International Multidisciplinary Research Journal - ISSN 2424-7073
Agriculture & Climate Change – (July 2019): Pages 58-64

Detecting Multicollinearity of Binary Logistic Regression Model: An Analysis of Motorcycle Accidents in Sri Lanka

N.A.M.R. Senaviratna[1] and T.M.J.A. Cooray[2]
[1]*The Open University of Sri Lanka, Nugegoda, Sri Lanka*
[2]*University of Moratuwa, Moratuwa, Sri Lanka*

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 can cause unstable estimates and inaccurate variances which 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 are 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 which consists of two levels namely grievous and non-grievous. Explanatory variables were accident cause, time, road surface, weather condition, light condition and location. Multicollinearity 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, combining variables into an index, and testing hypothesis about sets of variables. It is safely concluded that without increasing sample size, to omit one of the correlated variables can reduce multicollinearity considerably.

Keywords: Logistic Regression, Multicollinearity, VIF.

Introduction

Binary logistic regression is used to model the relationship between dichotomous dependent variable and multiple independent variables which are either continuous or categorical. It 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:

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

where π is the probability of the outcome of interest or event, β_0 is the intercept, β_1, …, β_n are regression coefficients, x_1, x_2, …, x_n are independent variables.

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

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

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

Method and Materials

One of the assumptions in logistic regression is predictor variables should be uncorrelated. 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 [1].

Variable selection or reduction such as stepwise selection has been the most common approach to avoid multicollinearity, but optimal variable reduction requires accurate assessment of relative variable importance. Unfortunately, this assessment of variable importance is itself corrupted by multicollinearity. Hence, multicollinearity makes optimal variable selection very difficult in standard logistic regression modeling [6].

In this study we will focus on detection of multicollinearity problems among the explanatory variables and discuss few collinearity diagnostics commonly used in multiple logistic regression.

Correlation Coefficients

The diagnostic tool considered primarily for identifying multicollinearity-namely, the pairwise coefficients of simple correlation between the predictor variables is frequently helpful. Often, however, serious multicollinearity exists without being disclosed by the pairwise correlation coefficients [2]. The general rule of thumb is that if simple correlation coefficient between two regressors is greater than 0.8 or 0.9, the multicollinearity is a serious problem [3].

Tolerance

By definition tolerance of any specific explanatory variable is,

$$\text{Tolerance} = 1\text{-}R^2$$

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 [4].

VIF

The VIF shows that how much the variance of the coefficient estimate is being inflated by multicollinearity.

It is defined as the reciprocal of tolerance as,

$$VIF = \frac{1}{TOLERANCE}$$

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 [4].

Eigen Values and Condition Index

Sometimes eigenvalues and condition index are referred to when examining multicollinearity. The condition index (k) is the square root of the ratio of the largest eigen value (λ_{max}) to the eigen value of interest (λ_k), i.e.

$$k = \sqrt{\frac{\lambda_{max}}{\lambda_k}}$$

When there is no collinearity at all, the eigenvalues and condition indices will all equal one. As collinearity increases, eigen values will be both greater and smaller than unity. Eigen values close to zero indicate a multicollinearity problem and condition indices will be increased. 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.

Variance Proportions

The variance of each regression coefficient can be broken down across the eigen values and the variance proportion tells us the proportion of the variance of each predictor regression coefficient that is attributed to each eigen value. we are looking for predictors that have high proportions on the same small eigen value because this would indicate that the variances of their regression coefficients are dependent.

Results and Discussion

Correlation matrix of highly correlated explanatory variables presented in Table 1.

Table 1: Pearson Correlation matrix between 2 explanatory variables.

Variables		Time		Weather Condition		Light Condition
		DT	NT	CL	RA	GSL
Light Condition	DL	**0.971** (0.000)	**-0.971** (0.000)	0.095 (0.124)	-0.095 (0.124)	**-0.837** **(0.000)**
	GSL	**-0.862** (0.000)	**0.862** (0.000)	-0.092 (0.247)	0.092 (0.247)	1.000 (0.000)
Road Surface	D	0.088 (0.164)	-0.088 (0.164)	**0.966** (0.000)	**-0.966** (0.000)	-0.085 (0.321)
	W	-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

Table 1 shows 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. 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. Table 2 indicates the collinearity statistics for each levels of factors.

Table 2: Collinearity statistics.

