The Influence of Multicollinearity in Binary Logistic Regression and Additive Models

Mustafa Cavus*
Department of Statistics Science Faculty Anadolu University, Eskisehir, Turkey –
mustafacavus@anadolu.edu.tr

Betul Kan Kilinc
Department of Statistics Science Faculty Anadolu University, Eskisehir, Turkey –
bkan@anadolu.edu.tr

Abstract

The generalized linear modeling has appeared very useful in researches to create regression models with any distribution of the dependent variable. It can be count, binary or ordinal. The binary logistic regression is a generalized linear model having a logit link function. In the present study we examine and compare from a statistical point of view the binary logistic models and additive logistic models in the presence of multicollinearity. As the collinearity inflates the variances of the parameter estimates and therefore it results incorrect inferences about the relationships between explanatory and response variables, etc. It also leads to instable estimated coefficients in generalized additive models. For this purpose, highly correlated explanatory variables with a binary response variable are fitted by generalized linear model and generalized additive logistic models (GALM) in a conducted simulation study for moderate to larger sample sizes. The results are compared by the existing two approaches using simulated datasets. Finally, their advantages and drawbacks are pointed out in conclusion.

Keywords: multicollinearity; logistic; additive.

1. Introduction

When the distribution of the response variable is not normal then performing hypothesis tests is not proper for general linear models. The generalized linear models (GLM) is an extension of general linear model and does not require normality or homogeneity in variances. The response variable in GLM follow any probability distribution from exponential family such as the Normal, Binomial, Poisson, Gamma, Negative Binomial, and etc. It can be linearly modelled with covariate by using link functions. The binary logistic regression is a GLM having a logit link acting on a binary response variable.

The logistic regression model is used to predict probability of possible outcome of response variable. Maximum likelihood estimation has gained widespread use for estimating model parameters but it is found that multicollinearity among the explanatory variables inflates the variances of the estimators. Urgan and Tez (2008) used Liu Estimator in logistic regression when the data has collinear explanatory variables. In their study a Liu type estimator is proposed that will have smaller mean squared error than the maximum likelihood estimator[2]. Godinez-Jaimes (2012) examined the effect of collinearity and the lack of overlap in the data on the logistic regression model in a simulation study with different estimators. After the simulation study, it is obtained that the degree of overlap and the level of collinearity strongly affect the bias and mean squared error of the maximum likelihood, Firth’s and Rousseeuw and Christmann’s estimators[3]. Ariffin and Midi (2014) investigated the performance of logistic ridge regression estimation technique in the presence of multicollinearity and high leverage points. The outcome of the study is that logistic ridge regression estimator fails to provide better parameter estimates in the presence of multicollinearity [4].

Petirini (2012) was to assess the degree of multicollinearity and to identify the variables involved in linear dependence relations in additive models. For this, the animal breeding data is used and VIF is used for obtaining the multicollinearity. According the result of this study, the variables associated with additive and non-additive effects are involved multicollinearity, partially due to the natural connection between these covariables as fractions of the biological types in breed composition[5]. Ma and Yan (2014) seek to inspect the nonparametric characteristics connecting the age of the driver to the relative risk of being an at-fault vehicle, in order to discover a more precise and smooth pattern of age impact with logistic additive models. There are some results achieved after this study. Briefly, these results are about age effect, sex effect and interaction effects on relative risk of being at-fault vehicle[6]. Prince and Aghajanian (2009) present an approach to gender classification based on constructing additive sums of non-linear functions of the data are then passed through the logistic function. According to previously published work, the performance of the gender classification detector increases 10% to 87%[7].

In our study we examine additive logistic regression and binary logistic regression in the presence multicollinearity. We provide GAM with at least two nonparametric smooth variables for a dictohomus response. Motivation for our study comes from the level of collinearity affects the model parameters.

In the second part this study, some essential information is introduced for Logistic Regression Model, Generalized Additive Model and Generalized Additive Logistic Model. In the simulation study, some regression models are constructed with multicollinearity data with respect to different parameters. The behaviour of the GLM and GAM in the presence of multicollinearity is analysed and interpreted.

2. Methods

In this section we formally state the acknowledge of the technical details and notation.

