Communications for Statistical Applications and Methods
2017, Vol. 24, No. 2, 155–162
https://doi.org/10.5351/CSAM.2017.24.2.155
Print ISSN 2287-7843 / Online ISSN 2383-4757

Case influence diagnostics for the significance of the linear regression model

Whasoo Bae[a], Soyoung Noh[b], Choongrak Kim[1,b]

[a]Department of Statistics, Inje University, Korea;
[b]Department of Statistics, Pusan National University, Korea

Abstract

In this paper we propose influence measures for two basic goodness-of-fit statistics, the coefficient of determination R^2 and test statistic F in the linear regression model using the deletion method. Some useful lemmas are provided. We also express the influence measures in terms of basic building blocks such as residual, leverage, and deviation that showed them as increasing function of residuals and a decreasing function of deviation. Further, the proposed measure reduces computational burden from $O(n)$ to $O(1)$. As illustrative examples, we applied the proposed measures to the stackloss data sets. We verified that deletion of one or few influential observations may result in big change in R^2 and F-statistic.

Keywords: coefficient of determination, F-statistic, goodness-of-fit, influential observations

1. Introduction

Research on regression diagnostics were initiated by Cook (1977) that have continued to most existing statistical models with a few exceptions. Good reviews on representative results in linear regression diagnostics are Belsley *et al.* (1980), Chatterjee and Hadi (1986), and Cook and Weisberg (1982). Diagnostic results in Box-Cox transformation model (Box and Cox, 1964) have also been done by Cook and Wang (1983), Hinkley and Wang (1988), Kim *et al.* (1996), and Tsai and Wu (1990). Walker and Birch (1988) derived a version of Cook's distance in the ridge regression (Hoerl and Kennard, 1970), and Kim *et al.* (2015) proposed a type of Cook's distance in the lasso regression model.

Studies on regression diagnostics have also been done in regards to nonparametric regression. In the spline smoothing model, Eubank (1985), Kim (1996), and Silverman (1985) suggested versions of Cook's distance, and Kim *et al.* (2001) suggested Cook's distance in the local polynomial regression, and Fung *et al.* (2002) and Kim *et al.* (2002) studied detection of influential observations in the semiparametric model. Bae *et al.* (2008) also studied diagnostic issues in the varying coefficient model.

In this paper, we study diagnostic issues in goodness-of-fit measures in the linear regression model. We develop influence measures for goodness-of-fit statistics such as the F-statistic and the coefficient of determination R^2. Many of the suggested influence measures in many statistical models are concerned with detecting influential observations on estimators of regression coefficients. As far as we know, influence measures for goodness-of-fit statistics have not yet been studied. By using the deletion

[1] Corresponding author: Department of Statistics, Pusan National University, 2, Busandaehak-ro 63 beon-gil, Geumjeong-gu, Busan 46241, Korea. E-mail: crkim@pusan.ac.kr

Published 31 March 2017 / journal homepage: http://csam.or.kr

method, we also express the proposed influence measures for goodness-of-fit statistics in terms of basic building blocks such as residuals and deviations. This paper is organized as follows. In Section 2, influence measures for goodness-of-fit statistics are suggested and analytic expression for goodness-of-fit statistics in terms of basic building blocks are given. Illustrative examples based on real data sets are given in Section 3, and concluding remarks are given in Section 4.

2. Influence measures for goodness-of-fit statistics

2.1. Basic notations

Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y}$ is an n-vector of responses, $\boldsymbol{X}$ 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 Gaussian random variables with mean zero and unknown variance σ^2. We use y_i and $\boldsymbol{x}_i$ to denote the i^{th} row of $\boldsymbol{y}$ and $\boldsymbol{X}$, respectively, and use the subscript (K) to indicate the deletion of k observations in an index set $K = \{i_1, i_2, \ldots, i_k\}$. Therefore, $\boldsymbol{X}_{(i)}$ denotes the matrix $\boldsymbol{X}$ with the i^{th} row deleted. However, the subscript K indicates the corresponding k observations in an index set K. Therefore, $\boldsymbol{X}_K$ denotes the $k \times p$ submatrix of $\boldsymbol{X}$. After fitting the model by the method of least squares, we have $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, 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 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.

