응용통계연구 제 7권 1호
1994년 3월 pp. 75 - 82

A Ridge-type Estimator For Generalized Linear Models

Byoung Jin Ahn[1)]

ABSTRACT

It is known that collinearity among the explanatory variables in generalized linear models inflates the variance of maximum likelihood estimators. A ridge-type estimator is presented using penalized likelihood. A method for choosing a shrinkage parameter is discussed and this method is based on a prediction-oriented criterion, which is Mallows's C_L statistic in a linear regression setting.

1. INTRODUCTION

Collinearity has long been recognized as a potential source of problem in the estimation, computation and interpretation of linear model parameters. As is the case in linear regression, model fitting via generalized linear models(GLMs) is also sensitive to collinearities among the explanatory variables in the model. Schaefer(1986) show that collinearity among explanatory variables in logistic regression inflates the variance of Maximum Likelihood Estimator(MLE). Mackinnon and Puterman(1989) investigate the relationship between collinearity in GLMs and standard linear models.

Schaefer et al.(1984) derive a ridge logistic estimator, and show that a ridge logistic estimator has smaller mean squard error than MLE under certain conditions. Marx and Smith(1990) present a principle component estimator for GLMs, and show that it can be useful with the presence of an ill-conditioned information matrix. It is the objective of this paper to develop and present a ridge-type estimator, as an option to traditional MLE for GLMs. Both iterative and one-step ridge estimators are developed using penalized likelihood.

There has been a substantial amount of interest in choosing a shrinkage parameter for ridge regression. This paper also concerns choosing the shrinkage parameter of ridge estimator for GLMs. In case of MLE, Efron(1986) discusses the general measures of prediction error which can be applied to GLMs. We adopt same idea to ridge estimator and obtain a prediction-oriented criterion, which can be used in the choice of shrinkage parameter for GLMs.

2. RIDGE ESTIMATOR

1) Department of Applied Statistics, Kon-Kuk University, Seoul, 133-701, Korea.

We suppose that the independent observations y_i are members of a one parameter exponential family with density functions

$$f(y_i;\theta_i,\phi) = \exp[(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i,\phi)], \quad i=1,2,...,n \tag{2.1}$$

where b(.) and c(.) are known functions and the value of $a(\phi)$ is, initially at least, assumed to be known. The mean and variance of Y_i are given by

$$E(Yi) = \dot{b}(\theta_i) = \mu_i,$$
$$Var(Y_i) = a(\phi)\ddot{b}(\theta_i) = v_i.$$

Here a dot denotes differentiation.

Assume that the natural parameters θ_i are expressed as linear combinations of known p-dimensional covariate vectors xi and an unknown p-dimensional parameter vector β,

$$\theta_i = x_i'\beta, \quad i=1,2,....,n.$$

The GLMs used in this paper are the models which have, so called, canonical links.(section 2.2.4 of McCullagh and Nelder, 1983).

Hoerl and Kennard(1970) derive the ridge estimator motivated by the tendency of least squares estimators to be too large. So, we may define the penalized likelihood as follows:

$$L_k(\beta) = \sum_i \log f(y_i;\theta_i,\phi) - kq(\beta), \tag{2.2}$$

where $q(\beta)=\beta'\beta/2$ is a penalty function and k is a nonnegative constant which controls the amount of penalty. Green(1987) examines penalized likelihood estimation in the context of general regression problems, characterized as probability models with composite likelihood functions. The Maximum Penalized Likelihood Estimator (MPLE) is the solution of $\dot{L}k(\beta) = 0$.

Applying the Newton-Rapson procedure with Fisher's scoring technique,a sequence of approximations, $\{\beta_{t+1}\}$, are generated according to

$$\begin{aligned}\tilde{\beta}_{t+1} &= \tilde{\beta}_t + (X'\ \tilde{W}_t X + kI)^{-1}[X'(y - \tilde{\mu}_t)/a(\phi) - k\,\tilde{\beta}_t] \\ &= (X'\ \tilde{W}_t X + kI)^{-1} X'\ \tilde{W}_t Z_t,\end{aligned} \tag{2.3}$$

where $\tilde{W}_t = diag\{\ \tilde{v}_i/a^2(\phi)\}_t$ and $Z_t = X\ \tilde{\beta}_t + \tilde{W}_t^{-1}(y - \tilde{\mu}_t)/a(\phi)$. In linear regression, MPLE is $\tilde{\beta} = (X'X + k^*I)^{-1}X'y$ without iteration, if we take $k = k^*/\phi$.

We can obtain one-step ridge estimator $\hat{\beta}^1$ of $\tilde{\beta}$ using the MLE $\hat{\beta}$ of β as an initial value

$$\tilde{\beta}^1=(X'\ \hat{W}X+kI)^{-1}(X'\ \hat{W}X)\hat{\beta}, \tag{2.4}$$

where $\hat{W}$ is the estimated W using $\hat{\beta}$. This estimator is appealing since it is easy to obtain in practice using output from existing generalized linear model packages. In case of logistic regression, $\tilde{\beta}^1$ in (2.4) is equivalent to ridge logistic estimator derived by Schaefer et al.(1984).

