Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 5, July, 2021: 4347 - 4352

Research Article

Multicollinearity detection and feature selection in diagnosis of cardio related tests

Dr.L.V.Nandakishore[1], Dr.S.Aruna[2]

Abstract

In this paper, data set relating to the results of the various diagnostic tests done on patients to diagnose cardiac related problems with variables are considered. A linear regression equation is found with class as a response variable and the other variables as continuous predictors. The same was analyzed for multicollinearity by calculating the variance inflation factor (VIF). To get a better regression equation with no multicollinearity, the variables with high VIF values were eliminated. Feature reduction was applied to the data, and the outcomes were compared.

Keywords: Multicollinearity, VIF, regression analysis, feature reduction, heart data.

Introduction

The medical profession requires many parameters to diagnose a disease. When the number of tests are taken and interpreted to come to a decision whether the patient is to be diagnosed positive or otherwise it is found that statistically some of the results are correlated. These correlated variables can be reduced to arrive at a better conclusion regarding the status of the patient. Some of the parameters taken for diagnosis of cardio vascular problems are given below. These consist of 13 variables. The final decision is given in the last column class.Statistically these variables have to be analysed for multicollinearity and multicollinear variables can be removed to give a better accuracy with less number of variables. These variables are termed as predictor variables and if two or more variables are correlated it gives rise to multicollinearity in the data. This study's data set is sourced from

https://s3-eu-west-1.amazonaws.com/pstorage-plos-3567654/8937223/S1Data.csv

Table1

Attributes of the dataset

S.no	Attribute name
1	Time

[1] Department of Mathematics, Dr. M.G.R. Educational and Research Institute, Chennai,India.

[2] Department of Computer Science, Agurchand Manmull Jain College, Chennai, India.
aruna.s@amjaincollege.edu.in

2	Event
3	Gender
4	Smoking
5	Diabetes
6	Blood pressure
7	Anaemia
8	Age
9	Ejection fraction
10	Sodium
11	Creatinine
12	Platelets
13	CPK
14	Class

The Variance Inflation Factor (VIF) plays an important role in deciding collinearity among variables. The VIF values and their interpretation are given in table 2 below.
Table 2

VIF values and their interpretation.

VIF Values	Interpretation
1	Not correlated
1 -5	Moderately correlated
> 5	Highly correlated

Multicollinearity in regression arises if there is correlation of one or more predictor variables with other predictor variables. It increases the variance of the regression coefficients resulting in instability and wrong predictions of the response variable.

a) A high (above 10) variance influence factor indicates multicollinearity.
b) A VIF of one indicates no multicollinearity.
c) A VIF greater than 5 implies that the regression coefficient for that variable is wrongly estimated and hence errors in prediction.
d) The ratio of the model variance to the variance of a model that includes only that single independent predictor variable is called VIF of the regression model.
e) Using R^2, the coefficient of determination, we can find how well the given data points fit a curve.

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

The effects of multicollinearity can be reduced by some of these methods.

a) QR decomposition can be applied to remove the correlated predictors from the regression equation.
b) Partial Least Squares or Principal Components Analysis can be applied to reduce the predictors to uncorrelated variables.

c) The correlated variables are removed from the table and a Regression equation is formed after analyzing the VIF of the remaining variables.

Methodology

Using the given variables, a regression equation is found. Here class is taken as a response variable and the other variables as continuous predictors. The data consists of 300 instances. The given data is applied in Minitab 16 software to get the results for the interpretations. The table3 gives the Variance Inflation Factor the given variables. This is compared with values in table 2. It is observed from table 3, that VIF that time, event and gender have VIF greater than one, indicating moderate multicollinearity. These variables are removed and regression analysis is done again.

