Volume 9, Issue 06, June 2024 ISSN (online): 2456-0006
International Journal of Science Technology Management and Research
Available online at: www.ijstmr.com

Analysis Effect of Two or More Correlated Independent Variables to Predict Values of Dependent Variable

Muskan Raghuwanshi
ME(Software Engineering) 4th semester
Computer Science and Engineering Department
Jawaharlal Institute of Technology Borawan (Khargone)

Mr. Kamlesh Patidar
Assistant Professor
Computer Science and Engineering Department
Jawaharlal Institute of Technology Borawan (Khargone) M.P.

Abstract: ***There is several ways to handle multicollinearity like, 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, Increase sample size. Other things being equal, a bigger sample means reduced sampling error. The increased precision may overcome potential problems from multicollinearity; 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, define a new variable equal to a linear combination of the highly-correlated variables. In the proposed work our objective design a model and remove multicollinearity present in the model, for this 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 variables and finally use real life data set to identify multicollinearity using VIF (Variable Inflation Factors). Create a modal free from problem of multicollinearity***

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

I. INTRODUCTION

Data analysis is the exercise of gathering information and interpreting what it can mean. When conducting data analysis, experts collect raw data and use a variety of methods for interpreting the information it presents. There are five main types of data analysis that describe how people can use different types of data to reach conclusions and make decisions. Here's more information about the primary types of data analysis[11,12]:

- **Descriptive analysis**: Descriptive analysis determines what happened in a certain situation. This type of analysis typically involves ordering and adjusting data from different sources to interpret its meaning.
- **Exploratory analysis**: This type of analysis explores relationships between specific data points or sets. When engaging in exploratory analysis, you can find connections between pieces of information and create hypotheses to determine why they might relate to each other.
- **Predictive analysis:** Predictive analysis refers to developing a prediction for what might happen. This can involve considering results from an earlier analysis and exploring trends and patterns to make an estimation about what might occur in the future.
- **Diagnostic analysis:** Diagnostic analysis considers why something happened. When using diagnostic analysis, you can explore events that occur and the context that surrounds them to reach a solution as to why they might arise.
- **Prescriptive analysis**: This type of analysis attempts to predict how something might happen. Prescriptive analysis considers raw data that relates to trends or patterns and determines how it might produce a certain expected result.

II. MULTICOLLINEARITY

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. For example, let's assume that in the following linear equation[13,14]:

$$Y = W_0 + W_1 X_1 + W_2 X_2$$

Coefficient W_1 is the increase in Y for a unit increase in X_1 while keeping X_2 constant. But since X_1 and X_2 are highly correlated, changes in X_1 would also cause changes in X_2 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 X_2 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.
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 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.

III. DEALING 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[15,16,17].
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:

- 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.
- Increase sample size. Other things being equal, a bigger sample means reduced sampling error. The increased precision may overcome potential problems from multicollinearity.
- 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.
- 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.

IV. LITERATURE SURVEY

In 2017 Bager, Ali et al proposed "Addressing multicollinearity in regression models: a ridge regression application" They determined the most important macroeconomic factors which affect the unemployment rate in Iraq, using the ridge regression method as one of the most widely used methods for solving the multicollinearity problem. The solution adopted in our research is the ridge regression model, which 18 was tested for identifying the factors that could explain the unemployment rate in an Arabic developing country, namely Iraq. The study showed that the use of the ridge regression method in the cases when explanatory variables are affected by multicollinearity is one of the successful ways to solve this issue. Therefore, applying the ridge regression method in other studies is recommended, since it provides better estimators than the ordinary least square method when the explanatory variables are related, without omitting any of the explanatory variables[1].

In 2017 Dr Manoj Kumar Mishra et al proposed "A Study of Multicollinearity in Estimation of Coefficients in Ridge Regression". The traditional solution is to collect more data or to drop one or more variables. Collecting more data may often be expensive or not practicable in many situations and to drop one or more variables from the model to alleviate the problem of multicollinearity may lead to the specification bias and hence the solution may be worse than the disease in certain situations. One may be interested in squeezing out maximum information from whatever data one has at one's disposal. They applied test for a test for multicollinearity by extracting the VIF quantities. Therefore, a multicollinearity problem has been observed at our constructed model. The technique of RR is used to deal with the problem of multicollinearity at the constructed model. By using the SYSTAT package, all values of coefficients are estimated based on suitable values of k and estimate of the model has been discussed [2].