Factor	Model	Collinearity Statistics	
		Tolerance	VIF
Time	**DT**	**.045**	**19.456**
	NT	**.050**	**20.181**
Location	RD	.997	1.004
	BJ	.994	1.006
Accident cause	Cause1	.971	1.030
	Cause2	.965	1.023
	OT	.959	1.043
Road surface	**D**	**.059**	**15.457**
	RO	**.067**	**14.927**
Weather	**CL**	**.061**	**15.451**
	WO	**.067**	**14.944**
Light condition	**DL**	**.052**	**19.314**
	NSL	**.057**	**18.654**
	LO	.186	5.364

Table 2 observes the high tolerances for the variables accident cause and location but very low tolerances for the variables time, light condition, road surface and weather condition. Similarly, the variance inflation factor (VIF) corresponding to the explanatory variables accident cause and location are very close to 1, but for variables time, light condition, road surface and weather 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. Following Table 3 illustrates the eigen values, condition indices and variance proportions for all the predictors. 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. The collinearity diagnostics confirm that there are serious problems with multicollinearity. Eigen values in the bottom few lines are close to 0, indicating that the predictors are highly intercorrelated and that small changes in the data values may lead to large changes in the estimates of the coefficients. In this table on the average 90% of the variance in the regression coefficients of weather and road surface is associated with eigen value corresponding to the dimension 12. Similarly, time and light condition is associated with eigen value corresponding to the dimension 13 which clearly indicates dependency between the variables. Hence the result of this analysis clearly indicates that there is collinearity between weather and road surface as well as time and light condition.

Table 3: Collinearity diagnostics table.

Dimension	Eigenvalue	Condition Index	Variance Proportions														
			(Constant)	RD	BJ	Cause1	Cause2	OT	DT	NT	D	RO	CL	WO	LO	DL	NSL
1	4.550	1.000	.00	.01	.01	.00	.01	.01	.00	.01	.01	.00	.00	.00	.01	.00	.00
2	1.841	1.572	.00	.00	.00	.00	.00	.00	.00	.00	.00	.02	.00	.02	.00	.00	.00
3	1.090	2.043	.00	.00	.01	.14	.03	.14	.00	.01	.01	.00	.00	.00	.01	.01	.00
4	1.005	2.128	.00	.00	.00	.37	.70	.21	.00	.00	.00	.00	.00	.00	.00	.00	.00
5	.988	2.146	.00	.00	.01	.10	.02	.68	.00	.00	.00	.00	.00	.00	.00	.00	.13
6	.877	2.278	.00	.00	.01	.38	.09	.05	.00	.00	.00	.00	.00	.00	.00	.01	.02
7	.753	2.459	.00	.03	.74	.00	.06	.07	.00	.03	.03	.00	.00	.00	.03	.00	.00
8	.688	2.572	.00	.54	.02	.00	.28	.24	.00	.00	.00	.00	.00	.00	.00	.00	.00
9	.612	2.727	.00	.26	.04	.00	.17	.23	.00	.00	.00	.00	.00	.00	.00	.00	.00
10	.476	3.092	.00	.14	.15	.00	.05	.02	.00	.00	.00	.00	.04	.00	.00	.00	.00
11	.082	7.443	.02	.01	.00	.00	.01	.00	.03	.04	.04	.00	.00	.00	.04	.03	.02
12	.033	11.828	.00	.00	.00	.00	.00	.00	.00	.00	.96	.98	.97	.98	.00	.00	.00
13	.006	28.641	.98	.00	.00	.00	.00	.00	.96	.95	.00	.00	.02	.00	.81	.94	.92

Reducing Multicollinearity

Once the collinearity between variables has been identified, the next step is to find solutions in order to remedy this problem. In some cases, variables involved in multicollinearity can be combined into a single variable will solve the problem. If combining variables does not make sense, then some variables causing multicollinearity need to be dropped from the model. There is no statistical ground for omitting one variable over another. Examining the correlations between the variables and taking into account practical aspects and importance of the variables help in making a decision what variables to drop from the model. However, the solutions to multicollinearity are summarized as follows.

In this study, the sample size is large enough (n = 32926) and the dummy variables are created properly. In order to minimize the multicollinearity, dropping the highly correlated variables is a fine idea. Thus, first, variables time and road surface are removed from the data and repeat the analysis. However, collinearity still exists among the levels of light variable and weather condition. Then time and road surface variables are added and light condition and weather are removed from the analysis and repeat the analysis. Table 4 and Table 5 present the collinearity statistics and collinearity diagnostics respectively when light condition and weather variables are removed.

Table 4: Collinearity statistics of remained variables.

