Collinearity diagnostics of binary logistic regression model

Article *in* Journal of Interdisciplinary Mathematics · May 2013

DOI: 10.1080/09720502.2010.10700699

3 authors, including:

Saroje Kumar Sarkar
University of Rajshahi, Bangladesh.
23 PUBLICATIONS **440** CITATIONS
SEE PROFILE

Sohel Rana
Universiti Putra Malaysia
44 PUBLICATIONS **475** CITATIONS
SEE PROFILE

Collinearity diagnostics of binary logistic regression model

Habshah Midi [*]
S.K. Sarkar [†]
Sohel Rana [‡]

Laboratory of Applied and Computational Statistics
Institute for Mathematical Research
University Putra Malaysia
43400 Serdang, Selangor
Malaysia

Abstract

Multicollinearity is a statistical phenomenon in which predictor variables in a logistic regression model are highly correlated. It is not uncommon when there are a large number of covariates in the model. Multicollinearity has been the thousand pounds monster in statistical modeling. Taming this monster has proven to be one of the great challenges of statistical modeling research. Multicollinearity can cause unstable estimates and inaccurate variances which affects confidence intervals and hypothesis tests. The existence of collinearity inflates the variances of the parameter estimates, and consequently incorrect inferences about relationships between explanatory and response variables. Examining the correlation matrix may be helpful to detect multicollinearity but not sufficient. Much better diagnostics are produced by linear regression with the option tolerance, Vif, condition indices and variance proportions. For moderate to large sample sizes, the approach to drop one of the correlated variables was established entirely satisfactory to reduce multicollinearity. On the light of different collinearity diagnostics, we may safely conclude that without increasing sample size, the second choice to omit one of the correlated variables can reduce multicollinearity to a great extent.

Keywords and phrases : *Collinearity, tolerance, variance inflation factor, condition index, eigen value, variance proportion.*

[*] *E-mail*: habshahmidi@hotmail.com

[†] *E-mail*: sarojeu@yahoo.com

[‡] *E-mail*: srana_stat@yahoo.com

Journal of Interdisciplinary Mathematics
Vol. 13 (2010), No. 3, pp. 253–267
© Taru Publications

1. Introduction

Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable. It is a variation of ordinary linear regression which is used when the response variable is a dichotomous variable and the independent variables are continuous, categorical, or both. Unlike ordinary linear regression, logistic regression does not assume that the relationship between the independent variables and the dependent variable is linear. It is a useful tool for analyzing data that includes categorical response variables. The use of logistic regression modeling has exploded during the last decade. From its original acceptance in epidemiological research, the method is now commonly employed in many fields. Social scientists and demographers frequently want to estimate regression model in which the dependent variable is dichotomous. Nowadays, most researchers are aware of the fact that it is not appropriate to use ordinary linear regression for a dichotomous dependent variable and the usage of logistic regression as a predictive model is a better choice. The difference between logistic and linear regression is reflected both in the choice of a parametric model and in the assumptions. Once this difference is accounted for, the methods employed in an analysis using logistic regression follow the same general principles used in linear regression. Thus, the techniques used in linear regression analysis will motivate our approach to logistic regression [3].

In order for our analysis to be valid, our model has to satisfy the assumptions of logistic regression. When the assumptions of logistic regression analysis are not met, we 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, before we can use our model to make any statistical inference, we need to check the underlying assumptions involved in logistic regression [12]. The general assumptions involved in logistic regression analysis are

- The true conditional probabilities are a logistic function of the explanatory variables.
- No important variables are omitted.
- No extraneous variables are included.
- The explanatory variables are measured without error.
- The observations are independent.

- The explanatory variables are not linear combinations of each other.
- Errors are binomially distributed.

