Journal of Management
Vol. V, No.1, October 2009, pp 41-49

Assessment of Anomalous Observations in Liu Estimator

Chen Jianbao
Department of Planning and Statistics
Xiamen University, China.
jbjy2006@126.com

And

Aboobacker Jahufer
Department of Mathematical Sciences
South Eastern University of Sri Lanka.
jahufer@yahoo.com

Abstract

Liu linear regression model building analysis is frequently applied for fitting model to near multicollinearity data sets. When atypical observations exist in a data set, they may exert undue influence on the result of the analysis. This paper studies an assessment of the minor perturbation of the Liu estimator in the ridge type linear regression model using Cook's (1986) method to detect anomalous observations in the data set when mean squared error is known or unknown, and perturbation of individual explanatory variable. Example based on Longley data are used for illustration.

Keywords: Influential Observations; Liu Estimator; Local Influence; Cook's Method.

Introduction

The fact that a small subset of the data can have a disproportionate influence on the estimated parameters or predictions is of concern to users of regression analysis, for, if this is the case, it is quite possible that the model-estimates are based primarily on this data subset rather than on the majority of the data.

Atypical observations in a data set may exert undue influence on a statistical analysis, leading to a misleading result; it is therefore of importance to identify such observations. In the context of ridge type estimators, an influential observation may cause a substantial change in the biasing parameter and prediction. Much effort has been devoted to the detection of influential observations in ridge type estimators. Among others, Shi and Wand (1999) studied the local influence of minor perturbations on the ridge estimator in the ridge regression model. The diagnostics under the perturbation of variance and explanatory variables are derived. Also, they developed a new technique for the detection of influential observations on the ridge biasing parameter. Moreover, Shi (1997) used local influence in principal component analysis. Walker and Birch (1988) analyzed the influence of observations in ridge regression using case deletion method. The main objective of the current study is to assess local influential observations in Liu estimator using the local influence approach.

The local influence approach was proposed by Cook (1986) as a general method for assessing the influence of minor perturbations of a statistical model, and the approach has been applied to a number of influence analysis problems (see, e.g., Beckman et al., 1987; Schall and Dunne, 1992; Thomas and Cook, 1990). The method makes use of differential geometry techniques to assess the change of the likelihood function due to perturbations of the model.

This paper is composed of five sections: Section 1 describes the introduction, Section 2 gives the background and definition, Section 3 derives the detecting local influential observations for Liu estimator, and Section 4 provides example and results. Discussion is given in the last section.

Backgrounds and Definition

A multiple linear regression model in matrix form can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (1)$$

where $\mathbf{Y}$ is an $n \times 1$ observation random vector, $\mathbf{X}$ is an $n \times p$ known matrix, β is a $p \times 1$ vector of unknown parameters, ε is an $n \times 1$ error vector with $E(\varepsilon)=0$ and $Var(\varepsilon)=\sigma^2\mathbf{I}_n$ and $\mathbf{I}_n$ is an identity matrix of the order n. The ordinary least squares estimator (OLSE) of β is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and the estimator of σ^2 is $s^2 = \dfrac{\mathbf{e}'\mathbf{e}}{(n-p)}$, where residual vector $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.

The existence of multicollinearity in the linear regression model can lead to a very sensitive least-squares

estimates, therefore mixed estimation and ridge regression are suggested to mitigate the effect of multicollinearity.

Many authors noted, the influence of the observations on ridge regression is different from the corresponding least squares estimate, and multicollinearity can even disguise anomalous data (Belsley et al. 1980, Belsley, 1991, Steece 1986, Walker and Birch 1988, Shi 1997). One of biased estimators of β used when multicollinearity is present in the data, the ridge regression estimator introduced by Hoerl (1964), is defined by

$$\hat{\boldsymbol{\beta}}_R = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}, \qquad (2)$$

where **I** identity matrix and k>0 is the so called ridge estimator biasing parameter. This estimator is more stable than the least-squares estimator.

Besides to overcome near multicollinearity, Liu (1993) combined the Stein (1956) estimator with ordinary ridge regression estimator (ORRE) to obtain what we call the Liu estimator. The Liu estimator is defined by

$$\hat{\boldsymbol{\beta}}_d = (\mathbf{X}'\mathbf{X} + \mathbf{I})^{-1}(\mathbf{X}'\mathbf{Y} + d\hat{\boldsymbol{\beta}}), \qquad (3)$$

where d is the Liu estimator biasing parameter.