Factor	Model	Collinearity Statistics	
		Tolerance	VIF
Time	DT	.980	1.028
	NT	.975	1.026
Location	RD	.998	1.004
	BJ	.997	1.003
Accident cause	Cause1	.639	1.565
	Cause2	.633	1.580
	OT	.638	1.560
Road surface	D	.993	1.004
	RO	.992	1.008

Table 5: Collinearity diagnostics table.

Dimension	Eigenvalue	Condition Index	Variance Proportions									
			(Constant)	RD	BJ	Cause1	Cause2	OT	DT	NT	D	RO
1	4.412	1.000	.00	.01	.01	.02	.01	.00	.02	.00	.00	.01
2	1.035	2.064	.00	.00	.02	.42	.02	.26	.02	.21	.08	.00
3	.981	2.121	.00	.00	.00	.00	.01	.16	.01	.55	.08	.00
4	.941	2.166	.00	.01	.01	.00	.00	.01	.00	.18	.78	.00
5	.754	2.420	.00	.09	.78	.04	.00	.01	.02	.02	.02	.00
6	.652	2.601	.00	.77	.02	.11	.00	.00	.09	.00	.01	.00
7	.572	2.777	.00	.00	.00	.02	.00	.00	.61	.02	.02	.00
8	.475	3.048	.01	.09	.15	.35	.65	.08	.20	.02	.00	.04
9	.133	5.766	.01	.00	.00	.00	.04	.37	.00	.00	.00	.39
10	.047	9.695	.98	.02	.01	.05	.28	.10	.02	.00	.00	.56

After dropping the variables light condition and weather, the warning signs due to multicollinearity were not observed in the output. It can be observed from Table 4 that the tolerances for all the predictors are close to 1 and all the VIF values are smaller than 2.5 which indicates that multicollinearity is not a cause for further concern.

Table 5 displays the eigen values, condition indices and distribution of variance proportions across the different dimensions of eigen values after dropping two correlated variables. According to the informal rule of thumb, all the condition indices are lower than 15 and we may conclude that multicollinearity is not a concern anymore.

For this reduced model we can see that each predictor has most of its variance loading onto a different dimension. (RD has 77% of variance on dimension 6, BJ has 78% of the variance on dimension 5, Cause1 has 42% of the variance on dimension 2, Cause2 has 65% of the variance on dimension 8, OT has 37% of the variance on dimension 9, DT has 61% of the variance on dimension 7, NT has 55% of the variance on dimension 3, D has 78% of the variance on dimension 4 and RO has 56% of the variance on dimension 10). There were no such predictors that have significantly high proportion of variances on the same small eigen value. This also indicates that the variances of the regression coefficients are independent and multicollinearity is not a concern. Therefore, it can be 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 reducing the collinearity problems may produce stable and unbiased model to predict the outcome variable.

Conclusion

In this study, different measurements including tolerance, VIF, condition index, eigen values and variance proportions to detect multicollinearity in logistic regression model are discussed. After discovering the existence of multicollinearity, the range of solutions available such as dropping variables, increasing the sample size and testing hypothesis about sets of variables. Since increasing sample size is expensive and difficult, dropping correlated variables method is applied to this analysis. After dropping the correlated variables in this study, it can be seen that multicollinearity is not a problem anymore. Thus, binary logistic regression can be applied for further analysis.

References

Allison P.D. 2001. *Logistic Regression Using the SAS system: Theory and Applications* Cary, NC: SAS Institute Inc.

Kutner M.H. *Applied Linear Statistical Models*. Irwin, McGraw-Hill, 2005.

Mayers R.H. *Classical and Modern Regression with Applications*. PWS-Kent Publishing Company, 1990.

Midi H., Sarkar S. K., Rana S. *Collinearity diagnostics of binary logistic regression model*. Journal of Interdisciplinary Mathematics, pp253-267.

Sarkar S.K., Midi H., Rana S. *Detection of outliers and influential observations in binary logistic regression: An empirical study.* Journal of Applied Sciences, pp26-35.

Pampel F.C. *Logistic Regression: A Primer* Sage Publications, 2000.

Appendix
List of Abbreviation

Abbreviation	Description	Abbreviation	Description
BJ	Bend/Junction	RD	Road
Cause 1	Speeding	D	Dry
Cause2	Aggressive/ negligent driving	DL	Daylight
OT	Others	DT	Day Time
GSL	Night, Good street lighting	NT	Night Time
NSL	Night, no street lighting		