Original Research Article

Diagnosing Multicollinearity of Logistic Regression Model

ABSTRACT

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 affects confidence intervals and hypothesis tests. In this study some diagnostic measurements are discussed to detect multicollinearity namely tolerance, Variance Inflation Factor (VIF), condition index and variance proportions. Motorcycle accident data were used to evaluate diagnostic measurements. Secondary data used from 2014 to 2016 in this study were acquired from the Traffic Police headquarters, Colombo in Sri Lanka. 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.

Keywords: Logistic regression, Multicollinearity, Tolerance, Variance Inflation Factor, Condition index

1. INTRODUCTION

Binary logistic regression is used to model the relationship between dichotomous dependent variable and multiple independent variables which are either continuous or categorical. There are some assumptions under binary logistic regression which are required to satisfy to give a valid result [1].

- Linearity: Explanatory variables should have a linear relationship with the logit of the response variable.
- Independent errors: Errors should not be correlated.
- Multicolinearity: Explanatory variables should not be highly correlated with each other.
- There should be no outliers, high leverage values or highly influential points.

One of the assumptions in logistic regression is explanatory variables should not be highly correlated with each other. The logistic regression model must satisfy the assumptions in order to valid the results. Unless model may have problems, such as biased coefficient estimates or very large standard errors for the logistic regression coefficients, and these problems may lead to invalid statistical inferences. Therefore, it is needed to check the underlying assumptions involved in logistic regression before making any statistical inference [2]. Therefore, it is important to test the multicolinearity in logistic regression to valid the results. In this study we focus on detection of multicolinearity problems among the explanatory variables.

2. MATERIAL AND METHODS

Binary logistic regression model estimates the probability of occurrence of an event by fitting data to a logistic curve. The dependent variable is the population proportion or probability that the resulting outcome is equal to 1. Parameters obtained for the independent variables can be used to estimate odds ratios for each of the independent variables in the model. The specific form of the logistic regression model is:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... \beta_n x_n}} \quad (1)$$

where π is the probability of the outcome of interest or event, β_0 is the intercept, $\beta_1, \ldots, \beta_n$ are regression coefficients, $x_1, x_2, \ldots, x_n$ are independent variables.

The transformation of the conditional mean $\pi(x)$ logistic function is known as the logit transformation:

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \quad (2)$$

The importance of the logit transformation is that it is linear in its parameters and may range from $-\infty$ to $+\infty$.

2.1 Pearson Correlation Coefficient

Usually we use Pearson's correlation coefficient to measure the strength of the association between two variables. The general rule of thumb is that if correlation coefficient between two variables is greater than 0.8 or 0.9, the multicollinearity is a serious problem. Formula of sample correlation coefficient is described as follows.

$$r = \frac{\sum x_i y_i - n\overline{xy}}{\sqrt{\left(\sum x_i^2 - n\overline{x}^2\right)}\sqrt{\left(\sum y_i^2 - n\overline{y}^2\right)}} \quad (3)$$

2.2 Tolerance

Furthermore, multicolinearity can mainly be detected with the help of tolerance and its reciprocal, called variance inflation factor (VIF). The tolerance is the percentage of the variance in a given predictor that cannot be explained by the other predictors.

By definition tolerance of any specific explanatory variable is

$$\text{Tolerance} = 1-R^2 \quad (4)$$

where R^2 is the coefficient of determination for the regression of that explanatory variable on all remaining independent variables. Tolerance close to 1 indicates that there is little multicollinearity, whereas a value close to zero suggests that multicollinearity may be a threat. There is no formal cutoff value to use with tolerance for determining presence of multicollinearity [2]. Myers [3] suggests a tolerance value below 0.1 indicates serious collinearity problem and Menard [4,5] suggests that a tolerance value less than 0.2 indicates a potential collinearity problem. As a rule of thumb, a tolerance of 0.1 or less is a cause for concern.

2.3 VIF

The VIF is defined as the reciprocal of tolerance as

$$VIF = \frac{1}{TOLERANCE} \quad (5)$$

VIF shows that how much the variance of the coefficient estimate is being inflated by multicollinearity. Like tolerance there is no formal cutoff value to use with VIF for determining the presence of multicollinearity. Values of VIF exceeding 10 are often regarded as indicating multicollinearity, but in weaker models, which is often the case in logistic regression; values above 2.5 may be a cause for concern [2].
From equation (2), VIF shows us how much the variance of the coefficient estimate is being inflated by multicollinearity. The square root of VIF tells us how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other explanatory variables in the equation. Like tolerance there is no formal cutoff value to use with VIF for determining the presence of multicollinearity. Values of VIF exceeding 10 are often regarded as indicating multicollinearity, but in weaker models, which is often the case in logistic regression; values above 2.5 may be a cause for concern .