Regression Analysis:

Analysis of Variance

```
Source          DF     SS      MS     F      P
Regression       13   5.2390  0.4030  1.92  0.028
Residual Error  285  59.9383  0.2103
Total           298  65.1773
```

The regression equation without removing the correlated factors is
Results = 0.958 + 0.000653 Time + 0.0376 Event - 0.101 Gender + 0.0877 Smoking + 0.0501 Diabetes - 0.0400 BP + 0.0295 Anaemia + 0.00376 Age - 0.00466 Ejection.Fraction - 0.00571 Sodium + 0.0674 Creatinine + 0.000018 CPK

Table 3

VIF for variables before removing multi collinearity

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.9577	0.8834	1.08	0.279	
Time	0.000653	0.000418	1.56	0.119	1.491
Event	0.0376	0.07439	0.51	0.614	1.715
Gender	-0.10064	0.06444	-1.56	0.119	1.345
Smoking	0.08769	0.06439	1.36	0.174	1.285
Diabetes	0.0501	0.05549	0.9	0.367	1.065
BP	-0.03995	0.05744	-0.7	0.487	1.069
Anaemia	0.02952	0.05583	0.53	0.597	1.087
Age	0.003759	0.002388	1.57	0.117	1.143
Ejection.Fraction	-0.00466	0.002432	-1.92	0.056	1.174
Sodium	-0.00571	0.006345	-0.9	0.369	1.111
Creatinine	0.06735	0.02745	2.45	0.015	1.142
Platelets	-0.00000035	0.00000028	-1.25	0.212	1.046
CPK	0.00001796	0.00002838	0.63	0.527	1.075