One of the attractive features about logistic regression analysis is that it is so much like the ordinary linear regression analysis. Unfortunately, some of the less pleasant features of linear regression analysis also carry over to logistic regression analysis. One of these is multicollinearity, which occurs when there are strong linear dependencies among the explanatory variables. The basic point is that, if two or more variables are highly correlated with one another, it is hard to get good estimates of their distinct effects on some dependent variable. The violation of the above sixth assumption introduced multicollinearity in the analysis. Multicollinearity is a statistical phenomenon in which two or more predictor variables in a multiple logistic regression model are highly correlated or associated. More clearly, a set of variables is collinear or ill conditioning if there exists one or more linear relationships among the explanatory variables. In this situation the coefficient estimates may change erratically in response to small changes in the model or the data. Although multicollinearity does not bias coefficients, it does make them unstable [2]. We have perfect multicollinearity if the correlation between two independent variables is equal to 1 or -1. In practice, we rarely face perfect multicollinearity in a data set. More commonly, the issue of multicollinearity arises when there is a high degree of correlation between two or more explanatory variables. 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. Multicollinearity does not reduce the predictive power or reliability of the model as a whole; it only affects calculations regarding individual predictors. That is, a multiple logistic regression model with correlated predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with others [10].

Multicollinearity has been the thousand pounds Monster in statistical modeling. Problems related to multicollinearity error are seen in all predictive modeling approaches and not just in logistic regression. The overfitting error that is associated with multicollinearity can be very costly in science applications. Yet, taming this monster has proven to be one of the great challenges of statistical modeling research. 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 [4].

Thus the existence of multicollinearity inflates the variances of the parameter estimates. That may result particularly for small and moderate sample sizes, in lack of statistical significance of individual explanatory variables while the overall model may be strongly significant. Multicollinearity may also result in wrong signs and magnitudes of logistic regression coefficient estimates, and consequently incorrect conclusions about relationships between explanatory and response variables. Multicollinearity can result in several more problems. For example, the partial logistic regression coefficient due to collinearity may not be estimated precisely. Multicollinearity makes it tedious to assess the relative importance of the explanatory variables in explaining the variation caused by the response variable. In the presence of high multicollinearity, the confidence intervals of the coefficients tend to become very wide and the statistics tend to be very small. Then it becomes difficult to reject the null hypothesis of any study when multicollinearity is present in the data. Therefore, multicollinearity is considered a disturbance that causes volatility in data [8].

In this study we will focus on detection of multicolinearity problems among the explanatory variables and introduce few collinearity diagnostics commonly used in multiple logistic regression. Using various collinearity diagnostics we would like to eliminate the consequences of multicollinearity and rectify the binary logistic regression model to draw valid statistical inferences regarding the regression parameters.

2. Materials and methods

The Bangladesh Demographic and Health Survey (BDHS-2004) is part of the worldwide Demographic and Health Surveys program, which is designed to collect data on fertility, family planning, maternal and child health. The BDHS-2004 is intended to serve as a source of population and health data for policymakers and the research community. The survey was conducted under the right of the National Institute for Population Research and Training (NIPORT) of the ministry of health and family welfare. Macro International Inc. of Calverton, Maryland, USA provided technical assistance to the project as part of its International Demographical and Health Surveys program and financial assistance was also provided by The United States Agency for International Development (USAID). A total of 10,523 households were selected for the sample, of which 10,500 were successfully interviewed. In those households, 11,601 women were identified as eligible for the individual interview and interviews were completed for 11,440 of them. But in this analysis there are only 2,212 eligible women those are able to bear and desire more children are considered. The women under sterilization, declared in fecund, divorced, widowed, having more than and less than two living children are not involved in the analysis. Those women who have two living children and able to bear and desire more children are only considered here during the period of global two children campaign.

In BDHS-2004, there are three types of questionnaires, namely the household's, women's and men's. The information's obtained from the field are recorded mainly in their respective data files. In this study, the information's corresponding to the women's data file is used. The variable age of the respondent, fertility preference, place of residence, level of education, working status of women, highest year of education, and expected number of children are considered in the analysis. The variable fertility preference involving responses corresponding to the question, would you like to have (a/another) child? The responses are coded 0 for 'no more' and 1 for 'have another' is considered as desire for children which is the binary response variable (Y) in the analysis. The age of the respondent (X_1), place of residence (X_2) is coded 0 for 'urban' and 1 for 'rural', educational level (X_3) is coded 0 for 'no education' 1 for 'primary level' 2 for 'secondary level' and 3 for 'higher level', highest year of education (X_4), working status of respondent (X_5) is coded 0 for 'not working' and 1 for 'working', and expected number of children (X_6) is coded 0 for 'two or less' and 1 for 'more than two' are considered as covariates in the binary logistic regression model. By reference cell coding technique, the three design variables X_{3}_n, X_{3}_p and X_{3}_s corresponding to 'no education', 'primary' and 'secondary' level of education are created for X_3 using 'higher' level of education as reference category.

