available online at http://pubs.sciepub.com/ajams/8/2/1 Published by Science and Education Publishing DOI:10.12691/ajams-8-2-1.

[7]. Alhassan Umar Ahmad, U A Study of Multicollinearity Detection and Rectification under Missing Values Turkish Journal of Computer and Mathematics Education Vol.12 No.1S. (2021), 399- 418. Received: 11 January 2021

[8]. 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 Medicine Unlocked Volume 33, 2022, 101088

[9]. 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/

[10]. Michael Olusegun Akinwande Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis Open Journal of Statistics, 2015, 5, 754-767 Published Online December 2015 in SciRes. http://www.scirp.org/journal/ojs.

[11]. Dr. Manoj Kumar Mishra A Study of Multicollinearity in Estimation of Coefficients in Ridge Regression Asian Journal of Technology and Management Research (AJTMR) ISSN: 2249-0892 Volume 07– Issue 02, Dec 2017.

[12]. 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.

[13]. Katerina M. Marcoulides1 and Tenko Raykov Evaluation of Variance Inflation Factors in Regression Models Using Latent Variable Modeling Methods Educational and Psychological Measurement 2019, Vol. 79(5) 874–882 The Author(s) 2018 Article reuse guidelines: sagepub.com/journals-permissions DOI: 10.1177/0013164418817803 journals.sagepub.com/home/epm.

[14]. N. A. M. R. Senaviratna1* and T. M. J. A. Cooray Diagnosing Multicollinearity of Logistic Regression Model Asian Journal of Probability and Statistics 5(2): 1-9, 2019; Article no.AJPAS.51693 ISSN: 2582-0230.

[15]. Jong Hae Kim Multicollinearity and misleading statistical results Department of Anesthesiology and Pain Medicine, School of Medicine, Daegu Catholic University, 33 Duryugongwon-ro 17-gil, Namgu, Daegu 42472, Korea Tel: +82-53-650-4979, Fax: +82-53-650-4517 Email: usmed12@gmail.com ORCID: https://orcid.org/0000-0003-1222-0054 Received: March 3, 2019. Revised: May 17, 2019. Accepted: July 8, 2019. Korean J Anesthesiol 2019 December 72(6): 558-569 https://doi.org/10.4097/kja.19087.

[16]. P. Sai Shankar, J.V. Narasimham and G. Ananthan Application of principal component regression analysis in agricultural studies www.researchjournal.co.in International Research Journal of Agricultural Economics and Statistics ISSN-2229-7278γVolume 10 | Issue 1 | March, 2019 | 59-64.

[17]. Shishodiya Ghanshyam Singh1 · S. Vasantha Kumar1 Dealing with Multicollinearity Problem in Analysis of Side Friction Characteristics Under Urban Heterogeneous Traffic Conditions Arabian Journal for Science and Engineering https://doi.org/10.1007/s13369-020-05213-y Received: 18 June 2020 / Accepted: 7 December 2020 © The Author(s) 2021.

[18]. Solly Matshonisa Seeletse and Motlalepula Grace Phalwane Dealing with Multicollinearity in Regression Analysis: A Case in Psychology J. Eng. Applied Sci., 15 (13): 2693-2703, 2020 Page No.: 2693-2703 Volume: 15, Issue 13, 2020 ISSN: 1816-949x Journal of Engineering and Applied Sciences Copy Right: Medwell Publications. Accepted: 27 February 2021; Published online: 5 April 2021.

[19]. Mariella Gregorich , Susanne Strohmaier Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution Gregorich, M.; Strohmaier, S.; Dunkler, D.; Heinze, G. Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution. Int. J. Environ. Res. Public Health 2021, 18, 4259. https://doi.org/10.3390/ijerph18084259 Academic Editors: Jimmy T. Efird and Paul B. Tchounwou Received: 17 March 2021 Accepted: 15 April 2021 Published: 17 April 2021.