2.1. Generalized Additive Model

Generalized Additive Model (GAM) is a linear model in which the explanatory variables depend linearly on smooth functions of explanatory variables. GAMs were originally developed by Hastie and Tibshrani to blend the properties of GLM with additive models.[1]

A generalized additive model has the form given in Eq. (1)

$$E(Y|X_1, X_2, \dots, X_k) = f(X_1, X_2, \dots, X_k) = f_0 + f_1(X_1) + \ \dots + \ f_k(X_k) = f_0 + \sum_{i=1}^{k} f_k(X_k) \quad (1)$$

Here Y is the response variable and $X_1, X_2, \dots, X_k$ are predictors. Note that it replaces $\sum_{i=1}^{k} \beta_i X_i$ in general linear model with $\sum_{i=1}^{k} f_i(X_i)$ where f_i' s are unspecified nonparametric functions. If Eq. (1) is fitted by the expansion of basis functions, then the least squares method could be used for fitting. However, we use scatter plot smoothers to estimate the *p* functions in Eq. (1).

2.2. Binary Logistic Regression Analysis

Logistic Regression Analysis (LRA) is one of the mostly used statistical tool to determine the relationship among the binary response variable and explanatory variables. It is used also to categorize the response variable as a probabilistic classification tool. Most regression models predict the average value of response variable taking into account explanatory variables. Unlike them, LRA predicts the probability of possible outcome of response variable.

There are three most used logistic regression model are popular in research which are binary, ordinal and multinomial logistic regression models. We introduce Binary Logistic Regression Model (BLRM) because this study is only constructed on a binary response variable. BLRM is being preferred when the response variable has two different outcome. For example, the response variable has two outcome such as damaged-healthy product, occurred-not occurred case, bad-good weather and etc.

Regression line can be constructed as in Eq. (2) when the response variable is dichotomous:

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i \quad (2)$$

where $x_i' = [1, x_{i1}, x_{i2}, \ldots, x_{ik}]$ and $\beta = [\beta_0, \beta_1, \beta_2, \ldots, \beta_k]$ response variable y_i takes values 0 and 1 which are coded of two outcome. The response is distributed as binomial variable so that the response variable can not be calculated without using logit function. Where $g(x) = x_i'\beta$, the expected value of the response variable can be calculated as in Eq. (3) with logit function:

$$E(y) = \frac{e^{g(x)}}{1 + e^{g(x)}} = \frac{1}{1 + e^{-g(x)}} \quad (3)$$

2.3. Generalized Additive Logistic Model

Generalized Additive Logistic Model (GALM) can be considered as a mixture of additive and logistic models. In this model, the probability of possible outcome of response variable which is explained by smooth functions of explanatory variables. Recall the logistic regression model in Section 2.2, the expected value of binary response variable is $E(x) = P(Y = 1|X)$ to the explanatory variables in the regression model with logit function as in Eq. (4) :

$$log\left(\frac{E(x)}{1 - E(x)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (4)$$

The GALM replaces each linear term by a smooth function form as in Eq. (5) :

$$log\left(\frac{E(x)}{1 - E(x)}\right) = f_0 + f_1(X_1) + \ f_2(X_2) + \cdots + \ f_k(X_k) \quad (5)$$

where f_i’ s are unspecified nonparametric functions. The nonparametric form of f_i’ s makes the model more flexible. The GALM is a specific example of a GAM. Hence, the expected value of the response variable $E(x)$ is related to an additive function of the explanatory variables by a link function of g in Eq. (6) [8] :

$$g[E(x)] = f_0 + f_1(X_1) + \ f_2(X_2) + \cdots + \ f_k(X_k) \quad (6)$$

3. Simulation Study

We use a simulation design with n=50,100,500, p=3 and the negligible, moderate and small pairwise correlations among two covariates. The first two covariates are generated from a normal distribution with means 0, and standard deviations 1. The third covariate is generated from a uniform distribution, and the correlation coefficient 0.10, 0.50, 0.90. The response variable is generated from a binomial distribution with various coefficients such as $\beta_0 = 0.5, \beta_1 = 3, \beta_2 = -5$ and $\beta_3 = 6$. We fit models by GLM and GAM. R Software version 3.1.2 is used for simulation study. For the basis for the smooth terms, the thin plate regression spline basis used in each GAM models. The results are then compared in terms of AIC and Deviance values. Table 1 shows the estimated AIC and Deviance values of GLM and GAM models.

4. Results of the Analysis

Two models are compared by using GLM and GAM. For the different sizes of samples and correlations, deviance and AIC values are summarized in Table 1. Multivariate normal distribution is used to generate correlated variables. The algorithm is repeated for 1000 times. Additionally, the VIF values are also produced in order to check whether the multicollinearity exist or not.