2.4 Eigen values, condition index and variance proportions

Moreover, eigen values for the scaled, uncentered cross-product matrix, condition indices and the variance proportions for each explanatory variable is used to identify multicollinearity. If any eigen value is larger than others, then of the regression parameters can be greatly affected by small changes in the explanatory variables or outcome. If the eigen values are fairly similar then the fitted model is likely to be unchanged by small changes in the measured variables [6].
The condition indices are computed as the square root of the ratio of the largest eigenvalue to the eigen value of interest. It is defined as

$$K = \sqrt{\frac{\lambda_{max}}{\lambda_k}} \quad (6)$$

where λ_{max} and λ_k are the maximum and the k^{th} eigen values respectively. When there is no collinearity at all, the eigen values, condition indices will equal unity. As collinearity increases, eigen values will be both greater and smaller than unity. If one or more of the

eigenvalues are small (close to zero) and the corresponding condition number large, then we have an indication of multicollinearity. There is no hard and fast rule about how much larger a condition index needs to be indicated collinearity problems. An informal rule of thumb is that if the condition index is 15, multicollinearity is a concern; if it is greater than 30, multicollinearity is a very serious concern [2].
The variance of each regression coefficient can be broken down across the eigen values. The variance proportion explains the proportion of the variance of each regression coefficient that is attributed to each eigen value [2].

2.5 Data

Motorcycle accident data were used to evaluate diagnostic measurements. Secondary data used from 2014 to 2016 in this study which consists of 32926 accidents were acquired from the Traffic Police headquarters, Colombo in Sri Lanka. In this study, it is considered only the road accidents involved motorcyclists at fault. The response variable is severity of accidents which consists of two levels namely grievous and non-grievous accidents. Explanatory variables were accident cause, time, road surface, weather condition, light condition, age of motorcyclist and location. Except age of motorcyclist variable, other variables are categorical. Dummy variables are created for those categorical variables.

3. RESULTS AND DISCUSSION

The correlation coefficients among the explanatory variables can be used as first step to identify the presence of multicollinearity. Correlation matrix of highly correlated explanatory variables presented in Table 1. It illustrates that the correlation coefficients between variables light and time as well as road surface and weather are highly correlated with each other and indicated them as bold. These high correlation coefficients signify the presence of severe multicollinearity between the explanatory variables light condition and time of accident as well as road surface and weather condition.

Table 1 : Pearson Correlation matrix between 2 explanatory variables

Variables		Time		Weather Condition		Light Condition
		Day	Night	Clear	Rainy	Night, no Street Lighting
Light Condition	Daylight	**0.971** (0.000)	**-0.971** (0.000)	0.095 (0.124)	-0.095 (0.124)	**-0.837** **(0.000)**
	Night, no Street Lighting	**-0.862** (0.000)	**0.862** (0.000)	-0.092 (0.247)	0.092 (0.247)	1.000 (0.000)
Road Surface	Dry	0.088 (0.164)	-0.088 (0.164)	**0.966** (0.000)	**-0.966** (0.000)	-0.085 (0.321)
	Wet	-0.088 (0.164)	0.088 (0.164)	**-0.966** (0.000)	**0.966** (0.000)	0.085 (0.326)

Cell value: correlation coefficient
p value

Examining the correlation matrix may be helpful but not sufficient. It is quite possible to have data in which no pair of variables has a high correlation, but several variables together may

be highly interdependent. Much better diagnostics are produced by tolerance and VIF values. Table 2 indicates the collinearity statistics.

Results of Table 2 observe that the high tolerances for the variables vehicle type, gender, validity of license, accident cause, alcohol test, weekday/weekend, location and age of driver but very low tolerances for the variables time and light condition. Similarly, the variance inflation factor corresponding to the explanatory variables vehicle type, gender, validity of license, accident cause, alcohol test, weekday/weekend, location and age of driver are very close to 1, but for variables time and light condition, the VIF are larger than 2.5. Using these collinearity statistics, it can be concluded that the data almost certainly indicates a serious collinearity problem.