If we adopt $q(\beta)=\beta(X'WX)\beta/2$ as a penalty function, the MPLE of β is a Stein-type estimator

$$\tilde{\beta}_{t+1}=C(X'\ \hat{W}_tX)^{-1}X'\ \hat{W}_tZ_t,$$

where $C=(1+k)-1$ and $0<C<1$. (Stein, 1960)

3. A PREDICTION-ORIENTED CRITERION

Efron(1986) discusses the accuracy of the model for predicting future observation, when MLE is used. For the choice of Shrinkage parameter k, same idea can be adopted to ridge estimator.

The (scaled) deviance play a central role to assessment of goodness-of-fit, and it can be expressed as

$$D(y,\tilde{\mu})=2\sum_i \log f(y_i;y_i,\phi)/f(y_i;\mu_i,\phi), \tag{3.1}$$

where $\tilde{\mu}_i$ is the estimated mean at x_i using $\tilde{\beta}$ and $f(y_i;\mu_i,\phi)$ is the probobility function of y_i. If we denote the estimate of the natural parameter by $\tilde{\theta}=h(\tilde{\mu})$, the deviance can be written

$$D(y,\tilde{\mu})=2\sum_i[h(y_i)y_i-h(\tilde{\mu}_i)y_i-b(h(y_i))+b(h(\tilde{\mu}_i))]/a(\phi). \tag{3.2}$$

The deviance in (3.2) tends to underestimate the true prediction error because the data have been used twice, both to fit the model and to check its accuracy. The true deviance D* of a prediction vector $\tilde{\mu}$ is defined to be

$$\begin{aligned} D^* &\equiv E_f(D(y_f,\tilde{\mu})) \\ &=2\sum_i[E_f(h(y_f)y_f)-h(\tilde{\mu}_i)\mu_i-E_f(b(h(y_f)))+b(h(\tilde{\mu}_i))]/a(\phi). \end{aligned} \tag{3.3}$$

Here $y_f=(y_{f_1},y_{f_2},.....y_{f_n})$ is a hypothetical new data vector, with same distribution but independent of the original vector $y=(y_1,y_2,....,y_n)$, which gave $\tilde{\mu}$.

The expectation of the difference between D^* and $D(y,\tilde{\mu})$ is the downward bias of

$D(y,\tilde{\mu})$ as an estimate of the true prediction error. From (3.2) and (3.3), we obtain

$$\begin{aligned} E(D^*-D(y,\tilde{\mu})) &= 2E(\sum_i h(\mu_i)(y_i-\mu_i))/a(\phi) \\ &= 2E(\tilde{\beta}'X'(y-\mu))/a(\phi). \end{aligned} \tag{3.4}$$

The following approximation formula can be obtained from the equation (2.3) used for determining $\tilde{\beta}$.

$$\tilde{\beta} \cong \beta + (X'WX+kI)^{-1}(X'(y-\mu)/a(\phi)-k\beta). \tag{3.5}$$

The equation (3.5) provides an approximation of (3.4).

$$\begin{aligned} E(D^*-D(y,\tilde{\mu})) &\cong 2E((y-\mu)'X(X'WX+kI)^{-1}X'(y-\mu))/a^2(\phi) \\ &= 2tr(W^{1/2}X(X'WX+kI)^{-1}X'W^{1/2}). \end{aligned} \tag{3.6}$$

Thus, we have an estimator of D^*, namely,

$$\widehat{D^*} = D(y,\tilde{\mu}) + 2tr(H), \tag{3.7}$$

where $H = \widehat{W}^{1/2}X(X'\widehat{W}X+kI)^{-1}X'\widehat{W}^{1/2}$. Monitoring the changes of D^* in (3.7) as the values of k vary, we can choose the shrinkage parameter k which minimizes $\widehat{D^*}$.

If we take $k=0$, the D^* in (3.7) is given by

$$D^* = D(y,\tilde{\mu}) + 2p, \tag{3.8}$$

where $\tilde{\mu}$ is the estimated mean using MLE $\hat{\beta}$. Hence, $\widehat{D^*}$ is equivalent to Akaike's Information Criterion(AIC). (Akaike,1973).

In linear regression setting, $\widehat{D^*}$ is given by

$$\widehat{D^*} = \sum_i (y_i - x_i'\beta)^2/\sigma^2 + 2tr(H_L),$$

where $\tilde{\beta} = (X'X+k^*I)^{-1}X'y$, $H_L = X(X'X+k^*I)^{-1}X'$, and $Var(Y_i)=\sigma^2$. If we replace σ^2 with $\widehat{\sigma^2} = y'(I-X(X'X)-1X')y/(n-p)$, $D^*$is equivalent to Mallows's C_L statistic (1973).