S = 0.458595   R-Sq = 8.0%   R-Sq(adj) = 3.8%

Residual Plots for Results

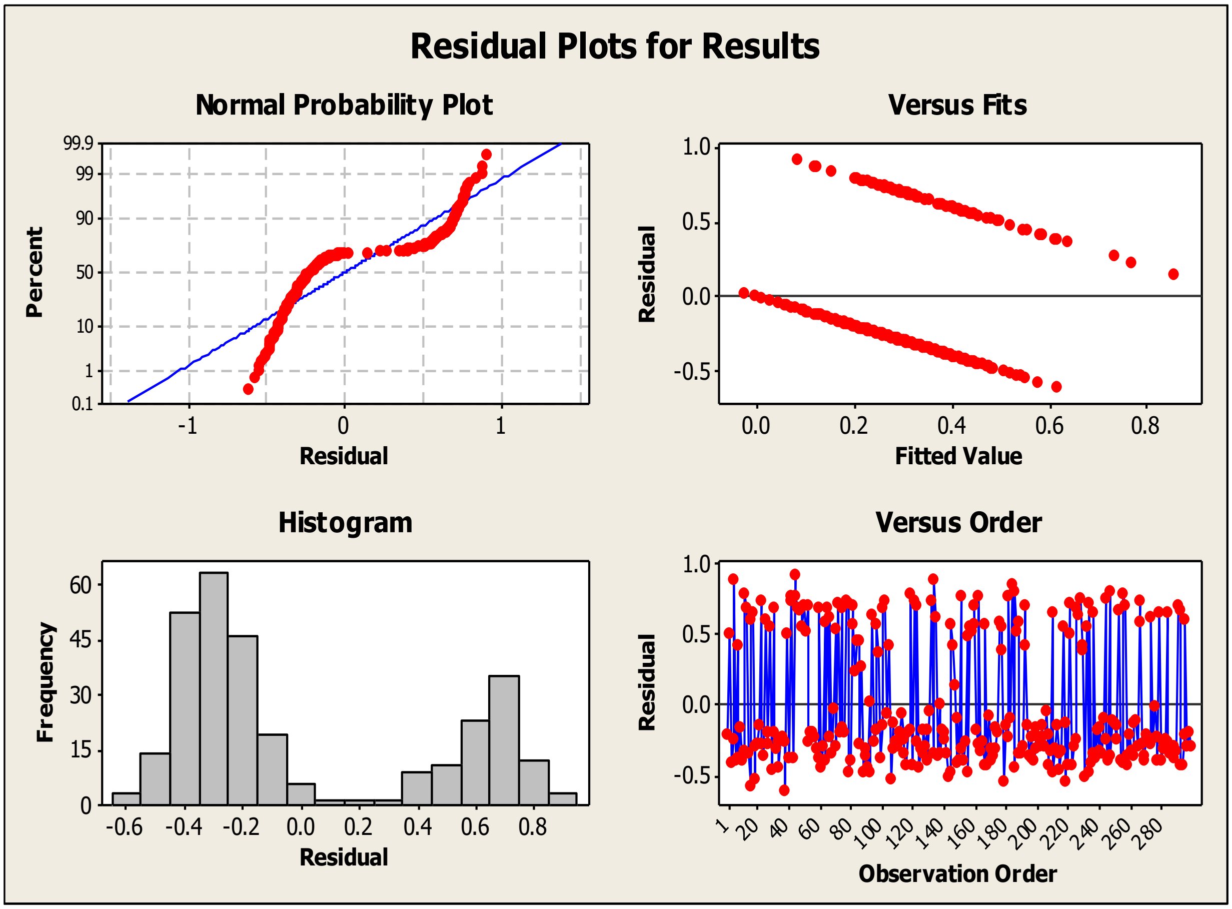

Figure 1 Residual plots before feature selection

The residual histogram follows an approximate normal distribution and the observation order shows that there is no non randomness in the data and there is non- correlation of residuals. From the normal probability plot is observed that there is not much deviation of residuals from the straight line indicating normalcy. The graph of fitted values and residuals shows that the residuals show a random pattern and constant variance. The regression analysis is done removing the collinear variables, time event and gender and equations are rewritten to remove multicollinearity.

Regression Analysis after removing collinear variables:

Table 4

VIF for variables after removing multi collinearity

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.968	0.876	1.11	0.27	
Smoking	0.03927	0.05833	0.67	0.501	1.047

Diabetes	0.05939	0.05545	1.07	0.285	1.056
BP	-0.0495	0.05638	-0.88	0.381	1.022
Anaemia	0.022	0.05544	0.4	0.692	1.064
Age	0.00305	0.002311	1.32	0.188	1.063
Ejection.Fraction	-0.004425	0.002308	-1.92	0.056	1.05
Sodium	-0.005241	0.006327	-0.83	0.408	1.096
Creatinine	0.06633	0.02665	2.49	0.013	1.069
Platelets	-0.00000029	0.00000028	-1.06	0.288	1.027
CPK	0.00001283	0.00002822	0.45	0.65	1.055

All VIF are approximately equal to one.

S = 0.460270   R-Sq = 6.4%   R-Sq(adj) = 3.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	10	4.1648	0.4165	1.97	0.037
Residual Error	288	61.0125	0.2118		
Total	298	65.1773			

The new regression equation is
Results = 0.968 + 0.0393 Smoking + 0.0594 Diabetes - 0.0495 BP + 0.0220 Anaemia + 0.00305 Age - 0.00443 Ejection.Fraction - 0.00524 Sodium + 0.0663 Creatinine + 0.000013 CPK