A prediction criterion suggested by Liu (1993) selects d by minimizing C_L statistics introduced by Mallows (1973) and it is given by

$$C_L = \frac{SSR_d}{s^2} + 2tr(\mathbf{H}_d) - n, \qquad (4)$$

where SSR_d is the sum of the squares of residuals from Liu estimation, s^2 is the estimator of σ^2 from least squares regression, and the hat matrix of Liu estimator is

$$\mathbf{H}_d = \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{I})^{-1}(\mathbf{X}'\mathbf{X} + d\mathbf{I})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Detecting Local Influential Observations for Liu Estimator

Perturbation of the Liu Estimator

We start by giving a brief sketch of the local influence approach suggested by Cook, (1986) and Lawrance, (1991). The assumption of constant variance in model (1) is perturbed. This perturbation scheme is a better way to handle cases badly modeled (Lawrance, 1988). The distribution of ε under perturbation becomes

$$\varepsilon_\omega \sim N(0, \sigma^2\mathbf{W}^{-1}), \qquad (5)$$

where **W**=diag(ω) is a diagonal matrix with diagonal elements of $\omega' = (\omega_1, \omega_2, ..., \omega_n)$. Let $\omega = \omega_0 + al$, where ω_0 = 1, ω denote the an n × 1 vector of case-weights for the regression model (1), $a \in R^1$ and l is a fixed nonzero vector of unit length in R^q.

Local Influential Analysis for Liu Estimator When σ^2 is known

It is assumed that the mean squared error σ^2 is known, then the relevant part of the multiplicative perturbed likelihood model for Liu estimator is

$$L(\varepsilon_\omega) = \prod_{i=1}^{n}(2\pi\sigma^2)^{\frac{-n}{2}}|\mathbf{W}|^{\frac{1}{2}}\exp\left[-\frac{1}{2\sigma^2}\mathbf{e}_d'W^{-1}\mathbf{e}_d\right], \qquad (6)$$

where $\mathbf{e}_d = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_d)$.

Based on Cook (1986), the influential patterns for Liu estimator can be identified by studying the eigenvector $l_{\max}^m$ associated with maximum eigenvalue of the normal

curvature matrix C_d and it is given by

$$C_d = 2\left|l'\Delta'\ddot{L}^{-1}\Delta l\right|. \qquad (7)$$

where Δ and $\ddot{L}$ are p × n and p × p component matrices for Liu estimator respectively.

The component matrices can be estimated from the equation (7) as,

$$\Delta = \frac{\partial^2 L(\varepsilon_\omega)}{\partial\hat{\beta}_d\partial\omega}\Big|_{\hat{\beta}_d,\omega_0,\sigma^2} = \frac{1}{\sigma^2}\mathbf{X}'\mathbf{D}(\hat{e}_d)$$

and

$$\ddot{L} = \frac{\partial^2 L(\varepsilon_\omega)}{\partial\hat{\beta}_d^2}\Big|_{\hat{\beta}_d,\omega_0,\sigma^2} = -\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}$$

where $D(\hat{e}_d) = diag(\hat{e}_{d1},\ldots,\hat{e}_{dn})$.

Therefore, C_d is

$$C_d = 2\left|l'\Delta\ddot{L}^{-1}\Delta\right| = 2\left|l'\left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{D}(\hat{e}_d)\right)'\left(-\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}\right)^{-1}\left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{D}(\hat{e}_d)\right)\right|$$

$$= \frac{2}{\sigma^2}\left|l'\mathbf{D}(\hat{e}_d)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}(\hat{e}_d)l\right|$$

$$C_d = \frac{2}{\sigma^2}\left|l'\mathbf{D}(\hat{e}_d)\mathbf{P}_X\mathbf{D}(\hat{e}_d)l\right|, \qquad (8)$$

where $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and $\|l\| = 1$.