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 approach does not require any additional survey or conducting extra crop cutting experiments (CCE) for crop production estimate at the district level. They demonstrated an application of small area estimates by using RR model. Least squares (LS) method is the oldest techniques for estimating the parameters of linear regression model under some assumptions.. Time series data on production of rice, area under rice, irrigated area under rice and fertilizer consumption pertaining to the

period 1990- 91 to 2002-03 for Uttarakhand state of India is taken from the Bulletin of Agricultural statistics, published by the government of Uttarakhand, India[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. Players from 4 major leagues of Europe are examined, and by applying Breusch – Pagan test for homoscedasticity, a reasonable regression model within 0.10 significance level is built, and the most and the least affecting factors are explained in detail. They evaluated football players with several criteria in different significance levels within the consideration of multicollinearity. 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 is carefully observes how those explanatory variables inter correlated with one another and to the response variable. Multicollinearity was discovered to work with a severe proportion using arrays of correlation analysis procedure which affects the performance of the explanatory variables present in the model by making it less independent and more redundant as it should not be. Complete elimination of collinearity is not possible but they reduced it is degree of intensity to enhance the performance of independent variables and error term in the model Multicollinearity from is not a problematic some time 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 2019 N. A. M. R. Senaviratna et al proposed" Diagnosing Multicollinearity of Logistic Regression Model". 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. The response variable is accident severity that consists of two levels particularly grievous and non-grievous. Multicolinearity is identified by correlation matrix, tolerance and VIF values and confirmed by condition index and variance proportions. The range of solutions available for logistic regression such as increasing sample size, dropping one of the correlated variables and combining variables into an index. It is safely concluded that without increasing sample size, to omit one of the correlated variables can reduce multicollinearity considerably. The problem of multicollinearity arises when one explanatory variable is not a linear function of another explanatory variable. The presence of multicollinearity specifies the biased coefficient estimates and very large standard errors for the logistic regression coefficients[6].

In 2020 Noora Shrestha et al proposed "Detecting Multicollinearity in Regression Analysis" Multicollinearity occurs when the multiple linear regression analysis includes several variables that are significantly correlated not only with the dependent variable but also to each other. Multicollinearity makes some of the significant variables under study to be statistically insignificant. They discussed on the three primary techniques for detecting the multicollinearity using the questionnaire survey data on customer satisfaction. The first two techniques are the correlation coefficients and the variance inflation factor, while the third method is eigenvalue method. It is observed that the 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 [7].

In 2021 Alhassan Umar Ahmad et al U.V proposed. "A Study of Multicollinearity Detection and Rectification under Missing Values". They discussed the consequences of missing observations on data-based multicollinearity were analyzed. Different missing values have a different effect on multicollinearity in the system of multiple regression models. Similarly, the comparison was done to investigate each response of multicollinearity on each pattern of the missing values with the same informatics data. They found that tolerance and variance inflation factor fluctuates due to the missing of information from the sample analyzed at different percentages of the missing values. They was 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 and arbitrary types as observed from the analysis , it brings out categorically that no missing values are small no matter how it is, can change the nature of correlation, Tolerance, and variance inflation factors which finally in return will affect the linear relationships among the response and predictor variables and end up producing a severe multicollinearity[8].

In 2021 Mariella Gregorich et al proposed "Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution". 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. They demonstrated, using two examples, how the diagnostic tools for collinearity and near-collinearity may fail in guiding the analyst in selecting the most appropriate way of handling collinearity. Hence, it is the aim of the analysis that should guide decisions on keeping or omitting a variable in a model. When statistical modeling is used to pursue a predictive aim, two highly correlated independent variables will lead to high variance in the predictions, even if both variables are relevant for prediction. In small samples, it may then be beneficial to omit one of the pair in order to decrease that variance, even if this incurs some new bias in the predictions[9].

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". They, 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. They applied the CRISP-DM extension for the medical domain methodology to design and construct the dataset. We experimented with the JBRCA dataset to solve many problems and issues related to the dataset's construction during a one-year journey. Data encoding, different data types, scaling, balancing, and multicollinearity problems were among the few. We provided solutions for the most raised issues by applying the most common techniques used in DM and ML. We demonstrated that it is indispensable for the recurrent breast cancer prediction system to have an original, extensive, versatile, adequately classified, and richly annotated reference dataset. At the end of the study, the JBRCA dataset has 20 attributes and 7562 instances[10].

V. PRAPOSED APPROACH

To detect the multicollinearity and identify the variables involved, linear regressions must be carried out on each of the variables as a function of the others. We used the following steps and need to calculate following things e

- **The R^2 of each of the models:-** If the R^2 is 1, then there is a linear relationship between the dependent variable of the model (the Y) and the explanatory variables (the Xs).