Here, we consider the influence of k observations on the coefficient of determination R^2 and the F-test statistic. Recall that

$$R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}},$$

where $\text{SST} = \sum(y_i - \bar{y})^2$ is the total sum of squares (SST), and $\text{SSE} = \sum(y_i - \hat{y}_i)^2$ is the error sum of squares (SSE). Also, the F-test statistic is given by

$$F = \frac{\text{SST} - \text{SSE}}{(p-1)s^2}.$$

Therefore, both statistics, R^2 and F, are functions of SST, SSE, and s^2.

2.2. Useful Lemmas

Let $R^2_{(K)}$ and $F_{(K)}$ be the coefficient of determination and the F-statistic, respectively, based on $(n-k)$ observations after deleting k observations in a set K. Now, to detect influential observations on R^2 and F, we suggest influence measures as: Define $\Delta D_K = R^2 - R^2_{(K)}$ and $\Delta F_K = F - F_{(K)}$ as influence measures for the coefficient of determination and the F-statistic, respectively. Hence, they can be rewritten as

$$\Delta D_K = \frac{\text{SSE}_{(K)}}{\text{SST}_{(K)}} - \frac{\text{SSE}}{\text{SST}}, \tag{2.1}$$

and

$$\Delta F_K = \frac{(n-p)\,(\text{SST} - \text{SSE})}{(p-1)\text{SSE}} - \frac{(n-p-k)\left(\text{SST}_{(K)} - \text{SSE}_{(K)}\right)}{(p-1)\text{SSE}_{(K)}}, \tag{2.2}$$

where $\mathrm{SST}_{(K)}$ and $\mathrm{SSE}_{(K)}$ denote the SST and the SSE, respectively, based on $(n-k)$ observations after deleting k observations in a set K.

Lemma 1. *$SST_{(K)} = SST - 1/(n-k)(\sum_{j\in K} d_j)^2 - \sum_{j\in K} d_j^2$, where $d_j = y_j - \bar{y}$.*

Proof: Let $\bar{y}_K = (1/k)\sum_{j\in K} y_j$ and $\bar{y}_{(K)} = (1/(n-k))\sum_{j\notin K} y_j$. Now,

$$\begin{aligned}\mathrm{SST}_{(K)} &= \sum_{j\notin K}\left(y_j - \bar{y}_{(K)}\right)^2 \\ &= \sum_{j=1}^{n}\left(y_j - \bar{y}_{(K)}\right)^2 - \sum_{j\in K}\left(y_j - \bar{y}_{(K)}\right)^2.\end{aligned}$$

Since $\sum_{j=1}^{n}(y_j - \bar{y}) = 0$ and $\bar{y} - \bar{y}_{(K)} = (1/(n-k))\sum_{j\in K} d_j$, we have

$$\begin{aligned}\mathrm{SST}_{(K)} &= \mathrm{SST} + \frac{n}{(n-k)^2}\left(\sum_{j\in K} d_j\right)^2 - \sum_{j\in K} d_j^2 - \frac{2}{n-k}\left(\sum_{j\in K} d_j\right)^2 - \frac{k}{(n-k)^2}\left(\sum_{j\in K} d_j\right)^2 \\ &= \mathrm{SST} - \frac{1}{n-k}\left(\sum_{j\in K} d_j\right)^2 - \sum_{j\in K} d_j^2.\end{aligned}$$

□

Lemma 2. *$SSE_{(K)} = s^2(n-p-r_K^2)$, where $r_K^2 = \boldsymbol{e}_K{}'(\boldsymbol{I}-\boldsymbol{H}_K)^{-1}\boldsymbol{e}_K/s^2$, $\boldsymbol{e}_K = (e_{i_1},\ldots,e_{i_k})'$, and $\boldsymbol{H}_K = \boldsymbol{X}_K(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_K$.*