Local Influential Analysis for Liu Estimator When σ^2 is unknown

When σ^2 is unknown, a similar calculation for $\theta' = (\beta_d', \sigma^2)$ yields

$$\Delta = \begin{pmatrix} \mathbf{X}'\mathbf{D}(e_d)/\hat{\sigma}^2 \\ \mathbf{e}_{sq}'/2\hat{\sigma}^4 \end{pmatrix} \qquad (9)$$

where $\hat{\sigma}^2$ is the maximum likelihood estimator of σ^2 and $\mathbf{e}_{sq}$ is the an n × 1 vector with elements e_{di}^2 where e_{di} is the i-th residual from the Liu estimator and $e_{di} = y_i - \hat{y}_{di}$. Since

$$\ddot{L} = -\begin{pmatrix} \mathbf{X}'\mathbf{X}/\hat{\sigma}^2 & 0 \\ 0 & n/2\hat{\sigma}^4 \end{pmatrix} \qquad (10)$$

the analogous result for θ is

$$C_d = \frac{2}{\sigma^2}\left|l'\left(\mathbf{D}(e_d)\mathbf{P}_X\mathbf{D}(e_d) + \mathbf{e}_{sq}\mathbf{e}_{sq}'/2n\hat{\sigma}^2\right)l\right| \qquad (11)$$

General analytic expressions for l_{max} are not known for (8) or (11).

If only β is of interest, the curvature is given by (8) with σ^2 replaced with $\hat{\sigma}^2$. Hence the ith curvature is given by

$$C_{di} = \frac{2}{\hat{\sigma}^2}e_{di}^2\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i + \frac{1}{n\hat{\sigma}^4}e_{di}^4, \qquad (12)$$

where $\mathbf{x}_i$ is the ith row of matrix **X**.

Perturbation of the Explanatory Variables

Here we assume that σ^2 is known. The following results can be easily adapted for the situation in which σ^2 is unknown and only β_d is of interest by replacing σ^2 and $\hat{\sigma}^2$.

Let s_j, j =1, ..., p, denote scale factors to account for the different measurement units associated with the columns of **X**, then the perturbed log-likelihood for Liu estimator $L(\beta_d \mid \omega)$ is constructed from (1) with **X** replaced by

$$\mathbf{X}_\omega = \mathbf{X} + \mathbf{W}\mathbf{S}, \qquad (13)$$

where $\mathbf{W} = (\omega_{ij})$ is an n × p matrix of perturbations and $\mathbf{S} = \text{diag}(s_1, \ldots, s_p)$. The perturbed log-likelihood is given by

$$L(\beta_d, W) = \frac{1}{2\sigma^2}[Y-(X+WS)\beta_d][Y-(X+WS)\beta_d] \quad (14)$$

The curvature C_{ij} associated with the perturbation ω_{ij} is given by

$$C_{ij} = \frac{2\Delta_{ij}'(X'X)^{-1}\Delta_{ij}}{\sigma^2}. \quad (15)$$

Here Δ_{ij} is the i[th] column of the matrix $\Delta_j = s_j(\mathbf{d}_j\hat{e}_d' - \hat{\beta}_d\mathbf{X}_j)$, where s_j is the j[th] scale factor, $\mathbf{d}_j$ is a p × 1 vector with a 1 in the j[th] position and zeros elsewhere, $\hat{e}_d$ is the vector of Liu estimator residuals, and $\hat{\beta}_{dj}$ is the j[th] Liu estimator parameter.

From the equation (15), $\Delta_{ij}'(X'X)^{-1}\Delta_{ij}/\sigma^2$ is a potentially large np × np matrix and determining its eigenvalues may be an unpleasant task. However, it can be shown that the nonzero eigenvalues of $\Delta_{ij}'(X'X)^{-1}\Delta_{ij}/\sigma^2$ are

$$\frac{1}{\sigma^2}\left[\mathbf{e}_d'\mathbf{e}_d\lambda_i + \sum_i \hat{\beta}_{dj}^2 s_j^2\right], \quad (16)$$

where λ_i is the i[th] eigenvalue of $\mathbf{S}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{S}$, i =1, ..., p. Thus

$$C_{\max} = \frac{2}{\sigma^2}\left[\mathbf{e}_d'\mathbf{e}_d\lambda_{\max} + \sum_j \hat{\beta}_{dj}^2 s_j^2\right]. \quad (17)$$

The above results can be used in situations in which less than p explanatory variables are perturbed by setting $s_j = 0$ for the unperturbed variables. In particular, when only the i[th] column of X is perturbed, $s_j = 0$ for j ≠ i and $\ddot{\mathbf{F}} = \Delta_i'(X'X)^{-1}\Delta_i\sigma^2$ where Δ_i is given in above with j = i.

