Communications of the Korean Statistical Society Vol. 15, No. 2, 2008, pp. 229–234

Cook-Type Influence Measure in Constrained Regression Models

Myung Geun Kim[1)]

Abstract

A Cook-type distance is considered for investigating the influence of observations in constrained regression models. Its exact sampling distribution is derived, which is used for judging whether each observation is influential or not. A numerical example is provided for illustration.

Keywords: Constrained regression; Cook-type distance; influence.

1. Introduction

Cook's distance (Cook, 1977; Cook and Weisberg, 1982) in linear regression has been used for measuring the influence of observations in estimating regression coefficients. It is based on the confidence ellipsoid for regression coefficients. It can be considered as a scaled difference between the estimates of regression coefficients computed with and without some observations. It is generalized to constrained regression models by Kim (2007) in which some relevant works can be found. However, it has two main drawbacks. One is that the exact sampling distribution of Cook's distance is not available. Hence we can not use the cut-off values for deciding which observations are influential and the use of Cook's distance should be a rule of thumb. The other is that the same scaling matrix is used for scaling all of the differences between the estimates computed with and without some observations, even though the covariance matrices of the differences do not have the same form.

In this work we will suggest a Cook-type distance in constrained regression models that will overcome the defects of the usual Cook's distance mentioned in the previous paragraph. In Section 2 some results for constrained regression are reviewed. In Section 3 a Cook-type distance in constrained regression is suggested. In Section 4 the sampling distribution of the Cook-type distance suggested in Section 3 is derived. In Section 5 a numerical example is given for illustration.

1) Professor, Depatment of Mathematics Education, Seowon University, 231 Mochung-Dong, Cheongju, Chungbuk 361-742, Korea. E-mail: mgkim@seowon.ac.kr

2. Preliminaries

The constrained regression of our interest in which some linear restrictions are imposed on regression coefficients can be expressed as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with

$$\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c},$$

where $\boldsymbol{y}$ is an n by 1 vector of response variables, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is an n by p matrix of fixed independent variables, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^T$ is a p by 1 vector of unknown regression parameters, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an n by 1 vector of unobservable errors, $\boldsymbol{A}$ is a known q by p ($q \leq p$) matrix of rank q and $\boldsymbol{c}$ is a known q by 1 vector. Further, we assume that the unobservable errors ε_r ($r = 1, \ldots, n$) are independent and identically distributed as a normal distribution with mean zero and unknown variance σ^2.

The least squares estimator(LSE) of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\tilde{\boldsymbol{\beta}} - \boldsymbol{c}),$$

where $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ (Searle, 1971, Section 3.6). We define

$$\boldsymbol{H} = (h_{ij}) = \tilde{\boldsymbol{H}} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T,$$

where $\tilde{\boldsymbol{H}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Then it is easy to show that $\boldsymbol{H}$ is symmetric and idempotent. The residual vector for constrained regression is written as $\boldsymbol{e} = (e_1, \ldots, e_n)^T = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Then we get another expression for the residual vector

$$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{c}.$$

We can easily show that the distribution of $\boldsymbol{e}$ is an n-variate normal with zero mean vector and the covariance matrix given by

$$\mathrm{cov}(\boldsymbol{e}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H}).$$

When we denote the error sum of squares by $\mathrm{SSE} = \boldsymbol{e}^T\boldsymbol{e}$, $\hat{\sigma}^2 = \mathrm{SSE}/(n - p + q)$ becomes an unbiased estimator of σ^2.

