The Korean Communications in Statistics **Vol. 14** No. 2, 2007, pp. 281–286

Influence Analysis of Constrained Regression Models

Myung Geun Kim[1]

Abstract

Cook's distance is generalized to the multiple linear regression with linear constraints on regression coefficients. It is used for identifying influential observations in constrained regression models. A numerical example is provided for illustration.

Keywords: Constrained regression; Cook's distance; influence.

1. Introduction

Diagnostic methods for the usual regression models have been suggested by many authors and some of them can be found in Cook and Weisberg (1982), Chatterjee and Hadi (1986), and Barnett and Lewis (1994). Constrained regressions are used widely in the field of econometrics, for example in the estimation of Cobb-Douglas production functions (Chipman and Rao, 1964). However, very few works have been done for constrained regression. Paula (1993, 1999) considered influence analysis of inequality-constrained models and Kim (2003) suggested a local influence method of detecting outliers.

Cook's distance for the usual regression models has been used for identifying influential observations that have a great effect on the estimates of regression coefficients (Cook, 1977; Cook and Weisberg, 1982) and it has been generalized to various statistical models, for example to structural equation models (Lee and Wang, 1996). However, no method of detecting influential observations based on Cook-type distance is available for constrained regression models. The counterparts of some diagnostic statistics, such as Andrews-Pregibon statistic, variance ratio, *etc,* suggested in unconstrained regression are not available in constrained regression.

[1] Professor, Department of Applied Statistics, Seowon University, 231 Mochung-Dong, Cheongju, Chungbuk 361-742, Korea.
E-mail : mgkim@seowon.ac.kr

In this work we will suggest a diagnostic method based on Cook's distances for constrained regression. In Section 2 some results for constrained regression are reviewed. In Section 3 Cook's distance is generalized to constrained regression and the generalized Cook's distance is derived. In Section 4 a numerical example is provided for illustration. In Section 5 concluding remarks are made.

2. Preliminaries for Constrained Regression

In this section we will review some results for the multiple linear regression with linear constraints on regression coefficients.

We consider the following constrained regression

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

with

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

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 ($q \le p$) 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 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 $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. The residual vector for constrained regression is written as $\boldsymbol{e} = (e_1, \ldots, e_n)^T = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Similarly, we let $\tilde{\boldsymbol{e}} = (\tilde{e}_1, \ldots, \tilde{e}_n)^T = \boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}$ for unconstrained regression. Denoting the hat matrix by $\tilde{\boldsymbol{H}} = (\tilde{h}_{ij}) = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ for unconstrained regression, we define

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

We note that $\boldsymbol{H}$ is symmetric and idempotent. Further, we put $\hat{\sigma}^2 = \boldsymbol{e}^T\boldsymbol{e}/n$ which is the maximum likelihood estimator of σ^2. More details can be found in Chap. 4 of Seber (1977).

3. A Generalized Cook's Distance for Constrained Regression

It is clear that the least squares estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. We can easily compute the covariance matrix of $\hat{\boldsymbol{\beta}}$ as

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{V},$$

where $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X})^{-1} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Since $\boldsymbol{AV}$ becomes a zero matrix, the covariance matrix of $\hat{\boldsymbol{\beta}}$ is singular. Hence based on the idea for Cook's distance, a generalized Cook's distance on the estimates $\hat{\boldsymbol{\beta}}$ for constrained regression models can be defined as

$$CD_i = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T[\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})]^-(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}) = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T\boldsymbol{V}^-(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{\hat{\sigma}^2}$$

ignoring unimportant constant terms, where $\hat{\boldsymbol{\beta}}_{(i)}$ is the LSE of $\boldsymbol{\beta}$ computed without the i-th case and $[\text{cov}(\hat{\boldsymbol{\beta}})]^-$ is a generalized inverse of $\text{cov}(\hat{\boldsymbol{\beta}})$ (for its definition, refer to Schott, 1997). A large value of CD_i indicates that the i-th case is influential in estimating $\boldsymbol{\beta}$. At present it is not easy to directly find a closed form of a generalized inverse of $\boldsymbol{V}$ and thus it seems difficult to directly compute the generalized Cook's distance CD_i by finding $[\text{cov}(\hat{\boldsymbol{\beta}})]^-$. We will see later that luckily this difficulty can be avoided.

In order to get a closed form of expression for the generalized Cook's distance CD_i, first we need to compute

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

where the subscript (i) indicates the removal of the i^{th} case in computing the corresponding quantity as in the above paragraph. We easily compute

$$\tilde{\boldsymbol{\beta}}_{(i)} = \tilde{\boldsymbol{\beta}} - \frac{\tilde{e}_i(\mathbb{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i}{1 - \tilde{h}_{ii}},$$

$$(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1} = (\boldsymbol{X}^T\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}}{1 - \tilde{h}_{ii}}.$$