Residual Plots for Results

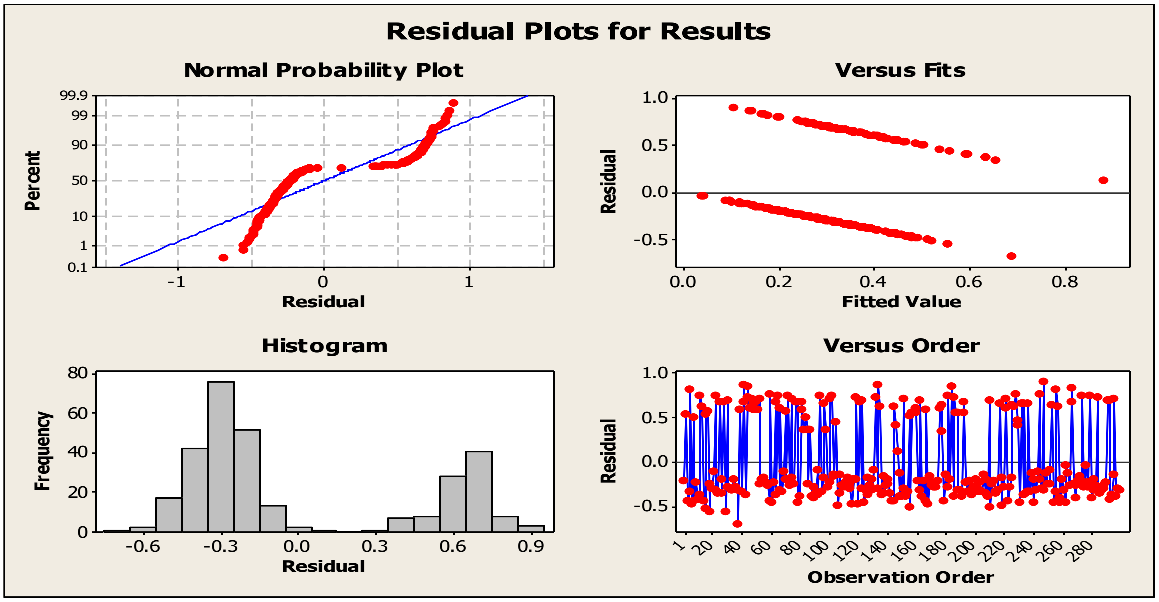

Figure 2 Residual plots after feature selection

Conclusion

Data set relating to diagnostic results of various contributing factors for cardio related diseases is analysed. A linear regression equation is found with class as a response variable and the other variables as continuous predictors. The same was analysed for multicollinearity and the variables with VIF greater than one time, event and gender were removed to give a better linear regression equation devoid of multicollinearity.

The regression equation is
Results = 0.958 + 0.000653 TIME + 0.0376 Event - 0.101 Gender + 0.0877 Smoking
+ 0.0501 Diabetes - 0.0400 BP + 0.0295 Anaemia + 0.00376 Age
- 0.00466 Ejection.Fraction - 0.00571 Sodium + 0.0674 Creatinine
+ 0.000018 CPK

It was found to have multicollinear variables which were removed and the regression equation rewritten as

The regression equation is
Results = 0.968 + 0.0393 Smoking + 0.0594 Diabetes - 0.0495 BP + 0.0220 Anaemia
+ 0.00305 Age - 0.00443 Ejection.Fraction - 0.00524 Sodium
+ 0.0663 Creatinine + 0.000013 CPK

References

1. Jamal I. Daoud (2017), Multicollinearity and Regression Analysis *J. Phys.:*
a. *Conf. Ser.* 949 012009.
2. Carl F. Mela and Praveen K. Kopalle (2002), The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations. *Journal of Applied Economics*, 43, 667-677.
3. N.O Adeboye, I. S Fagoyinbo and T.O Olatayo (2014), Estimation of the Effect of Multicollinearity on the Standard Error for Regression Coefficients, *IOSR Journal of Mathematics (IOSR-JM)* e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 10, Issue 4, pp 16-20.
4. Nandakishore L.V and S. Aruna (2021), “Multicollinearity detection and Regression Analysis of intensity of stroke in patients”,*International Journal of Grid and Distributed Computing*, Vol.1.
5. Farrar and Glauber (1967), "Multicollinearity in regression analysis", *Review of Economics and Statistics*, 49, pp. 92-107.
6. Weisberg (2005), "Applied Linear Regression" John Wiley & Sons, New York.
7. Nandakishore, L.V and S.Aruna (2021), “Analysis of Time Series Trends and ARIMA models to forecast covid 19 cases”,*Psychology and Education*, Vol.58.
8. Midi H, Sarkar S, Rana S. (2013), Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*; pp 253-267.
9. N.A.M.R.Senaviratna, T.M.J.A.Cooray, Diagnosing Multicollinearity of Logistic Regression Model Asian Journal of Probability and Statistics, 5(2), pp1-9.