Table 2: Collinearity statistics

Variables	Categories	Collinearity Statistics	
		Tolerance	VIF
Gender	Male	.994	1.030
	Female	.992	1.022
Validity of license	With license	.945	1.023
	Without license	.985	1.016
Time	Day time	**.045**	**19.456**
	Night	**.050**	**20.181**
Weekday/Weekend	Weekday	.997	1.003
	Weekend	.993	1.102
Location	Bend/Junction	.994	1.006
	Road	.997	1.004
Accident cause	Speeding	.971	1.030
	Aggressive driving	.965	1.023
	Others	.959	1.043
Road surface	Dry	**.059**	**15.457**
	Wet	**.067**	**14.927**
Weather condition	Clear	**.061**	**15.451**
	Rainy	**.067**	**14.944**
Light condition	Daylight	**.052**	**19.314**
	Night, no street lighting	**.057**	**18.654**
	Others	**.186**	**5.364**
Age	Age	.984	1.017

The collinearity diagnostics are also checked to confirm the multicollinearity and displayed in Table 3. It can be seen that a large deviation in the final two factors, with the eigenvalue resulting very close to zero and the condition index resulting quite large in comparison. Furthermore, it is observed that the largest condition index is 28.641, which is beyond the range of our rules of thumb and indicate a cause for serious concern. According to the table 3, variance in the regression coefficients of time and light condition is associated with eigen value corresponding to the dimension 11 and variance in the regression coefficients of surface and weather is associated with eigen value corresponding to the dimension 10 which clearly indicate dependency between the variables. Hence the result of this analysis clearly indicates that there is collinearity between light condition and time of accident as well as road surface and weather condition. This dependency results in the model becoming biased.

Table 3: Collinearity diagnostics

Dimension	Eigenvalue	Condition Index	Variance Proportions										
			(Constant)	Validity of license	Weekday/ weekend	Location	Gender	Accident cause	Time	Surface	Age	Weather	Light
1	4.550	1.000	.00	.01	.01	.01	.00	.01	.00	.00	.00	.00	.00
2	1.841	1.572	.00	.00	.00	.00	.00	.00	.00	.02	.00	.02	.00
3	1.090	2.043	.00	.01	.00	.01	.14	.03	.00	.00	.00	.00	.01
4	1.005	2.128	.00	.00	.00	.00	.37	.78	.00	.00	.00	.00	.00
5	.988	2.146	.00	.00	.00	.01	.10	.02	.00	.00	.92	.00	.00
6	.877	2.278	.00	.00	.00	.01	.38	.09	.00	.00	.00	.00	.01
7	.753	2.459	.00	.03	.03	.74	.00	.06	.00	.00	.00	.00	.00
8	.688	2.572	.00	.00	.54	.02	.00	.28	.00	.00	.00	.00	.00
9	.612	2.727	.00	.40	.26	.04	.00	.17	.00	.00	.00	.00	.00
10	.033	11.828	.00	.50	.14	.15	.00	.05	.00	.97	.04	.96	.00
11	.006	28.641	.98	.04	.01	.00	.00	.01	.96	.00	.00	.00	.94

3.1 Solutions to Multicollinearity

Once the collinearity between variables has been identified, the next step is to find solutions in order to remedy this problem. There are a few solutions to overcome this such that combining variables, increasing sample size, omitting highly correlated variables, ridge regression, principal component analysis [7]. Since combining variables does not make sense and increasing sample size is not possible, here, we focus for the omitting highly correlated variables. Any of the collinear variables could be omitted. There is no statistical ground for omitting one variable over another. Thus, first, time is removed from the data and repeat the analysis. However, collinearity still exists among the levels of light variable. Then time is added and light condition is removed and repeat the analysis. Moreover, weather condition is removed to reduce the multicollinearity between road surface and weather condition. Results are presented in Table 4.
According to Table 4, tolerances for all the predictors are very close to 1 and all the VIF values are smaller than 2.5. Therefore, it can be concluded that multicollinearity is not a concern when one of the correlated variables is omitted.