Proof: First, we note that

$$\begin{aligned}\mathrm{SSE}_{(K)} &= \sum_{j\notin K}\left(y_j - \mathbf{x}_j'\hat{\boldsymbol{\beta}}_{(K)}\right)^2 \\ &= \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{(K)}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{(K)}\right) - \left(\boldsymbol{y}_K - \boldsymbol{X}_K\hat{\boldsymbol{\beta}}_{(K)}\right)'\left(\boldsymbol{y}_K - \boldsymbol{X}_K\hat{\boldsymbol{\beta}}_{(K)}\right).\end{aligned}$$

Also, recall that

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

therefore, we have

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

and

$$\boldsymbol{y}_K - \boldsymbol{X}_K\hat{\boldsymbol{\beta}}_{(K)} = \boldsymbol{e}_K + \boldsymbol{H}_K(\boldsymbol{I} - \boldsymbol{H}_K)^{-1}\boldsymbol{e}_K.$$

Finally, we have

$$\mathrm{SSE}_{(K)} = \mathrm{SSE} + \boldsymbol{e}_K'\left\{-\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{H}_K)^{-1}\boldsymbol{H}_K\right\}\boldsymbol{e}_K,$$

and note that

$$\boldsymbol{I} + (\boldsymbol{I} - \boldsymbol{H}_K)^{-1}\boldsymbol{H}_K = (\boldsymbol{I} - \boldsymbol{H}_K)^{-1}.$$

Then,

$$\begin{aligned}\mathrm{SSE}_{(K)} &= \mathrm{SSE}\left(1 - \frac{r_K^2}{n-p}\right) \\ &= s^2\left(n - p - r_K^2\right)\end{aligned}$$

which completes the proof of Lemma 2. □

2.3. Influence measures

By Lemmas 1 and 2, we have

$$\begin{aligned}\Delta D_K &= R^2 - R_{(K)}^2 \\ &= \frac{\mathrm{SSE}\left(1 - \frac{r_K^2}{n-p}\right)}{\mathrm{SST} - \frac{1}{n-k}\left(\sum_{j\in K} d_j\right)^2 - \sum_{j\in K} d_j^2} - \frac{\mathrm{SSE}}{\mathrm{SST}},\end{aligned}$$

and

$$\begin{aligned}\Delta F_K &= F - F_{(K)} \\ &= \frac{\frac{\mathrm{SST}-\mathrm{SSE}}{p-1}}{\frac{\mathrm{SSE}}{n-p}} - \frac{(n-p-k)\left(\mathrm{SST} - \frac{1}{n-k}\left(\sum_{j\in K} d_j\right)^2 - \sum_{j\in K} d_j^2 - \mathrm{SSE}\left(1 - \frac{r_K^2}{n-p}\right)\right)}{(p-1)\left(\mathrm{SSE}\left(1 - \frac{r_K^2}{n-p}\right)\right)}.\end{aligned}$$

As a special case, we consider deleting a single observation, the i^{th} observation, say. Then, we can easily express them as: For notational simplicity, let

$$q_i = \frac{d_i^2}{\mathrm{SST}}$$

be the proportion of the i^{th} deviation, and $r_i = e_i/(s\sqrt{1-h_{ii}})$ be the i^{th} internally standardized residual, and $r_i^* = e_i/(s_{(i)}\sqrt{1-h_{ii}})$ be the i^{th} externally standardized residual. Then, we have

$$\begin{aligned}\Delta D_i &= \frac{\mathrm{SSE}_{(i)}}{\mathrm{SST}_{(i)}} - \frac{\mathrm{SSE}}{\mathrm{SST}} \\ &= \left(1 - R^2\right)\cdot\frac{\left(\frac{n}{n-1}q_i - \frac{r_i^2}{n-p}\right)}{1 - \frac{n}{n-1}q_i}.\end{aligned}$$