3. Cook-Type Distance for Constrained Regression

We denote by $\hat{\boldsymbol{\beta}}_{(i)}$ the LSE of $\boldsymbol{\beta}$ computed without the i^{th} case. From Kim (2007), we have

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{e_i}{1 - h_{ii}}\boldsymbol{V}\boldsymbol{x}_i, \tag{3.1}$$

where $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X})^{-1} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Note that $\boldsymbol{V}$ is singular since $\boldsymbol{A}\boldsymbol{V}$ becomes a zero matrix. The covariance matrix of $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$ is computed as

$$\mathrm{cov}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}) = \frac{\sigma^2}{1 - h_{ii}}\boldsymbol{V}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{V}.$$

The rank of $\boldsymbol{V}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{V}$ is one (for its proof, refer to Mardia *et al.*, 1979, p. 471) for nonnull vector $\boldsymbol{V}\boldsymbol{x}_i$. *Hence the covariance matrix of* $\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}$ *is singular. For nonnull vector* $\boldsymbol{V}\boldsymbol{x}_i$, the only nonzero eigenvalue of $\boldsymbol{V}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{V}$ is $\boldsymbol{x}_i^T\boldsymbol{V}^2\boldsymbol{x}_i$ and its associated nonnormalized eigenvector is $\boldsymbol{V}\boldsymbol{x}_i$ (for their derivations, refer to Mardia *et al.*, 1979, Corollary A.6.2.1). A geometrical meaning of this result is that $\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}$ lies entirely along the axis determined by $\boldsymbol{V}\boldsymbol{x}_i$. That is, $\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}$ is totally determined by $\boldsymbol{V}\boldsymbol{x}_i$, which coincides with the real computation given in (3.1). Hence it is reasonable to scale the difference between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$ with respect to the coordinate axis $\boldsymbol{V}\boldsymbol{x}_i$. However, the conventional Cook's distance is obtained by scaling all of the differences $\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}$ by the same matrix, for example $\boldsymbol{X}^T\boldsymbol{X}$ in the unconstrained regression models.

Let c_i be the length of $\boldsymbol{V}\boldsymbol{x}_i$. Then $c_i^2=\boldsymbol{x}_i^T\boldsymbol{V}^2\boldsymbol{x}_i$. A generalized inverse (for its definition, refer to Searle, 1971) of $\boldsymbol{V}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{V}$ is easily computed as

$$(\boldsymbol{V}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{V})^- = c_i^{-4}\boldsymbol{V}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{V}.$$

In view of the previous paragraph, a Cook-type distance between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$ can be defined by and computed as

$$D_i=(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)})^T[\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)})]^-(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)})=\frac{e_i^2}{\hat{\sigma}^2(1-h_{ii})}.$$

A large value of D_i indicates that the i^{th} case is influential in estimating $\boldsymbol{\beta}$. In the next section the sampling distribution of the Cook-type statistic D_i is derived, which is used for deciding which observations are influential.

4. Sampling Distribution of the Cook-Type Statistic D_i

In what follows the subscript (i) indicates the removal of the i^{th} case in computing the corresponding quantity as in $\hat{\boldsymbol{\beta}}_{(i)}$. Then $\mathrm{SSE}_{(i)}$ is the error sum of squares computed *without the* i^{th} *case. Since* $\mathrm{SSE}_{(i)}=(\boldsymbol{y}_{(i)}-\boldsymbol{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)})^T(\boldsymbol{y}_{(i)}-\boldsymbol{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)})$, a little computation shows that

$$\mathrm{SSE}_{(i)}=\mathrm{SSE}-\frac{e_i^2}{1-h_{ii}}. \tag{4.1}$$

Since $\boldsymbol{e}=(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}$, we have $\mathrm{SSE}=\boldsymbol{\varepsilon}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}$. Let $\boldsymbol{q}_i$ be the $n\times 1$ vector whose i^{th} element is one and whose other elements are all zero. Since the i^{th} residual can be expressed as $e_i=\boldsymbol{q}_i^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}$, its square is $e_i^2=\boldsymbol{\varepsilon}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{q}_i\boldsymbol{q}_i^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}$. From (4.1), we get

$$\mathrm{SSE}_{(i)}=\boldsymbol{\varepsilon}^T\boldsymbol{Q}_i\boldsymbol{\varepsilon},$$

where $\boldsymbol{Q}_i=\boldsymbol{I}-\boldsymbol{H}-(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{q}_i\boldsymbol{q}_i^T(\boldsymbol{I}-\boldsymbol{H})/(1-h_{ii})$. We can easily see that $\boldsymbol{Q}_i$ is symmetric and idempotent. Hence the rank of $\boldsymbol{Q}_i$ is equivalent to its trace, $\mathrm{tr}(\boldsymbol{Q}_i)=n-p+q-1$. By Corollary 2.1 in p. 58 of Searle (1971), $\mathrm{SSE}_{(i)}/\sigma^2$ is distributed as a chi-squares distribution with degrees of freedom $n-p+q-1$.