There are certain reasons or causes to occur multicollinearity. Multicollinearity is caused by an inaccurate use of design variables; it is caused by the inclusion of a variable which is computed from other variables in the equation; multicollinearity can also result from the repetition of the same kind of variables; it is generally occurs when the variables are highly and truly correlated to each other.

In some situation, when no pair of variables is highly correlated, but several variables are involved in interdependencies, it may not be sufficient. It is better to use multicollinearity diagnostic statistics produced by linear regression analysis. For nominal explanatory variables, we should create dummy variables for each category except the reference category which are generally known as design variables. Use the dependent variable from logistic regression analysis or any other variable that is not one of the explanatory variables, as a dependent variable in the linear regression. The collinearity diagnostic statistics are based on the explanatory variables only, so the choice of the dependent variable does not matter. There are certain signals which help the researcher to detect the degree of multicollinearity. One such signal is if the individual outcome of a statistic is not significant but the overall outcome of the statistic is significant. In this instance, the researcher might get a mix of significant and insignificant results that show the presence of multicollinearity. Mathematically multicollinearity can mainly be detected with the help of tolerance and its reciprocal, called *variance inflation factor* (VIF). By definition tolerance [2] of any specific explanatory variable is

$$\text{Tolerance} = \text{Tol} = 1 - R^2 \,. \tag{1}$$

where R^2 is the coefficient of determination for the regression of that explanatory variable on all remaining independent variables. The variance inflation factor [2] is defined as the reciprocal of tolerance as

$$\text{Vif} = \frac{1}{\text{Tol}} = \frac{1}{1 - R^2} \,. \tag{2}$$

To determine the impact of multicollinearity on estimated standard error of the regression coefficient, let $U = (X_1, X_2, \ldots, X_p)$ be the set of all explanatory variables and $V = (X_1, X_2, \ldots X_{k-1}, X_{k+1}, \ldots X_p)$ be the set of all explanatory variables except X_k. The standard error of the regressor corresponding to the explanatory variable X_k is denoted as $s_{\hat{\beta}_{X_k}}$ and defined [2] as

$$\begin{aligned} s_{\hat{\beta}_{X_k}} &= \sqrt{\frac{1 - R^2_{YU}}{(1 - R^2_{X_k V})(N - k - 1)}}\, \frac{s_Y}{s_{X_k}} \\ &= \sqrt{\frac{1 - R^2_{YU}}{\text{Tol}_k (N - k - 1)}}\, \frac{s_Y}{s_{X_k}} = \sqrt{\text{Vif}_k}\, \sqrt{\frac{1 - R^2_{YU}}{(N - k - 1)}}\, \frac{s_Y}{s_{X_k}} \,. \end{aligned} \tag{3}$$

The equation (3) tells us how much inflated the standard error of the coefficient is compared to what it would be if the variable were uncorrelated with any other variable in the model. The larger $R^2_{X_k V}$ is (more highly correlated X_k is with the other explanatory variables in the model), the larger the standard error will be. Indeed, if X_k is perfectly correlated with the other explanatory variables, the standard errors become infinite and the solution to the model becomes indeterminate. This is referred to as the problem of multicollinearity. The problem is that, as the X's become more highly correlated, it becomes more and more difficult to determine which X is actually producing the effect on Y. Recall from equation (1) $1 - R^2_{X_k V}$ is referred to as tolerance of X_k. A 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. Myers [10] suggests a tolerance value below 0.1 indicates serious collinearity problem and Menard [11] 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.

From equation (2) the variance inflation factor 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 [7].

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 [5].

Correlation matrix presented in Table 1 is obtained from the SPSS output and the correlation coefficients among the explanatory variables can be used as first step to identify the presence of multicollinearity [1]. It is mentioned earlier about the rule of thumb that if simple correlation coefficient is greater than 0.8 or 0.9 then multicollinearity is a serious concern. The simple correlation coefficients between three design variables of X_3 and X_4 are highly correlated and indicated them as bold in Table 1. These high correlation coefficients signify the presence of severe multicollinearity between the explanatory variable X_3 and X_4.

