Journal of the Korean Statistical Society 44 (2015) 271–279

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Case influence diagnostics in the lasso regression

Choongrak Kim [a,*], Jungsu Lee [a], Hojin Yang [a], Whasoo Bae [b]

[a] Department of Statistics, Pusan National University, Pusan, 609-735, Republic of Korea
[b] Department of Data Science, Inje University, Gimhae, 621-749, Republic of Korea

ARTICLE INFO

Article history:
Received 14 February 2014
Accepted 27 September 2014
Available online 6 November 2014

AMS 2000 subject classifications:
62J20
62J07

Keywords:
Cook's distance
Generalized cross-validation
Influential observations
Shrinkage
Subset selection

ABSTRACT

Using the diagnostic results in the ridge regression model, we propose an approximate version of Cook's distance in the lasso regression model since the analytic expression of the lasso estimator is not available. Also, we express the proposed Cook's distance in terms of basic building blocks such as residuals and leverages. We verify that the proposed statistic successfully detects potentially influential observations on estimators of regression coefficients and on the model selection in the lasso regression model. An illustrative example based on a real dataset is given.

1. Introduction

Studies on regression diagnostics continued for more than 30 years since Cook (1977), and they focused on most of the existing statistical models with few exceptions. Good references in the linear regression diagnostics are Cook and Weisberg (1980), Belsley, Kuh, and Welsch (1980), and Chatterjee and Hadi (1986) among others. Also, Cook and Wang (1983), Hinkley and Wang (1988), Tsai and Wu (1990), and Kim, Storer, and Jeong (1996) developed diagnostic methods in the Box–Cox transformation model (Box & Cox, 1964). Walker and Birch (1990) derived influence measures in the ridge regression (Hoerl & Kennard, 1970). In the spline smoothing model, Eubank (1985), Silverman (1986), and Kim (1996) suggested a version of Cook's distance. Kim, Lee, and Park (2001) defined Cook's distance in the local polynomial regression, and Fung, Zhu, Wei, and He (2002) and Kim, Park, and Kim (2002) studied influence diagnostics in the semiparametric model. Recently, Bae, Hwang, and Kim (2008) developed diagnostic issues in the varying coefficient model.

In this paper, we study diagnostic issues in the lasso regression (Tibshirani, 1996). Most influence measures, suggested so far in many statistical models, are concerned about detecting influential observations on estimators of regression coefficients. In the lasso regression, the first interest is, of course, detecting influential observations on estimators of regression coefficients, and the second interest is detecting influential observations on estimate of shrinkage parameter. Especially in the lasso regression, one or few influential observations on estimates of regression coefficients can also be influential on the estimator of shrinkage parameter, so that model selection results based on the lasso will be different due to one or few observations. We use the deletion method to define a type of Cook's distance in the lasso regression. Analytic expression is not possible in the lasso regression. To overcome this difficulty in the lasso regression, we use the diagnostic result of

* Corresponding author.
E-mail address: crkim@pusan.ac.kr (C. Kim).

http://dx.doi.org/10.1016/j.jkss.2014.09.003
1226-3192/

ridge regression and adapt it to the lasso regression. This paper is organized as follows. In Section 2, relevant notations and results in the linear and ridge regression diagnostics are introduced. An approximate version of Cook's distance in the lasso regression is derived in Section 3. Numerical studies on the proposed statistic are done, and an illustrative example based on a real data is given in Section 4, and concluding remarks are given in Section 5.

2. Linear and ridge regression diagnostics

Consider the linear regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y}$ is an n-vector of responses, $\mathbf{X}$ is an $n \times p$ full column matrix of known covariates, $\boldsymbol{\beta}$ is a p-vector of unknown coefficients, and $\boldsymbol{\epsilon}$ is an n-vector of independent random variables with mean zero and unknown variance σ^2. We use y_i and $\boldsymbol{x}_i$ to denote the ith row of $\boldsymbol{y}$ and $\boldsymbol{X}$, respectively, and use the subscript (i) to indicate the deletion of the ith observation. Thus, $\boldsymbol{X}_{(i)}$ denotes the matrix $\boldsymbol{X}$ with the ith row deleted. After fitting the model by the method of least squares, we have $\hat{\boldsymbol{\beta}}^{lse} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, and $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$, where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\mathbf{X}'$ is the hat matrix. Let the residual vector be $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ and $s^2 = \boldsymbol{e}'\boldsymbol{e}/(n-p)$ be the unbiased estimator of σ^2. Let $\hat{\boldsymbol{\beta}}^{lse}_{(i)}$ be the least squares estimator of $\boldsymbol{\beta}$ calculated with the ith case deleted. Miller (1974) showed that