We can easily check that $\boldsymbol{q}_i^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Q}_i = \boldsymbol{0}$. By Theorem 3 in p. 59 of Searle (1971), the two quadratic forms e_i^2 and $\text{SSE}_{(i)}$ are independent. Hence the i^{th} residual R_i can be defined as

$$R_i^2 = \frac{e_i^2}{(1-h_{ii})\sigma^2} \Big/ \frac{\text{SSE}_{(i)}}{(n-p+q-1)\sigma^2} = \frac{e_i^2}{\hat{\sigma}_{(i)}^2(1-h_{ii})}, \tag{4.2}$$

where $\hat{\sigma}_{(i)}^2 = \text{SSE}_{(i)}/(n-p+q-1)$, which is distributed as an F distribution with degrees of freedom 1 and $n-p+q-1$.

Using (4.1), it is easy to find the following identity

$$\frac{\text{SSE}}{\text{SSE}_{(i)}} = 1 + \frac{R_i^2}{n-p+q-1}. \tag{4.3}$$

Further, from (4.2) we have

$$R_i^2 = \frac{n-p+q-1}{1-h_{ii}} \cdot \frac{e_i^2}{\text{SSE}} \cdot \frac{\text{SSE}}{\text{SSE}_{(i)}}.$$

Hence we get the following relationship between D_i and R_i^2 by using (4.3)

$$1 - \frac{D_i}{n-p+q} = \left[1 + \frac{R_i^2}{n-p+q-1}\right]^{-1}. \tag{4.4}$$

By Theorem 7 in p. 319 of Rohatgi (1976), the left-hand side in (4.4) is distributed as a beta distribution as follows

$$1 - \frac{D_i}{n-p+q} \sim \text{Beta}\left(\frac{n-p+q-1}{2}, \frac{1}{2}\right).$$

It is well known that if a random variable Z is distributed as $\text{Beta}(a,b)$, then $1-Z$ is distributed as $\text{Beta}(b,a)$. Hence the distribution of $D_i/(n-p+q)$ is

$$\frac{D_i}{n-p+q} \sim \text{Beta}\left(\frac{1}{2}, \frac{n-p+q-1}{2}\right).$$

5. A Numerical Example

We illustrate the use of the Cook-type distance D_i for identifying influential observations by using the stackloss data (Brownlee, 1965) which comprises 21 measurements on a single response variable and three independent variables. The stackloss data set can also be found in Table 6.3 of Chatterjee and Hadi (1988). Some detailed analysis of this data set was made by Kim (1998).

We will fit the constrained regression model with the following linear relationship $5\beta_2 + 43\beta_3 = 0$ to the stackloss data. We can use the usual F-test (see, Searle, 1971, Chapter 3) to check if this relationship holds for the stackloss data. The value of the F-test statistic is computed as $8.66 * 10^{-5}$ and its associated p-value is 0.993. Hence we