Also, for the influence of the i^{th} observation on F, we have

$$\begin{aligned}\Delta F_i &\equiv F - F_{(i)} \\ &= F\cdot\left\{1 - \frac{1 - \left(\frac{n}{n-1}d_i^2 - r_i^2 s^2\right)/\mathrm{SSR}}{r_i^2/r_i^{*2}}\right\}.\end{aligned}$$

Table 1: Values of ΔD_i, ΔF_i with the standardized residual r_i, the leverage h_{ii}, and the proportion of deviation q_i in the stackloss data

ΔD_i		ΔF_i		r_i		h_{ii}		q_i	
Obs.	Value	Obs.	Value	Obs.	Value	Obs.	Value	Obs.	Value
21	−0.035	21	−38.9	21	−2.638	17	0.412	1	0.290
1	0.027	1	18.4	4	1.882	2	0.318	2	0.183
2	0.017	2	13.9	3	1.546	1	0.302	3	0.183
4	−0.014	4	−8.4	1	1.193	21	0.285	16	0.054
3	0.006	3	7.2	9	−1.046	7	0.219	4	0.053
6	−0.005	16	6.7	12	0.969	8	0.219	15	0.044
9	−0.005	18	6.3	6	−0.965	12	0.217	17	0.044
16	0.005	19	5.7	11	0.884	14	0.206	18	0.044
12	−0.004	17	5.0	7	−0.834	10	0.200	19	0.035
18	0.004	14	4.5	15	0.809	15	0.190	13	0.021
7	−0.003	13	4.0	2	−0.716	3	0.175	14	0.015
11	−0.003	15	4.0	17	−0.611	19	0.175	12	0.010
19	0.003	10	3.2	5	−0.542	18	0.161	10	0.006
17	0.002	20	3.0	8	−0.485	13	0.158	11	0.006
5	−0.001	8	2.9	13	−0.480	11	0.155	8	0.003
8	−0.001	5	2.4	20	0.454	9	0.140	9	0.003
13	0.001	7	1.0	10	0.437	16	0.131	20	0.003
14	0.001	11	1.0	16	0.299	4	0.129	21	0.003
15	0.001	12	0.6	19	−0.203	20	0.080	7	0.001
20	−0.001	9	−0.5	18	−0.153	6	0.077	5	0.000
10	0.000	6	0.0	14	−0.017	5	0.052	6	0.000

All the statistics are given in a descending order in absolute values. Obs. = observations.

Based on ΔD_i and ΔF_i, we see that r_i^2 and q_i play key roles in determining influential observations on R^2 and F. Both R^2 and F increase as we delete i^{th} observation with large residual r_i^2; however, they decrease as we delete i^{th} observation with large q_i. This result is quite interesting in the sense that the Cook's distance is an increasing function of h_{ii}, but both R^2 and F are a decreasing function of q_i. Recall that q_i can be regraded as the i^{th} leverage of the response, while h_{ii} is the leverage of covariates.

Remark 1. In regression diagnostics, "Influential on what?" is a really important question for a particular influence measure. For example, the Cook's distance (Cook, 1977) is an influence measure to see the effect of observations on the (scaled) change of the estimates of regression coefficients. Therefore, influential observations detected by the Cook's distance may not be influential to other estimators such as the change of variance estimate or the F-statistic. Two influence measures, ΔD_i and ΔF_i, proposed in this paper are related to the significance of the postulated model. Therefore, ΔD_i deals with the influence on the change of the proportion of regression sum of squares among the SST. However, ΔF_i is concerned about the change of the F-statistic when testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$. Hence, influential observations based on ΔD_i may not be influential based on ΔF_i, and vice versa.

3. Numerical study

As an illustrative example for the proposed influence measures, we consider the stackloss data, consisting of one response variable Y and three explanatory variables X_1, X_2, X_3 with 21 observations. After fitting the response on three covariates using the linear model, we obtained that $R^2 = 0.9136$ and $F = 59.9$.

