한국통계학회논문집 제2권 2호
1995년 12월 pp. 303-309

Multicollinearity in Logistic Regression [1)]

Jong-Han Lee [2)] and Myung-Hoe Huh [3)]

Abstract

Many measures to detect multicollinearity in linear regression have been proposed in statistics and numerical analysis literature. Among them, condition number and variance inflation factor(VIF) are most popular. In this study, we give new interpretations of condition number and VIF in linear regression, using geometry on the explanatory space. In the same line, we derive natural measures of condition number and VIF for logistic regression. These computer intensive measures can be easily extended to evaluate multicollinearity in generalized linear models.

KEY WORDS: linear regression, logistic regression, multicollinearity measures, condition number, variance inflation factor, random permutation.

1. Introduction

Multicollinearity is one of the major problems in linear regression. It inflates the variance of estimated regression parameters and makes the parameter estimates highly sensitive to small perturbations in the data. So multicollinearity has long been recognized as a potential source of problems in the estimation, testing and interpretation of linear regression parameters. Among measures to detect multicollinearity in the data, condition number KAPPA and variance inflation factor VIF are most popular(Belsley et al., 1991), although it is not easy to interpret condition number in statistical contexts(Stewart, 1987).

Recently, the multicollinearity in generalized linear model has received a growing concern of statisticians. Mackinnon and Puterman(1989) and Weissfeld and Sereika(1991) proposed condition numbers for generalized linear models in a frame of linear regression with individual weights. We will return to their measures in the last section. Also, Segerstedt and Nyquist(1992) suggested a geometrical approach to study the mechanism which determines the condition of data matrix in generalized linear model. In this study, we will deal with the same topic, but in a different way leading to naturally interpretable measures.

[1)] This research was partially supported by Institute of Statistics, Korea University.

[2)] Daewoo Research Institute, Yoido-dong 34-3, Seoul 150-010, KOREA.

[3)] Dept. of Statistics, Korea University, Anam-dong 5-1, Seoul 136-701, KOREA.

2. Multicollinearity Measures for Linear Regression

We will consider a standard linear regression model

$$y = \alpha + X\beta + \varepsilon \quad , \quad \varepsilon \sim N(0, \sigma^2 I_n),$$

where y is a size n vector of dependent observations, X is an $n \times p$ matrix of explanatory observations, and so on.

For the moment, we assume that X is centered for notational convenience. We may observe that the variance of estimated η ($\equiv \alpha + x^t\beta$) is

$$Var(\hat{\eta} | x) = \sigma^2 (1/n + x^t (X^t X)^{-1} x) \quad ,$$

which achieves its minimum at $x = 0$. Now, examine the variance trace $Var(\hat{\eta} | x_*)$ of the estimated linear predictor $\hat{\eta}$ when x_* travels along an ellipsoidal boundary

$$E : x_*^t \, D_{x'x}^{-1} \, x_* = c$$

where $D_{x'x} = diag(X^t X)$, sharing the same diagonal elements with $X^t X$, c is a constant. Observe that the ratio of the maximum to the minimum "variance gain", i.e.

$$\begin{aligned} \rho &= \frac{Max_{x_* \in E} \; Var(\hat{\eta} | x_*) - Var(\hat{\eta} | 0)}{Min_{x_* \in E} \; Var(\hat{\eta} | x_*) - Var(\hat{\eta} | 0)} \\ &= \frac{Max_{x_* \in E} \; x_*^t (X^t X)^{-1} x_*}{Min_{x_* \in E} \; x_*^t (X^t X)^{-1} x_*} \\ &= \frac{Max_{u^t u = c} \; u^t D_{x'x}^{1/2} (X^t X)^{-1} D_{x'x}^{1/2} u}{Min_{u^t u = c} \; u^t D_{x'x}^{1/2} (X^t X)^{-1} D_{x'x}^{1/2} u} \end{aligned}$$

is equal to the ratio of the maximum to the minimum eigenvalue of

$$D_{x'x}^{1/2} \, (X^t X)^{-1} \, D_{x'x}^{1/2} .$$

Thus, we may define a multicollinearity index KAPPA in linear regression by the condition number, ratio of the maximum to the minimum singular value, of the centered-scaled matrix

$$\tilde{X} \; (\equiv X D_{x'x}^{-1/2}) \; .$$

This leads to a relationship

$$KAPPA = \rho^{1/2} \quad ,$$

which assigns to KAPPA^2 a useful statistical meaning - the "unbalancedness" in variance gain of the estimated linear predictor $\hat{\eta}$ on a regulated path in the explanatory space.