$$\hat{\boldsymbol{\beta}}^{lse} - \hat{\boldsymbol{\beta}}^{lse}_{(i)} = \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i e_i}{1 - h_{ii}}, \tag{1}$$

where h_{ij} is the ij-th element of the hat matrix $\boldsymbol{H}$, i.e., $h_{ij} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_j$. Cook's distance of the ith observation is defined as

$$C_i^{lse} = \frac{1}{p}(\hat{\boldsymbol{\beta}}^{lse} - \hat{\boldsymbol{\beta}}^{lse}_{(i)})' Cov(\hat{\boldsymbol{\beta}}^{lse})^{-1}(\hat{\boldsymbol{\beta}}^{lse} - \hat{\boldsymbol{\beta}}^{lse}_{(i)}).$$

Using the result of (1), it can be expressed as basic building blocks, i.e.,

$$C_i^{lse} = \frac{1}{p\sigma^2}\frac{{e_i}^2 h_{ii}}{(1-h_{ii})^2}. \tag{2}$$

For the influence of k observations, let $K = \{i_1, \ldots, i_k\}$ be an index set of size k. Also, let $\boldsymbol{X}_K$ be the $k \times p$ submatrix of $\boldsymbol{X}$ corresponding to the rows of the cases in K, and let $\boldsymbol{e}_K$ be the k-vector with k elements of $\boldsymbol{e}$ in K. Let $\hat{\boldsymbol{\beta}}_{(K)}$ be the estimate of $\boldsymbol{\beta}$ based on $n-k$ observations after deleting observations in a set K. Then, Cook's distance for k observations is defined as

$$C_K^{lse} = \frac{1}{p}(\hat{\boldsymbol{\beta}}^{lse} - \hat{\boldsymbol{\beta}}^{lse}_{(K)})' Cov(\hat{\boldsymbol{\beta}}^{lse})^{-1}(\hat{\boldsymbol{\beta}}^{lse} - \hat{\boldsymbol{\beta}}^{lse}_{(K)}).$$

If we let $\boldsymbol{H}_K = \boldsymbol{X}_K(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_K'$, then

$$\hat{\boldsymbol{\beta}}^{lse} - \hat{\boldsymbol{\beta}}^{lse}_{(K)} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_K'(\boldsymbol{I} - \boldsymbol{H}_K)^{-1}\boldsymbol{e}_K,$$

and, therefore,

$$C_K^{lse} = \frac{1}{p\sigma^2}\boldsymbol{e}_K'(\boldsymbol{I} - \boldsymbol{H}_K)^{-1}\boldsymbol{H}_K(\boldsymbol{I} - \boldsymbol{H}_K)^{-1}\boldsymbol{e}_K.$$

When σ^2 is unknown, it is often replaced by its unbiased estimator s^2.

Ridge regression was first introduced by Hoerl and Kennard (1970) as a way of dealing with multicollinearity in the linear regression. The ridge estimate of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_\theta^{ridge} = (\boldsymbol{X}'\boldsymbol{X} + \theta\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

where θ, called a ridge regression parameter, is a positive value to be estimated. To estimate θ, the GCV (generalized cross-validation) criterion defined as

$$\mathrm{GCV}_\theta = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{\theta,i})^2}{\{1 - tr(\boldsymbol{H}_\theta)\}^2},$$

is often used. Here

$$\boldsymbol{H}_\theta = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \theta\boldsymbol{I})^{-1}\boldsymbol{X}'$$

is, so-called, the ridge hat matrix and $\hat{y}_{\theta,i} = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_\theta^{ridge}$ is the ith fitted value. Cook's distance of the ith observation in ridge regression can be defined in the same form as in the linear regression model, i.e.,

$$C_i^{ridge} = \frac{1}{p}(\hat{\boldsymbol{\beta}}_\theta^{ridge} - \hat{\boldsymbol{\beta}}_{\theta(i)}^{ridge})' Cov(\hat{\boldsymbol{\beta}}_\theta^{ridge})^{-1}(\hat{\boldsymbol{\beta}}_\theta^{ridge} - \hat{\boldsymbol{\beta}}_{\theta(i)}^{ridge}),$$