Now, we compute the proposed influence measures $\Delta D_i, \Delta F_i$ with the standardized residual r_i, the leverage h_{ii}, and the proportion of deviation q_i in the stackloss data in Table 1. Table 2 also summarizes the five largest values of ΔD_K and ΔF_K when $k = 1, 2, 3$. From both tables, we note the following:

Table 2: Five largest values (in absolute) of ΔD_K and ΔF_K when $k = 1, 2, 3$ in the stackloss data

k	K	ΔD_K	K	ΔF_K
1	21	−0.035	21	−38.921
	1	0.027	1	18.357
	2	0.017	2	13.853
	4	−0.014	4	−8.373
	3	0.006	3	7.249
2	1, 2	0.079	4, 21	−96.900
	4, 21	−0.056	6, 21	−40.427
	1, 3	0.039	13, 21	−39.813
	6, 21	−0.039	9, 21	−37.202
	13, 21	−0.039	5, 21	−35.925
3	1, 2, 3	0.153	4, 13, 21	−121.462
	1, 2, 16	0.091	4, 6, 21	−100.135
	1, 2, 18	0.088	3, 4, 21	−99.731
	1, 2, 17	0.086	4, 15, 21	−98.469
	1, 2, 19	0.085	4, 5, 21	−93.470

From ΔD_i and ΔF_i, we see that four observations $\{21\}, \{1\}, \{2\}, \{4\}$ seem to be influential on R^2 and F. They also have large values of r_i, h_{ii}, and q_i. Among 4 influential observations, deletion of $\{21\}$ and $\{4\}$ increases R^2 and F, however, deletion of $\{1\}$ and $\{2\}$ decreases R^2 and F. This phenomenon occurs when $k = 2$, i.e., influential sets are $\{1, 2\}$ and $\{4, 21\}$. If we delete $\{4, 21\}$, then R^2 increases from 0.9136 to 0.9696, and F increases from 59.9 to 156.8.

We should also note that both R^2 and F increase as we delete i^{th} observation with large residual r_i^2; however, they decrease as we delete i^{th} observation with large q_i, i.e., when we delete $\{21\}$, observation with large residual r_i, both R^2 and F increase, however, they decrease as we delete observations either $\{1\}$, $\{2\}$, or $\{3\}$. This phenomenon occurs when we delete multiple cases of observations (Table 2).

Table 2 also indicates that observation $\{21\}$ is very influential on both the coefficient of determination and the F-statistic; therefore, when we delete observation $\{21\}$ the F-statistic increases from 59.9 to 98.8. We now evaluate $\Delta t_i = t - t_{(i)}$ for each covariate to see how the t-statistic for each covariate change as we delete one observation (Table 3). Here, we see that when we delete observation $\{21\}$, the t-statistic for X_1 increases from 5.307 to 7.481, while the t-statistic for X_2 decreases from 3.520 to 2.513. Therefore, when we find influential observations on the F-statistic, it is necessary to find which covariate is seriously affected by the influential observations.

4. Concluding remarks

Most of diagnostic issues are studied about regression models. In this paper, we proposed influence measures for two basic goodness-of-fit statistics, the coefficient of determination R^2 and the overall significance test statistic F. We have derived analytic formula for two measures and expressed them as basic building blocks such as residuals, leverages and deviations. From the analytic expressions, we found that both R^2 and F increase as we delete i^{th} observation with large residual r_i^2; however, they decrease as we delete i^{th} observation with large q_i. We also applied them to the real data set and verified that deletion of one or few influential observations may result in big change in R^2 and F.

It would be useful for further research to study the cutoff or threshold for the proposed influence measures. We also expect that the bootstrap method will be one other possible method.