On the other hand, variance inflation factor(VIF) can be formulated as the "relative variance" of the current estimate $\hat{\beta}_j$ compared to that in the ideal setting for the j-th explanatory variable under a certain restriction. Hereafter in this section, we will use changed notation for X: it is not pre-processed for centering. Define

$$VIF^j = \frac{Var(\hat{\beta}_j \, ; x^1, \cdots, x^j, \cdots, x^p)}{Min_{x^{*j} \in S^j} Var(\beta_j^* \, ; x^1, \cdots, x^{*j}, \cdots, x^p)} \quad , \quad j = 1, \cdots, p.$$

where x^j (*or* x^{*j}) is the j-th column of X(or X^*) and S^j is a sphere for the j-th explanatory variable. More specifically,

$$S^j : \| x^{*j} \| = \| x^j \| .$$

We can easily show that

$$VIF^j = 1/(1 - R_j^2) \quad , \quad j = 1, \cdots, p.$$

where R_j^2 is the coefficient of determination when x^j is regressed on $x^1, \cdots, x^{j-1}, x^{j+1}, \cdots, x^p$.

3. Generalizations: The Case of Logistic Regression

Consider the binary response

$$y_i \sim Bernoulli(p_i), \quad 0 < p_i < 1, \ i = 1, \cdots, n \ .$$

For the p_i part, set the logistic model with linear predictor :

$$p_i = \exp(\eta_i)/[1 + \exp(\eta_i)], \qquad \eta_i = \alpha + x_i^t \beta \quad .$$

Generalized linear model methodology tells us that(McCullagh and Nelder, 1989)

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = ((1:X)^t D_w (1:X))^{-1} (1:X)^t D_w z,$$

where z is the adjusted dependent observation vector and

$$D_w = diag(w_1, \cdots, w_n) \quad , \quad w_i = \hat{p}_i (1 - \hat{p}_i), \quad \hat{\eta}_i = \alpha + x_i^t \hat{\beta} \quad .$$

For the moment, we assume that the columns of X are centered with weights $w_1, \cdots, w_n$.

To define condition number, we use "variance gain" concept for $\hat{\eta}_i$ as follows: The ratio of the maximum to the minimum "variance gain" is given by

$$\rho = \frac{Max_{x_* \in E} \ Var(\hat{\eta} \mid x_*) - Var(\hat{\eta} \mid 0)}{Min_{x_* \in E} \ Var(\hat{\eta} \mid x_*) - Var(\hat{\eta} \mid 0)}$$

where

$$E : x_*^t D_{XD_wX}^{-1} x_* = c$$

and $D_{XD_wX} = diag(X^t D_w X)$. Using the fact

$$Var(\hat{\eta} \mid x_*) = 1/(\sum_{i=1}^n w_i) + x_*^t (X^t D_w X)^{-1} x_* \quad ,$$

we can show that

$$\rho = KAPPA^2 \ ,$$

where KAPPA is the condition number of $D_w^{1/2} X D_{XD_wX}^{-1/2}$, a specially weighted centered-scaled matrix of X.