Table 1

Pearson correlation matrix between the predictors for the data BDHS-2004

	X_6	X_{3_s}	X_2	X_5	X_1	X_{3_n}	X_4	X_{3_p}
X_6	1.000	−.005	−.049	.022	.129	−.008	.001	-.010
X_{3_s}		1.000	−.078	.178	.056	**.861**	**.705**	**.893**
X_2			1.000	−.025	.098	−.058	.021	−.070
X_5				1.000	−.097	.061	−.012	.108
X_1					1.000	−.018	−.062	.017
X_{3_n}						1.000	**.933**	**.959**
X_4							1.000	**.880**
X_{3_p}								1.000

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 linear regression with the option tolerance, Vif, eigen values, condition indices and variance proportions.

The SPSS output in Table 2 gives the collinearity statistics. In this table we observe the high tolerances for the variables X_1, X_2, X_5 and X_6 but very low tolerances for the three design variables of X_3 and X_4. Similarly the variance inflation factor corresponding to the explanatory variables X_1, X_2, X_5 and X_6 are very close to 1 but for three design variables $X_{3\_n}, X_{3\_p}, X_{3\_s}$ and X_4 the Vif are larger than 2.5. Using these collinearity statistics we may conclude that the data almost certainly indicates a serious collinearity problem.

Sometimes eigen values for the scaled, uncentered cross-product matrix, condition indices and the variance proportions for each predictor will be referred to examine as multicollinearity diagnostics. If any of the eigen values in the output are much larger than others then the uncentered cross-product matrix is said to be ill-conditioned, which means that the solution of the regression parameters can be greatly affected by small changes in the predictors or outcome. These values give us some idea as to how accurate our logistic regression model is? If the eigen values are fairly similar then the derived model is likely to be unchanged by small changes in the measured variables.

Table 2

Detection of multicollinearity based on collinearity statistics

Explana-	Unstandardized		Standardized			Collinearity	
tory	coefficient		coefficient	t-statistic	p-value	statistics	
variables	β	S.E.	β			Tolerance	Vif
Intercept	.600	.109		5.533	.000		
X_1	$-.008$	.001	$-.112$	-5.833	.000	.913	1.095
X_2	.077	019	.077	4.038	.000	.925	1.081
X_{3}_n	$-.109$	.101	$-.103$	-1.074	.283	.037	27.264
X_{3}_p	$-.063$	.077	$-.061$	$-.818$	.414	.061	16.510
X_{3}_s	$-.036$	.052	$-.034$	$-.692$	.489	.142	7.037
X_4	$-.021$	.008	$-.177$	-2.801	.005	.085	11.799
X_5	$-.049$	.022	$-.043$	-2.238	.025	.901	1.109
X_6	.463	.020	.424	22.733	.000	.973	1.028

The condition index are another way of expressing these eigen values and represent the square root of the ratio of the largest eigen value to the eigen value of interest. Mathematically, suppose $\lambda_{\max}$ and λ_k be the maximum and the kth eigen values respectively then condition index or condition number for the kth dimension is defined as

$$K = \sqrt{\frac{\lambda_{\max}}{\lambda_k}} \,. \tag{4}$$

When there is no collinearity at all, the eigen values, condition indices or condition number will all equal unity. 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. There is no hard and fast rules 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. The output of our study is exhibited in Table 3 as collinearity diagnostics. It is observed that the largest condition index is 33.161, which is beyond the range of our rules of thumb and indicate a cause for serious concern.

Final step in analyzing the output of Table 3 is to look at the 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. The proportions can be converted to percentages by multiplying them by 100 to simplify interpretation. In terms collinearity, 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. So we are interested mainly in the bottom few rows of the Table 3. In this table on the average 80% of the variance in the regression coefficients of X_3 and X_4 is associated with eigen value corresponding to the dimension 9 which clearly indicates dependency between the variables. Hence the result of this analysis clearly indicates that there is collinearity between X_3 and X_4 and this dependency results in the model becoming biased.