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

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

The sum squared regression is the sum of the residuals squared, and the total sum of squares is the sum of the distance the data is away from the mean all squared. As it is a percentage it will take values between 0 and 1.

The coefficient of determination, or R^2, is a measure that provides information about the goodness of fit of a model. In the context of regression it is a statistical measure of how well the regression line approximates the actual data. It is therefore important when a statistical model is used either to predict future outcomes or in the testing of hypotheses. There are a number of variants; the one presented here is widely used.

.

VI. COMPARATIVE ANALYSIS

Variation Inflation Factor after removing multicollinearity.

Table 1 Variation Inflation Factor after removing multicollinearity.

S No	Variables	VIF
0	Age	1.659637
1	Weight	2.256150
3	Dur	1.235620
4	Pulse	3.599913
5	Stress	1.739641

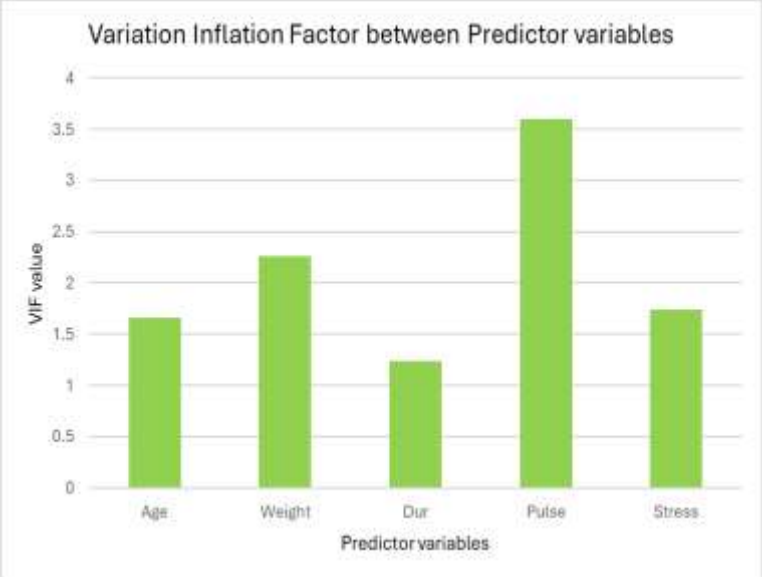

Figure 1 Bar graph for VIF Values after removing multicollinearity.

VII. CONCLUSION

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 have 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.

.

REFERENCE

1. Bager, Ali and Roman, Monica and Algedih, Meshal and Mohammed, Bahr Addressing multicollinearity in regression models: a ridge regression application Online at https://mpra.ub.uni-muenchen.de/81390/ MPRA Paper No. 81390, posted 16 Sep 2017 09:04 UTC.
2. 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.
3. 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.
4. Yunus Koloğlu, 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. 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.
7. 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.
8. 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; Accepted: 27 February 2021; Published online: 5 April 2021.
9. Mariella Gregorich , Susanne Strohmaier Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution. 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
10. 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.
11. 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
12. Sudhanshu K. Mishra Shapley value regression and the resolution of multicollinearity C-91, (Ground Floor) Avantika, Rohini Sector-1 Delhi – 110085 Email: mishrasknehu@hotmail.com.
13. Christopher Winship, Bruce Western Winship, Christopher, and Bruce Western. 2016. "Multicollinearity and Model Misspecification." Sociological Science 3: 627-649. Received: February 5, 2016 Accepted: March 5, 2016 Published: July 26, 2016 Editor(s): Jesper Sørensen, Olav Sorenson.
14. Hanan Duzan Solution to the Multicollinearity Problem by Adding some Constant to the Diagonal Journal of Modern Applied Statistical Methods May 2016, Vol. 15, No. 1, 752-773. Copyright © 2016 JMASM, Inc. ISSN 1538 – 9472.
15. 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.
16. Gary H. McClelland1 & Julie R. Irwin2 & David Disatnik3 & Liron Sivan Multicollinearity is a red herring in the search for moderator variables: A guide to interpreting moderated multiple regression models and a critique of Iacobucci, Schneider, Popovich, and Bakamitsos (2016) Behav Res (2017) 49:394–402 DOI 10.3758/s13428-016-0785-2.
17. N S M Shariff, H M B Duzan A Comparison of OLS and Ridge Regression Methods in the Presence of Multicollinearity Problem in the Data International Journal of Engineering & Technology, 7 (4.30) (2018) 36-38 International Journal of Engineering & Technology Website: www.sciencepubco.com/index.php/IJET.