Variance inflation factor can be similarly defined as in linear regression. With X for the notation of uncentered original matrix of explanatory observations, define

$$VIF^j = \frac{Var(\hat{\beta}_j ; x^1, \cdots, x^j, \cdots, x^p)}{Min_{x^{*j} \in S^j} Var(\beta_j^* ; x^1, \cdots, x^{*j}, \cdots, x^p)} \quad , \quad j = 1, \cdots, p.$$

But in this case, seemingly it is very hard to obtain VIF exactly. However, we may obtain its approximation by Monte-Carlo generation of random permutations of elements in the j-th predictor vector x^j. That is, we replace "minimization over S^j " in the denominator of VIF by "minimization over P^j_N ", where P^j_N is the set of N random permutations of n observed values for the j-th explanatory variable :

$$P^j_N : (x^*_{1j}, \cdots, x^*_{ij}, \cdots, x^*_{nj}) \quad \text{is a random permutation of} \quad (x_{1j}, \cdots, x_{ij}, \cdots, x_{nj}) \quad .$$

More specifically, approximate VIF's are obtained as follows: For the case $j = p$, without loss of generality,

STEP 1> For a given design matrix $(1:X)$ and response y, calculate MLE $\hat{\alpha}$ and $\hat{\beta}$ by the IWLS algorithm. Then its variance is approximately

$$Var\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = ((1:X)^t D_w (1:X))^{-1} \quad .$$

STEP 2> Generate a new design matrix

$$X^* \equiv (X^{(p)} : x^{*p}),$$

where x^{*p} is a random permutation of x^p. Suppose that the responses are to be generated from the logistic model initially fitted (with $\hat{\alpha}$ and $\hat{\beta}$ as model coefficients).

STEP 3> The new estimates α^* and β^* have approximate variance

$$Var\begin{pmatrix} \alpha^* \\ \beta^* \end{pmatrix} = ((1:X^*)^t D_{w^*} (1:X^*))^{-1}$$

where

$$D^*_w = diag(w^*_1, \cdots, w^*_n) \quad , \quad w^*_i = p^*_i \, (1 - p^*_i), \quad \eta^*_i = \hat{\alpha} + x^{*t}_i \hat{\beta} \quad ,$$

and x^{*t}_i is the i-th row of X^*.

STEP 4> Repeat STEP 2 and STEP 3 N times. Finally compute

$$VIF^p_N \equiv \frac{Var(\hat{\beta}_j ; x^1, \cdots, x^{p-1}, x^p)}{Min_{x^{*p} \in P^p_N} Var(\beta^*_j ; x^1, \cdots, x^{p-1}, x^{*p})} \quad , \quad j = 1, \cdots, p \quad ,$$

as an approximation of VIF^p.

In the next section, we will see that how such approximate VIF's are converging to some limits, as the number N of Monte-Carlo permutations increases.

4. Numerical Illustrations

As an illustration for the case of linear regression, we consider the Aerobic Fitness Data(SAS Institute, 1991), in which aerobic fitness is measured by the ability to consume oxygen. Consider the following linear regression model:

$$y = \alpha + \beta_1 x_1 + \cdots + \beta_6 x_6 + \varepsilon$$

where y is the dependent variable OXY, and $x_1, \cdots, x_6$ are explanatory variables AGE, WEIGHT, RUNTIME, RSTPULSE, RUNPULSE, MAXPULSE. The number of observations n is 31.

Table 1 lists VIF's and their approximations with the number N (=30, 100, 200, 1000) of Monte-Carlo permutations. We can see that approximate VIF's are getting closer to the exact VIF's as N increases. In this case, N = 100 seems working well.

Judging from condition number KAPPA=6.53, which can be obtained conveniently by SAS PROG REG with COLLINOINT option, the multicollinearity seems quite mild. However, we may note that VIF's for two variables RUNPULSE and MAXPULSE are of the magnitude 10, indicating the possibility that these two are not functioning independently in linear regression model.

Table 1. Variance Inflation Factor(VIF) and its approximations in linear regression of the Aerobic Fitness Data

Variables	VIF	Approximate VIF			
		N*=30	N=100	N=200	N=1000
AGE	1.5128	1.4768	1.4927	1.4927	1.5029
WEIGHT	1.1553	1.1506	1.1506	1.1506	1.1506
RUNTIME	1.5908	1.5142	1.5510	1.5654	1.5751
RESTPULSE	1.4155	1.3640	1.3868	1.4015	1.4015
RUNPULSE	8.4372	8.3968	8.3968	8.3968	8.4152
MAXPULSE	8.7438	8.3964	8.5914	8.6679	8.6685

* N is the number of Monte-Carlo generated random permutations for each predictor.