4. A SIMPLE EXAMPLE

As a simple example we fit logistic regression model to the artificial data displayed in Table 1. The notation $N(0,1)$ indicates a random variable following the standard normal distribution and $U(0,1)$ indicates a random variable following the uniform distribution on the interval [0,1]. In the table $X_1 \sim N(0,1)$ and X_2 was generated from X_1 as follows : $X_2 = X_1 + 0.7 \times N(0,1)$. Thus, X_1 and X_2 are highly correlated and correlation coefficient is 0.82. The outcome variable was generated by comparing a $U(0,1)$ variate, U, to the true probablity $\pi = [1+\exp(-1-X_1-X_2)]^{-1}$ as follows : if $U<\pi$ then Y=1, otherwise Y=0.

Table 1. Data displaying near collinearity between the explanatory variables.

Y	X_1	X_2
0	-1.07	-1.66
0	0.57	-0.36
1	-0.11	-0.17
1	0.31	1.31
1	1.09	2.03
1	0.61	0.62
0	0.08	-0.60
0	-0.77	-1.71
1	0.35	-0.18
1	-0.20	0.16
1	1.92	1.60
1	1.32	-1.34
1	-0.47	0.21
1	0.03	-0.56
1	-0.59	0.16
1	-0.36	0.32
1	2.24	2.55
0	-0.16	-1.03
1	0.99	0.09
0	-0.43	-0.37

The results of fitting logistic regression to various values of shrinkage parameter k are presented in Table 2. The estimated coefficients are one-step estimates given in

(2.4). The values of tr(H) decrease monotonically as k increases. However, the values of $D(y,\tilde{\mu})$ increse as k increases. Hence, as a result, the choice of k using $\widehat{D^*}$ is a trade-off problem and k=0.03 might be a proper choice.

Table 2. Estimated coefficients and the values of $\widehat{D^*}$ for various values of shrinkage parameter k.

k	$\tilde{\beta}_o$	$\tilde{\beta}_1$	$\tilde{\beta}2$	$D(y,\tilde{\mu})$	tr(H)	$\widehat{D^*}$
0.00	2.587	-2.500	4.599	11.884	3.00	17.884
0.01	2.353	-2.208	4.184	11.916	2.891	17.698
0.02	2.160	-1.986	3.842	11.996	2.800	17.596
0.03	1.998	-1.767	3.555	12.112	2.721	17.554
0.04	1.860	-1.597	3.310	12.250	2.653	17.557
0.05	1.741	-1.451	3.099	12.406	2.593	17.591
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
0.1	1.327	-0.954	2.368	13.329	2.369	18.023
0.2	0.914	-0.484	1.645	14.960	2.103	19.167
0.3	0.705	-0.267	1.282	16.286	1.933	20.152

5. CONCLUDING REMARKS

It is known that collinearity among the explanatory variables in GLMs seriously effects the MLE in that the variance of this estimator is inflated in much the same way that collinearity inflates the variance of the least aquares estimator in multiple regression. A ridge-type estimator and a method for chooing a shrinkage parameter are discussed. This method is based on a prediction oriented criterion, which is Mallows's CL satistic in a linear regression setting.

It seems necessary that some Monte Carlo study be done to compare the performance of the ridge estimator relative to MLE, and another method for choosing a shrinkage parameter also deserves further study.

REFERENCES

[1] Akaike,H.(1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory,* Akademiai Kiado, Budapest, 267-281.

[2] Efron,B.(1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association,* Vol. 81, 461-470.

[3] Green,P.J.(1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review,* Vol. 55, 245-259.

[4] Hoerl,A.E., and kennard,R.W.(1970). Rigde regression. : Biased estimation for non-orthogonal problems. *Technometrics,* Vol. 12, 55-67.

[5] Lee, A. H. and Silvapulle, M. J. (1988). Ridge estimation in logistic regression. *Communications in Statistics, Simulation,* Vol. 17, 1231-1257.

[6] Mackinnon,M.J. and Puterman,M.L.(1989). Collinearity in generalized linear models. *Communications in Statistics, Theory and Methods,* Vol. 18, 3463-3472.

[7] Mallows,C.W.(1973). Some comments on Cp. *Technometrics,* Vol. 15, 661-675.

[8] Marx,B.D., and Smith,E.F.(1990). Principal component estimation for generalized linear regression. *Biometrika,* Vol. 77, 23-31.

[9] MacCullagh,P., and Nelder,J.A.(1983). *Generalized Linear Models.* New York : Chapman and Hall.

[10] Schaefer,R.L.(1986). Alternative estimators in logistic resression when data are collinear. *Journal of Statistics, Computation and simulation,* Vol. 25, 75-91.

[11] Schaefer,R.L., Roi,L.D., and Wolfe,R.A.(1984). A ridge logistic estimator. *Communications in statistics, Theory and Methods,* Vol. 13, 99-113.

[12] Stein,C.M.(1960). *Multiple regression, contributions to probability and Statistics,* Stanford university press.

일반화 선형모형에서의 능형형태의 추정량

안병진[2)]

요약

일반화 선형모형에서도 회귀모형에서와 마찬가지로 다공선성이 존재할 경우 여러가지 문제가 발생한다. 이를 극복하기 위한 한가지 방법으로 능형형태의 추정량과 그의 축소 모수를 결정하는 방법에 대하여 다루었다.

2) (133-701) 서울특별시 성동구 모진동 93-1, 건국대학교 응용통계학과