Table 3

Detection of multicollinearity based on collinearity diagnostics

Dimension	Eigen	Condition	Variance proportion								
	value	index	Constant	X_1	X_2	X_3_n	X_3_p	X_3_s	X_4	X_5	X_6
1	4.727	1.000	.00	.00	.01	.01	.00	.00	.00	.01	.01
2	1.302	1.906	.00	.00	.01	.00	.00	.02	.01	.01	.01
3	1.008	2.165	.00	.00	.00	.00	.03	.01	.00	.02	.01
4	.802	2.428	.00	.00	.01	.00	.00	.01	.00	.49	.29
5	.625	2.750	.00	.00	.04	.00	.00	.00	.00	.26	.65
6	.304	3.941	.00	.02	.69	.00	.00	.01	.01	.11	.00
7	.193	4.954	.00	.00	.21	.00	.03	.18	.07	.10	.00
8	.034	11.757	.02	.92	.04	.08	.06	.02	.10	.01	.03
9	.004	33.161	.98	.06	.00	.90	.89	.74	.81	.01	.00

3. Solutions to 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. The problem with this should be obvious; there is no way of knowing which variable to omit. The resulting theoretical conclusions are meaningless because, statistically speaking; any of the collinear variables could be omitted. 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. Even if a predictor is removed, O. Connell and A. Ann [6] recommended replacing the omitted variable with another equally important predictor that does not have strong multicollinearity. They also suggested collecting more data to see whether the multicollinearity can be lessened. Indeed, it may not be feasible for the researchers to follow their recommendation due to different shortcomings of data collection. However, the solutions to multicollinearity are summarized as follows.

- The researcher should remove all but one of the highly correlated variables from the analysis. This means there is no need to retain redundant variables. Thus an explanatory variable may be dropped to produce a model with significant coefficients.
- It should make sure the researcher have not made any flagrant errors, as for example, improper use of design variables.
- If possible, the researcher should increase the sample size. This will usually decrease standard errors, and make it less likely that results are some sort of sampling fluke.
- All highly correlated variables can be deleted and re-introduced in the model in the form of an interaction or as a composite variable.
- Transform the independent variables by centering and this will reduce multicollinearity.
- Another possibility is that when there are several predictors involved in multicollinearity is to run a factor analysis on these predictors and to use the resulting factor scores as predictors.
- It may be that the best thing to do is simply to realize that multicollinearity is present, and be aware of its consequences.

In the present study, the sample size is large [9] enough $(n = 2212)$ and the design variables are created properly. In order to minimize the multicollinearity, dropping the offending variable is a fine idea and we should drop one of the highly correlated variables. Based on the information of the prior research we know that the level of education is not significantly associated with the outcome variable but the highest year of education is negatively associated with the outcome. To overcome the multicollinearity problem we may drop the level of education (X_3) from the covariates. Table 4 and Table 5 present the collinearity statistics and collinearity diagnostics respectively when variable X_3 is deleted.

After dropping the variable X_3, the warning signs due to multicollinearity were not observed in the output. That is, the individual coefficients are not substantially large; t-test suggests individual predictors are statistically significant and consistent with the overall F-test. It can be observed from Table 4 that the tolerances for all the predictors are very close to 1 and all the Vif values are smaller than 2.5. If we take the average of Vif values and this average is not substantially greater than 1 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 one of the 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 when one of the correlated variable is omitted.

Table 4

Detection of multicollinearity based on collinearity statistics

Explanatory variables	Unstandardized coefficient		Standardized coefficient	t-statistic	p-value	Standardized coefficient	
	β	S.E	β			Tolerance	Vif
Intercept	.507	.045		11.281	.000		
X_1	$-.009$	.001	$-.115$	-6.030	.000	.936	1.068
X_2	.077	.019	.077	4.034	.000	.931	1.074
X_4	$-.013$	.002	$-.110$	-5.757	.000	.932	1.073
X_5	$-.050$	.021	$-.045$	-2.410	.016	.968	1.033
X_6	.463	.020	.424	22.738	.000	.973	1.027

Table 5

Detection of multicollinearity based on collinearity diagnostics

Dimension	Eigen value	Condition index	Variance proportion					
			Constant	$X1$	$X2$	$X4$	$X5$	$X6$
1	3.886	1.000	.00	.00	.02	.02	.02	.02
2	.801	2.203	.00	.00	.01	.02	.40	.44
3	.607	2.529	.00	.00	.02	.07	.54	.40
4	.485	2.830	.00	.00	.31	.47	.02	.09
5	.197	4.438	.03	.07	.55	.42	.02	.01
6	.023	13.063	.96	.92	.09	.01	.01	.04

For this reduced model we can see that each predictor has most of its variance loading onto a different dimension (X_1 has 92% of variance on dimension 6, X_2 has 55% of the variance on dimension 5, X_4 has 47% of the variance on dimension 4, X_5 has 54% of the variance on dimension 3 and X_6 has 44% of the variance on dimension 2). 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, on the light of different collinearity diagnostics, we can safely conclude 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.