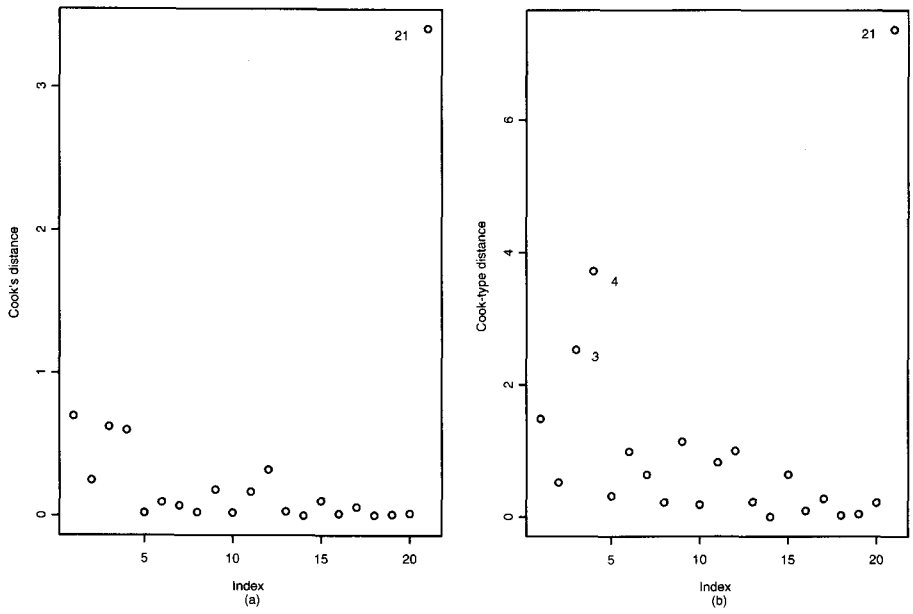

Figure 5.1: An index plot (a) of the CD_i and (b) of the D_i

can conclude that the relationship $5\beta_2 + 43\beta_3 = 0$ holds at any reasonable significance level.

For this constrained regression model, we compute the usual Cook's distances CD_i adapted to the constrained regression model by Kim (2007) and the Cook-type distances D_i derived in Section 3. An index plot of the usual Cook's distances CD_i is included in Figure 5.1(a), and that of the Cook-type distances D_i is provided in Figure 5.1(b). In both index plots, the influence of case 21 in estimating β is remarkable relative to the others: $CD_{21} = 3.41$ and $D_{21} = 7.36$. While the influence of the other cases except for case 21 based on the CD_i is not severe, cases 3 and 4 ($D_3 = 2.53$ and $D_4 = 3.71$) are also a little influential based on the D_i.

As is well known, the exact sampling distribution of each Cook's distance CD_i is not available. However, the exact sampling distribution of each Cook-type distance $D_i/(n-p+q)$ has been derived in Section 4. Hence we are in a position to judge whether each case is really influential or not. Since $n - p + q = 18$, each $D_i/(n - p + q)$ follows a beta distribution Beta(0.5, 8.5). For case 21, we have $D_i/(n-p+q) = 0.409$ which is the 99.7% percentile of Beta(0.5, 8.5). Influential cases are those that have a large influence in estimating regression coefficients, compared with the other cases. Hence it is reasonable to draw to a conclusion that case 21 is an influential case in estimating the regression coefficients for the above constrained regression model. For case 4, $D_i/(n-p+q) = 0.206$

and it is the 94.9% percentile. For case 3, $D_i/(n-p+q) = 0.141$ and it is the 88.6% percentile. At the 5% significance level, the influence of case 3 and 4 is not severe.

6. Concluding Remarks

The sampling distribution of Cook's distance adapted to constrained regression models is not available. Therefore we have no cut-off values that can be used for deciding which observations are influential and the use of Cook's distance should be a rule of thumb. In this work, a Cook-type distance in constrained regression models overcoming the defects of the usual Cook's distance has been suggested. The sampling distribution of the Cook-type distance is derived and it is useful for measuring the degree of the influence of anomalous data.

References

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering.* 2nd ed., John Wiley & Sons, New York.

Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression.* John Wiley & Sons, New York.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15–18.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman & Hall/CRC, New York.

Kim, M. G. (1998). Local influence on a test of linear hypothesis in multiple regression models. *Journal of Applied Statistics*, **25**, 145–152.

Kim, M. G. (2007). Influence analysis of constrained regression model. *The Korean Communications in Statistics*, **14**, 281–286.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, London.

Rohatgi, V. K. (1976). *An Introduction to Probability Theory and Mathematical Statistics.* John Wiley & Sons, New York.

Searle, S. R. (1971). *Linear Models.* John Wiley & Sons, New York.

[Received December 2007, Accepted January 2008]