Collinearity diagnostics for the remaining variables are also checked and indicated in the Table 5. According to the Table 5, all the condition indices are lower than 15 and it can be concluded that multicollinearity is not a concern when one of the correlated variables is omitted. It can be seen that each explanatory variable has most of its variance loading onto a different dimension (validity of license has 42% of variance on dimension 7, weekday/weekend has 77% of the variance on dimension 6, location has 78% of the variance on dimension 5, gender has 55% of the variance on dimension 3, accident cause has 65% of the variance on dimension 9, time has 65% of the variance on dimension 2,

surface has 78% of the variance on dimension 4 and age has 56% of the variance on dimension 8). There were no such variables that have significantly high proportion of variances on the same small eigen value. This also indicates that multicollinearity is not a concern.

Table Error! No text of specified style in document.**: Collinearity statistics for remained variables**

Variables	Categories	Collinearity Statistics	
		Tolerance	VIF
Gender	Male	.993	1.007
	Female	.992	1.022
Validity of license	With license	.978	1.010
	Without license	.981	1.020
Time	Day time	.980	1.028
	Night	.975	1.026
Weekday/Weekend	Weekday	.997	1.003
	Weekend	.998	1.002
Location	Bend/Junction	.997	1.003
	Road	.998	1.004
Accident cause	Speeding	.639	1.565
	Aggressive driving	.633	1.580
	Others	.638	1.560
Road surface	Dry	.993	1.004
	Wet	.992	1.008
Age	Age	.984	1.017

Table 5: Collinearity diagnostics for remaining variables

Dimension	Eigenvalue	Condition Index	Variance Proportions								
			(Constant)	Validity of license	Weekday/ weekend	Location	Gender	Accident cause	Time	Surface	Age
1	4.412	1.000	.00	.02	.01	.01	.00	.01	.02	.00	.01
2	1.035	2.064	.00	.01	.00	.02	.21	.02	.65	.08	.00
3	.981	2.121	.00	.00	.00	.00	.55	.01	.01	.08	.00
4	.941	2.166	.00	.00	.01	.01	.18	.00	.00	.78	.00
5	.754	2.420	.00	.04	.09	.78	.02	.00	.02	.02	.00
6	.652	2.601	.00	.11	.77	.02	.00	.00	.09	.01	.00
7	.572	2.777	.00	.42	.00	.00	.02	.00	.61	.02	.00
8	.475	3.048	.01	.35	.09	.15	.02	.04	.20	.00	.56
9	.133	5.766	.98	.00	.00	.00	.00	.65	.00	.00	.39

Therefore, it can be safely concluded that multicollinearity is no more a problem to fit the binary logistic regression model. Hence the intensive analysis and fitting of the binary logistic regression model after minimizing the collinearity problems may produce stable and unbiased model to predict the outcome variable [2].

4. CONCLUSION

One of the problems in binary logistic regression model typically arise is that explanatory variables of the logistic regression model are highly correlated among themselves. 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. Therefore, researchers always try to remove the multicollinearity among explanatory variables.

The range of solutions available for logistic regression, such as omitting highly correlated variables, ridge regression, combining variables into an index, and increasing the sample size. Since combining variables does not make sense and increasing sample size is very expensive solution to multicollinearity, even though it gives a viable solution, we focus for the omitting highly correlated variables. It can be seen that multicollinearity is not a problem after omitting highly correlated variables. Hence, reliable and valid predictive logistic regression model can be built based on the adequate inspection and measures of remedy taken against multicollinearity.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

[1] Field A. Discovering Statistics Using SPSS. 3rd ed. California: SAGE Publications Inc; 2009.

[2] Midi H, Sarkar S, Rana S. Collinearity diagnostics of binary logistic regression model. Journal of Interdisciplinary Mathematics. 2013; 253-267.

[3] Mayers RH. Classical and Modern Regression with Applications. PWS-Kent Publishing Company; 1990.

[4] Menard S. Applied Logistic Regression Analysis. 2nd ed. A Sage University paper; 2002.

[5] Menar S. An Introduction to Logistic Regression Diagnostics. 1st ed. Thousand Oaks: SAGE Publications Inc; 2011.

[6] Rana S, Midi H, Sarkar S. Validation and Performance Analysis of Binary Logistic Regression Model. Proceedings of the WSEAS International Conference on Environment, Medicine and Health Sciences. 2010; 51-55.

[7] Senaviratna NAMR, Cooray TMJA. Detecting Multicollinearity of Binary Logistic Regression Model. Second International conference on Multidisciplinary Research, Sri Lanka. 2018; 15..