Table 1: Results of the Simulation Study

				GLM		GAM	
Model	**n**	**r**	**VIF**	**AIC**	**Deviance**	**AIC**	**Deviance**
$y = 0.5 + 3x_1 - 5x_2 + 6x_3$ $x_1, x_2 \sim N$ *and correlated* $x_3 \sim U$	50	0.10	2.170	25.949	17.949	19.690	7.041
		0.50	5.012	26.448	18.448	20.276	7.631
		0.90	23.276	23.130	15.130	17.124	5.058
	100	0.10	2.360	46.349	38.349	41.703	28.668
		0.50	4.940	45.455	37.455	40.233	26.896
		0.90	16.556	42.631	34.631	37.671	24.468
	500	0.10	2.684	198.090	190.090	196.215	184.805
		0.50	4.403	205.178	197.178	203.357	191.750
		0.90	10.879	206.035	198.035	204.191	192.619
$y = 0.5 + 3x_1 - 5x_2$ $x_1, x_2 \sim N$ *and correlated*	50	0.10	2.194	29.612	23.612	26.418	17.329
		0.50	4.871	31.116	25.116	27.897	18.606
		0.90	15.983	36.696	30.696	33.837	24.287
	100	0.10	2.366	53.137	47.137	51.452	42.925
		0.50	4.855	57.616	51.616	55.909	46.890
		0.90	12.389	71.558	65.558	69.925	60.783
	500	0.10	2.703	237.035	231.035	235.945	227.688
		0.50	4.415	272.237	266.237	271.152	262.667
		0.90	10.820	335.645	329.645	334.626	326.254

In Table 1, the two models are given to be compared in terms of AIC and Deviance values. For n=50, n=100 and n=500 and for each pairwise correlations are 0.10, 0.50 and 0.90, AIC and Deviance values are smaller in GAM than GLM. It can be easily concluded from the table that as the sample size increases up to 500, the difference in AIC and Deviance values are quite closer in GAM an GLM models.

5. Conclusions

The generalized linear model is widely used in many studies when the normality of response variable does not hold. The generalized additive model is also popularized in fitting because of some flexibility. On the other hand, estimation of linear or generalized linear models in the presence of multicollinearity is a common problem. Hence we consider a simulation study to examine which approach is proper when multicollinearity exists.

For this aim, a binary response variable with explanatory variables are fitted by generalized linear model and generalized additive logistic model in a conducted simulation study for moderate to larger sample sizes at different correlation levels. According to the results, the deviance values in GAM are smaller than GLM at each correlation level for sample sizes of 50, 100 and 500. However, the difference in deviance values between GAM and GLM is decreasing when the sample sizes are increased. Likewise, the AIC values are smaller in GAM than GLM in each model. As a result, we might conclude that the difference in deviance and AIC values between GAM and GLM, are not effected by the level of correlation. Finally, GAM fits better for small sample sizes in the presence of multicollinearity in terms of AIC and deviance values when the response variable is binary.

References

[1] Hastie, T. J. & Tibshirani, R. J. (1990). Generalized Additive Models. Chapman & Hall/CRC. ISBN 978-0-412-34390-2.

[2] Urgan, N. & Tez, M. (2008). Liu Estimator in Logistic Regression when the data are collinear, 20th International Euro Mini Conference on Continuous Optimization and Knowledge-Based Technologies, page 323-327.

[3] Godines-James, F. & Ramirez-Valverde, G. & Reyes-Carreto, R. & Ariza-Hernandez, F. & Barrera-Rodriguez, E. (2012). Collinearity and Seperated Data in the Logistic Regression Model, vol.46, is.4, page 411-425.

[4] Ariffin, S.B. & Midi, H. (2014). The Effect of High Leverage Points on the Logistic Ridge Regression Estimator Having Multicollinearity. Proceedings of the 3rd International Conference on Mathematical Sciences, vol.1602, page 1105-1111.

[5] Petrini, J. & Dias, R. & Pertile, S. & Eler, J. & Ferraz, J. & Mourao, G. (2012). Degree of Multicollinearity and Variables Involved in Linear Dependence in Additive-Dominant Models. Pesquisa Agropecuária Brasileira, vol.47, no.12.

[6] Ma, L. & Yan, X. (2014). Examining the Nonparametric Effect of Drivers’ Age in Rear-End Accidents Through an Additive Logistic Regression Model. Accident Analysis and Prevention, vol.67, page 129-136.

[7] Prince, S. & Aghajanian, J. (2009). Gender Classification in Uncontrolled Settings Using Additive Logistic Models. 2009 IEEE International Conference on Image Processing Proceedings Book, pages 2557-2560.

[8] Hastie, T. & Tibshirani, R. & Friedman, J. (2008). The Elements of Statistical Learning. Standford, California, Springer.