A Study on Local Influence in Ridge Regression

Sungho Moon
Pusan University of Foreign Studies, Department of Statistics
55-1, Uam-Dong, Nam-Gu, Pusan 608-738, Korea
shmoon@taejo.pufs.ac.kr

Jaekyoung Shin
Changwon National University, Department of Statistics
9, Sarim-Dong, Changwon 641-773, Korea
jkshin@sarim.changwon.ac.kr

Hyunjeong Kim
Silla University, Department of General Education
1-1, Gwaebop-Dong, Sasang-Gu, Pusan 617-736, Korea
semikim@silla.ac.kr

Yutaka Tanaka
Okayama University, Department of Environmental and Mathematical Sciences
3-1-1, Tsushima Naka, Okayama 700-8530, Japan
tanaka@stat.ems.okayama-u.ac.jp

1. Introduction

We consider ordinary regression model, which is expressed as $\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}\beta_1 + \epsilon$, where $\mathbf{y}$ is an $(n \times 1)$ vector of dependent variable, $\mathbf{1}$ is an $(n \times 1)$ vector whose elements are all 1's, $\mathbf{X}$ is an $(n \times p)$ matrix of standardized $(\sum_i^n \mathbf{X}_{ij} = 0, \quad \sum_i^n \mathbf{X}_{ij}^2 = 1, \quad j = 1, \ldots, p)$ independent variables, and ϵ is an $(n \times 1)$ vector of error terms. The ridge regression estimator for the regression coefficient vector can be expressed as $\hat{\beta}_R = \left[\bar{y},\ \mathbf{y}^T\mathbf{X}V(n\Lambda + k\mathbf{I})^{-1}V^T\right]$, where Λ is the diagonal matrix of the eigenvalues of the covariance matrix $\mathbf{X}^T\mathbf{X}$ and V is the matrix of the associated eigenvectors. Shi and Wang(1999) considered the influence on the regression coefficients under the condition that ridge parameter k is fixed and on the influence on k, separately. But in this paper, we consider the influence on regression coefficients in two cases: One is the case where the ridge parameter is fixed, and the other unfixed. In the latter case the indirect effect through the change of the ridge parameter as well as the direct effect are taken into account on the regression coefficients.

2. Measuring Influence in RR

By using the result of Tanaka(1989) the influence function for the regression coefficient vector is derived as

$$\begin{aligned}\hat{\beta}_R^{(1)} &= \Bigg[(y_i-\bar{y}),\ -S_{yx}\Bigg\{\sum_s\sum_r(n\lambda_s+k)^{-1}(n\lambda_r+k)^{-1}(\mathbf{v}_s^TA^{(1)}\mathbf{v}_r)(\mathbf{v}_r\mathbf{v}_s^T+\mathbf{v}_s\mathbf{v}_r^T)\\ &\quad+\frac{\partial k}{\partial\omega_i}\left(\sum_s(n\lambda_s+k)^{-2}\mathbf{v}_s\mathbf{v}_s^T\right)\Bigg\}+S_{iyx}^{(1)}V(n\Lambda+k\mathbf{I})^{-1}V^T\Bigg],\end{aligned}$$

where $S_{yx}=n^{-1}\sum_i(\mathbf{x}_i-\bar{\mathbf{x}})(y_i-\bar{y})$, $A^{(1)}=S_D^{-1/2}S^{(1)}S_D^{-1/2}-2^{-1}S_D^{-1}S_D^{(1)}R-2^{-1}RS_D^{-1}S_D^{(1)}$, R is a correlation matrix of X and the superscript (1) means the influence function.

3. Numerical Example

Walker and Birch(1988) and Shi and Wang(1999) applied ridge regression analysis and their influence analyses to the Longley data(Longley, 1967). The data set consists of 6 explanatory variables and 16 observations. Walker and Birch(1988) used approximate deletion formulas and cases 16, 10, 4, 15 and 1 were found to be influential in ridge regression($k=0.0002$) in order of the amounts of influence, while Shi and Wang(1999)'s result modified those influential observations on the ridge estimator in order as 10, 4, 15, 16 and 1($k=0.0002$). In our result cases 5, 16, 3, 15 and 4 are found influential in this order.

REFERENCES

Longley, J.W. (1967). An appraisal of least squares programs for electronic computer from the point of view of the user. *J. Amer. Statist. Assoc.*, **62**, 819-841.

Shi, L. and Wang, X. (1999). Local influence in ridge regression. *Comp. Statist. & Data Analysis*, **31**, 341-353.

Tanaka, Y. (1989). Influence functions related to eigenvalue problems which appear in multivariate methods. *Comm. Statist.*, **A 18**, 3991-4010.

Walker, E. and Birch, J.B. (1988). Influence measures in ridge regression. *Technometrics*, **30**, 221-227.

RESUME

Influence functions are derived for the regression coefficients in ridge regression. In doing this we consider two cases: One is the case where the ridge parameter is fixed, and the other unfixed. In the latter case indirect effect through the change of the ridge parameter as well as direct effect are taken into account on the regression coefficients. A numerical example is shown for illustration and a comparison is made with the influence analysis based on Shi and Wang(1999)'s method.