4. Discussion and conclusion

We noted some key problems that typically arise when the predictor variables being considered for the logistic regression model are highly correlated among themselves. To the extent that one explanatory variable is a near but not perfect linear function of another explanatory variable, the problem of multicollinearity will occur in logistic regression, as it does in OLS regression. As the explanatory variables increase in correlation with each other, the standard errors of the logit (effect) coefficients will become inflated. Multicollinearity does not significantly change the estimates of the coefficients, only their reliability. High standard errors flag possible multicollinearity. In the presence of multicollinearity, the estimate of one variable's impact on the outcome variable controlling for the others tends to be less precise than if predictors were uncorrelated with one another. Also an analyst might falsely conclude that there is linear relationship between an explanatory variable and outcome variable. The collinear variables contain the same information about the outcome variable. If nominally different' measures actually quantify the same phenomenon then they are redundant. Alternatively, if the variables are accorded different names and perhaps employ different numeric measurement scales but are highly correlated with each other, then they suffer from redundancy. A principal danger of such data redundancy is that of over fitting in logistic regression analysis. The best logistic regression models are those in which the predictor variables each correlate highly with outcome variable but correlate at most only minimally with each other. Such a model is often called low noise and will be statistically robust.

Diagnosis of multicollinearity is easy. The big problem is what to do about it. The range of solutions available for logistic regression is pretty much the same as for linear regression, such as dropping variables, combining variables into an index, and testing hypothesis about sets of variables. A traditional approach to avoid multicollinearity error is simply to increase the sample size. This will always work. Because, unlike OLS regression, logistic regression uses *maximum likelihood estimation* (MLE) rather than *ordinary least squares* (OLS) to derive parameters. MLE relies on large-sample asymptotic normality which means that reliability of estimates declines when there are few cases for each observed combination of explanatory variables. With a large enough sample size; it can be assured that the researcher will have greatly diminished problems with multicollinearity error. Unfortunately, this is very expensive solution to multicollinearity, so it is rarely if ever a viable solution. Yet, it is a very important clue to how we might fix the problem. Usually, none of the potential fix-ups is very satisfying.

Since, multicollinearity does not bias result; it just produces large standard errors in the related explanatory variables. It is important to note that multicollinearity is a property of explanatory variables not the dependent variable. So, whenever the researchers suspect multicollinearity in a logit model, just estimate the equivalent model in linear regression and specify the collinearity diagnostics. The approaches that were used to solve the problems of multicollinearity are entirely satisfactory in the vast majority of cases, but occasionally it can create more serious multicollinearity. The reason is that by dropping variables and their linear combinations should actually be adjusted by the weight matrix used in the maximum likelihood algorithm. Thus reliable and valid predictive logistic regression models can be built based on the adequate inspection and measures of remedy taken against multicollinearity.

References

[1] A. Field, *Discovering Statistics Using SPSS*, 3rd edition, SAGE Publications, Inc., Thousand Oaks, California, 2009.

[2] D. A. Belsley, E. Kuh and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, Inc., New York, 1980.

[3] D.W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd edition, Wiley, New York, 2000.

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

[5] M. H. Kutner, C. J. Nachtsheim, J. Neter and W. Li, *Applied Linear Statistical Models*, 5th edition, McGraw-Hill, Irwin, 2005.

[6] O'Connell and A. Ann, Logistic regression models for ordinal response variables, Thousand Oaks, Sage Publications, CA, *Quantitative Applications in the Social Sciences*, Vol. 146 (2005).

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

[8] P. D. Allison, Comparing logit and probit coefficients across groups, *Sociological Methods and Research*, Vol. 28(1999), pp. 186-208.

[9] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford and A. Feinstein, A simulation of the number of events per variable in logistic regression analysis, *Journal of Clinical Epidemiology*, Vol. 99 (1996), pp. 1373–1379.

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

[11] S. Menard, *Applied Logistic Regression Analysis*, 2nd edition, A Sage University paper, 2002.

[12] UCLA, *Introduction to SAS*, UCLA: Academic Technology Services, Statistical Consulting group, 2007.

Received January, 2010