As an illustration for the case of logistic regression, we consider the Cancer Remission Data(Lee, 1974) which consists of patient characteristics and a clinical outcome whether cancer remission has occurred. The response variable is REMISS and explanatory variables are CELL, SMEAR, INFIL, LI, BLAST and TEMP.

Consider the logistic regression model

$$\log\ p/(1-p) = \alpha + \beta_1 x_1 + \cdots + \beta_6 x_6$$

where p is the probability of remission occurrence, and $x_1, \cdots, x_6$ are the explanatory variables from CELL to TEMP. The number of observations n is 27.

We calculate $\hat{\beta}$ by the iterative weighted least squares(IWLS) algorithm(McCullagh and Nelder, 1989), the diagonal matrix D_w of case weights $w_1,\cdots,w_n$, and, finally, the singular value decomposition of $D_w^{1/2} X D^{-1/2}_{X'D_wX}$, from which KAPPA = 48.70 is obtained. Judging from the condition number, we may suspect that the data in hand suffers from the multicollinearity.

In Table 2, approximate VIF's from N(=30, 100, 200, 1000) Monte-Carlo permutations are listed. We can see that the VIF's for CELL, SMEAR and INFIL are very large. So we may conclude that harmful effects from data ill-conditioning appear in regression coefficients corresponding to those three variables.

Table 2. Approximate VIF in logistic regression of the Cancer Remission Data

Variables	Approximate VIF			
	N=30	N=100	N=200	N=1000
CELL	16.876	21.142	21.142	22.290
SMEAR	46.591	47.992	50.633	54.943
INFIL	44.141	44.141	44.141	46.823
LI	1.908	1.971	2.388	2.388
BLAST	5.939	5.939	6.007	6.652
TEMP	1.865	2.065	2.139	2.162

* N is the number of Monte-Carlo generated random permutations for each predictor.

5. Concluding Remarks

In this study, we derived multicollinearity measures such as condition number and VIF for logistic regression that can be applied to generalized linear model of McCullagh and Nelder(1989) in a natural way. We would like to add two remarks.

In the same notations of Section 3 but with $n\times(p+1)$ matrix X that is not column-centered, Mackinnon and Puterman(1989) defined condition number for generalized linear models using extreme singular values of $D_w^{1/2} X$, while Weissfeld and Sereika's(1991) condition number was obtained from those of $D_w^{1/2} X D^{-1/2}_{X'D_wX}$, that would be the same matrix as ours if X is column-centered at respective weighted means and of the order $n\times$p, not containing *the column of 1's.* Both did not consider VIF. In contrast, we proposed a working definition of VIF.

Although our multicollinearity measures need some computations, it makes no serious problem even with personal computers. More details were studied by Lee(1994) in his doctoral dissertation at Statistics Department of the Korea University.

References

[1] Belsley, D.A., Kuh, E. and R.E. Welsch (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* John Wiley and Sons, New York.

[2] Lee, E.T. (1974). A Computer Program for Linear Logistic Regression Analysis, *Computer Programs in Biomedicine*, 80-92.

[3] Lee, J.-H. (1994). *Multicollinearity in Logistic Regression,* Unpublished Ph. D. Thesis. Department of Statistics, Korea University.

[4] Mackinnon, M.J. and Puterman, M.L. (1989). Collinearity in Generalized Linear Models, *Communications in Statistics-Theory and Methods*, 18(9), 3463-3472.

[5] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models,* 2nd Edition, Chapman and Hall, London.

[6] SAS Institute (1990). *SAS User's Guide: Statistics*, Version 6, 4th Edition. Cary, NC: SAS Institute, Inc.

[7] Segerstedt, B. and Nyquist, H. (1992). On the Conditioning Problem in Generalized Linear Models, *Journal of Applied Statistics*, Vol. 19, 513-526.

[8] Stewart, G.W. (1987). Collinearity and Least Squares Regression, *Statistical Science*, Vol. 2, 68-100.

[9] Weissfeld, L.A. and Sereika, S.M. (1991). A Multicollinearity Diagnostics For Generalized Linear Models, *Communications in Statistics-Theory and Methods,* 20(4), 1183-1198.