Table 3: Values of $\Delta t_i = t - t_{(i)}$, change in the t-statistic when the i^{th} observation is deleted, in the stackloss data

$X_0 \equiv 1$		X_1		X_2		X_3	
Obs.	Value	Obs.	Value	Obs.	Value	Obs.	Value
21	1.249	21	−2.174	21	1.007	7	0.052
17	−1.115	4	−0.771	4	0.438	8	−0.109
13	−0.294	1	0.581	13	0.364	1	−0.048
18	−0.292	7	0.562	3	−0.168	10	−0.130
19	−0.241	8	0.522	14	0.164	15	−0.185
14	−0.227	9	0.494	9	−0.162	17	−1.115
15	−0.185	10	0.454	19	0.135	11	−0.100
4	0.175	12	0.383	10	0.113	2	−0.037
10	−0.130	19	0.373	6	−0.106	12	−0.018
8	−0.109	3	0.288	18	0.103	4	0.175
11	−0.100	18	0.236	8	0.100	21	1.249
16	−0.092	11	0.182	20	0.090	14	−0.227
9	0.078	16	0.177	16	0.088	20	−0.056
5	−0.061	13	0.163	12	−0.085	9	0.078
20	−0.056	14	0.163	11	−0.081	13	−0.294
3	−0.055	6	0.148	1	0.068	5	−0.061
7	0.052	2	0.144	7	−0.053	18	−0.292
1	−0.048	5	0.127	5	0.051	6	0.029
2	−0.037	17	0.124	17	0.023	19	−0.241
6	0.029	20	0.118	15	0.011	16	−0.092
12	−0.018	15	0.016	2	0.005	3	−0.055

Obs. = observations.

Acknowledgement

This work was supported by the 2016 Inje University Research Grant.

References

Bae W, Hwang S, and Kim C (2008). Influence diagnostics in the varying coefficient model with longitudinal data, *Computational Statistics*, **23**, 185–196.

Belsley DA, Kuh E, and Welsch RE (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York.

Box GEP and Cox DR (1964). An analysis of transformations, *Journal of the Royal Statistical Society Series B (Methodological)*, **26**, 211–252.

Chatterjee S and Hadi AS (1986). Influential observations, high leverage points, and outliers in linear regression, *Statistical Science*, **1**, 379–393.

Cook RD (1977). Detection of influential observation in linear regression, *Technometrics*, **19**, 15–18.

Cook RD and Wang PC (1983). Transformations and influential case in regression, *Technometrics*, **25**, 337–343.

Cook RD and Weisberg S (1982). Characterizations of an empirical influence function for detecting influential cases in regression, *Technometrics*, **22**, 495–508.

Eubank RL (1985). Diagnostics for smoothing splines, *Journal of the Royal Statistical Society Series B (Methodological)*, **47**, 332–341.

Fung WK, Zhu ZY, Wei BC, and He X (2002). Influence diagnostics and outlier tests for semiparametric mixed models, *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **64**, 565–579.

Hinkley DV and Wang S (1988). More about transformations and influential cases in regression, *Technometrics*, **30**, 435–440.

Hoerl AE and Kennard RW (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55–67.

Kim C (1996). Local influence and replacement measure, *Communications in Statistics - Theory and Methods*, **25**, 49–61.

Kim C, Storer BE, and Jeong M (1996). Note on Box-Cox transformation diagnostics, *Technometrics*, **38**, 178–180.

Kim C, Lee Y, and Park BU (2001). Cook's distance in local polynomial regression, *Statistics and Probability Letters*, **54**, 33–40.

Kim C, Park BU, and Kim W (2002). Influence diagnostics in semiparametric regression models, *Statistics and Probability Letters*, **60**, 49–58.

Kim C, Lee J, Yang H, and Bae W (2015). Case influence diagnostics in the lasso regression, *Journal of the Korean Statistical Society*, **44**, 271–279.

Silverman BW (1985). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Tsai CL and Wu X (1990). Diagnostics in transformation and weighted regression, *Technometrics*, **32**, 315–322.

Walker E and Birch JB (1988). Influence measures in ridge regression, *Technometrics*, **30**, 221–227.

Received December 16, 2016; Revised January 18, 2017; Accepted January 20, 2017