Thus we can compute

$$[\boldsymbol{A}(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{A}^T]^{-1} = [\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1} - \frac{\boldsymbol{Q}_i}{1 - h_{ii}},$$

where

$$\boldsymbol{Q}_i = [\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}.$$

Then we easily get

$$\begin{aligned}&(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{A}^T]^{-1}\\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1} - \frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T\boldsymbol{Q}_i}{1-h_{ii}} + \frac{\boldsymbol{W}_i}{1-h_{ii}},\end{aligned}$$

where

$$\boldsymbol{W}_i = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}.$$

So it is easy to obtain

$$\begin{aligned}&(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\tilde{\boldsymbol{\beta}}_{(i)} - \boldsymbol{c})\\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\tilde{\boldsymbol{\beta}} - \boldsymbol{c})\\ &\quad -\frac{d_{1i}+\tilde{e}_i}{1-h_{ii}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\\ &\quad +\left(\frac{d_{1i}}{1-h_{ii}} - \frac{\tilde{e}_i(\tilde{h}_{ii}-h_{ii})}{(1-\tilde{h}_{ii})(1-h_{ii})}\right)(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\\ &\quad +\frac{1}{1-h_{ii}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T\boldsymbol{Q}_i\boldsymbol{c} - \frac{1}{1-h_{ii}}\boldsymbol{W}_i\boldsymbol{c},\end{aligned}$$

where

$$d_{1i} = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{A}\tilde{\boldsymbol{\beta}}.$$

Using the above results, a little more computation yields

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

Then the generalized Cook's distance reduces to

$$CD_i = \frac{e_i^2}{(1-h_{ii})^2\hat{\sigma}^2}\boldsymbol{x}_i^T\boldsymbol{V}\boldsymbol{V}^-\boldsymbol{V}\boldsymbol{x}_i = \frac{e_i^2}{(1-h_{ii})^2\hat{\sigma}^2}\boldsymbol{x}_i^T\boldsymbol{V}\boldsymbol{x}_i$$

whose second equality follows from the definition of a generalized inverse. From the definition of $\boldsymbol{H}$, we get

$$h_{ii} = \tilde{h}_{ii} - \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T[\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i.$$

Since $\tilde{h}_{ii} = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$, we have

$$h_{ii} = \boldsymbol{x}_i^T\boldsymbol{V}\boldsymbol{x}_i.$$

Hence the generalized Cook's distance becomes

$$CD_i = \frac{h_{ii}}{(1-h_{ii})^2}\frac{e_i^2}{\hat{\sigma}^2}.$$

4. A Numerical Example

We illustrate the use of the generalized Cook's distance CD_i to identify influential observations using the body fat data (Neter *et al.*, 1996, p. 261). This data set consists of twenty measurements on a single response variable (y) and three independent variables (X_1, X_2, X_3) for which the regression model can be written as

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \text{error}.$$

We consider the following linear relationship

$$50\beta_1 + 99\beta_3 = 0.$$

In order to check the adequacy of the above linear relationship for the body fat data, we can use the usual F-test for linear hypothesis given in Chap. 4 of Seber (1977). The value of the F-test statistic is 0.0003 and the associated p-value is 0.987. Hence it is reasonable to conclude that this linear relationship holds in the regression model for the body fat data. Figure 4.1 shows an index plot of the generalized Cook's distances for the constrained regression with the above linear relationship.

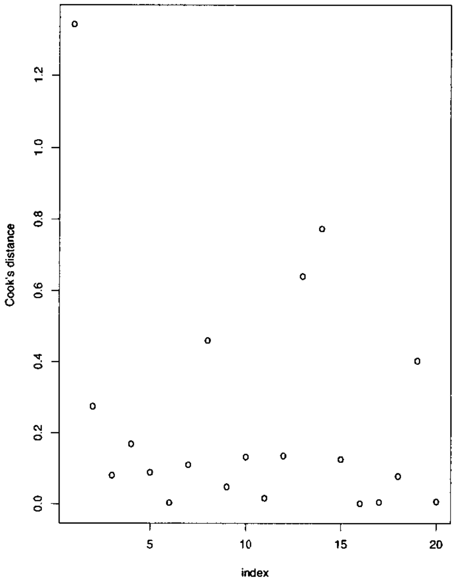

Figure 4.1: An index plot of the generalized Cook's distances

We see from Figure 4.1 that case 1 is the most influential in estimating $\boldsymbol{\beta}$. Case 14 is the next but its influence is not severe compared with that of case 1. We note that case 1 is also identified as an outlier using the local influence method of detecting outliers for constrained regression models suggested by Kim (2003).

5. Concluding Remarks

No method of detecting influential observations based on Cook-type distance is available for constrained regression models. In Section 3 Cook's distance was generalized to constrained regression models, and it may yield a clue to a useful diagnostic method of detecting influential observations for constrained regression.

References

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data.* 3rd ed., John Wiley & Sons, New York.

Chatterjee, S. and Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression (with discussions). *Statistical Science*, **1**, 379–416.

Chipman, J. S. and Rao, M. M. (1964). The treatment of linear restrictions in regression analysis. *Econometrica*, **32**, 198–209.

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, New York.

Kim, M. G. (2003). Detection of outliers in constrained regression. *The Korean Communications in Statistics*, **10**, 519–524.

Lee, S. -Y. and Wang, S. J. (1996). Sensitivity analysis of structural equation models. *Psychometrika*, **61**, 93–108.

Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996). *Applied Linear Regression Models.* 3rd ed., McGraw-Hill/Irwin.

Paula, G. A. (1993). Assessing local influence in restricted regression models. *Computational Statistics & Data Analysis*, **16**, 63–79.

Paula, G. A. (1999). Leverage in inequality-constrained regression models. *The Statistician*, **48**, 529–538.

Schott, J. R. (1997). *Matrix Analysis for statistics.* John Wiley & Sons, New York.

Seber, G. A. F. (1977). *Linear Regression Analysis.* John Wiley & Sons, New York.

[Received January 2007, Accepted March 2007]