Using this identity C_{max} is obtained from (17) as

$$C_{\max} = \frac{2s_i^2}{\sigma^2}\left[\|e_d\|\|r\|^{-2} + \hat{\beta}_{di}^2\right], \quad (18)$$

where $\|r\|^{-2} = \mathbf{d}_i'(X'X)^{-1}\mathbf{d}_i$.

Cook (1986, p. 147) proposed that the curvature values depend on an essentially arbitrary choice of the matrix **S**. The solution is to compute the scaled curvature that Schall and Dunne (1992) suggested where by a treatment of local influence in terms of scaled curvature is equivalent to a treatment in terms of curvature (8) with the canonical scaling matrix $\mathbf{S} = \sigma diag(\hat{\beta}_1^{-1},...,\hat{\beta}_p^{-1})$.

Therefore, when the i[th] individual explanatory variable is perturbed and corresponding diagnostic l_{max} can be obtained by finding the eigenvector corresponding to the largest absolute eigenvalue of matrix

$$\frac{2s_i^2}{\sigma^2}\left[(\mathbf{d}_i\hat{e}_d' - \hat{\beta}_{di}\mathbf{X}')(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{d}_i\hat{e}_d' - \hat{\beta}_{di}\mathbf{X}')\right]. \quad (19)$$

Example and Results

We used Longley (1967) data set to illustrate our methodology. This data set is already used by Cook (1977), Walker and Birch (1988), Shi and Wang (1999), Jahufer and Chen (2008) to identify influential cases. Hence, we can compare the influential cases detected by this method with previous studies. This data consist of 6 explanatory variables and 16 observations. The scaled condition number for this data set is 43,275 (Walker and Birch, 1988), which suggests the presence of strong

multicollinearity.

The detected most influential cases by Cook (1977) in OLSE, Walker and Birch (1988), Shi and Wang (1999) in ORRE using global and local influential method respectively, and Jahufer and Chen (2008) in modified ridge regression estimator (MRRE) using Cook's (1986) method are given in table 1.

Table 1: Most five influential observations

Cook (1977) in OLSE	Walker and Birch (1988) in ORRE – Global Method	Shi and Wang (1999) in ORRE –Local Method	Jahufer and Chen (2008) in MRRE – Local Method
5	16	10	10
16	10	4	4
4	4	15	15
10	15	16	5
15	1	1	16

Detection of Influential Cases When σ^2 is known

Here, we used Cook (1986) technique to identify the local influential observations for Liu estimator in linear ridge type regression model. The value of d that minimizes (4) for this data set is 0.9724.

we consider the influence of observations on the Liu estimator based on the eigenvector l_{max} associated with maximum eigenvalue of the normal curvature matrix C_d in (8) when we assumed the mean squared error σ^2 is known. The plot l_{max}^{m} against observation number is given in Figure 1. when the i^{th} element of l_{max} is found to be relatively large this indicates that perturbations in the weight ω_i of the i^{th} case may lead to substantial change in the results of the analysis and thus ω_i is relatively influential. In such situations, it will, of course, be important to investigate the i^{th} case to find the specific cause of the sensitivity.

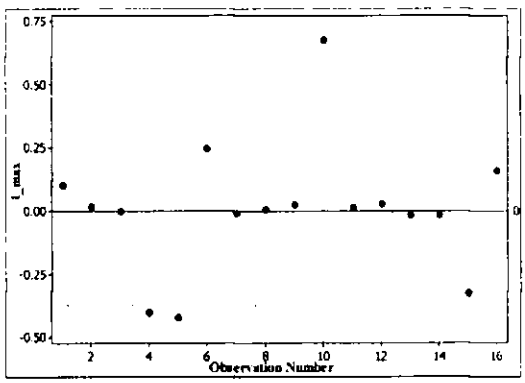


Figure 1: Index plot of l_{max} against Observation Number

From the above figure the most influential observations are cases 10, 5, 4, 15, 6 and 16 in this order. These influential observations are approximately the same for case deletion method but the order of magnitude is changed. Case 1 had been identified as high influential observation in Walker and Birch's (1988), Shi and Wang's (1999) paper however, this does not occur in Cook's (1977) and Jahufer and Chen (2008). Case 6 is not an influential point in

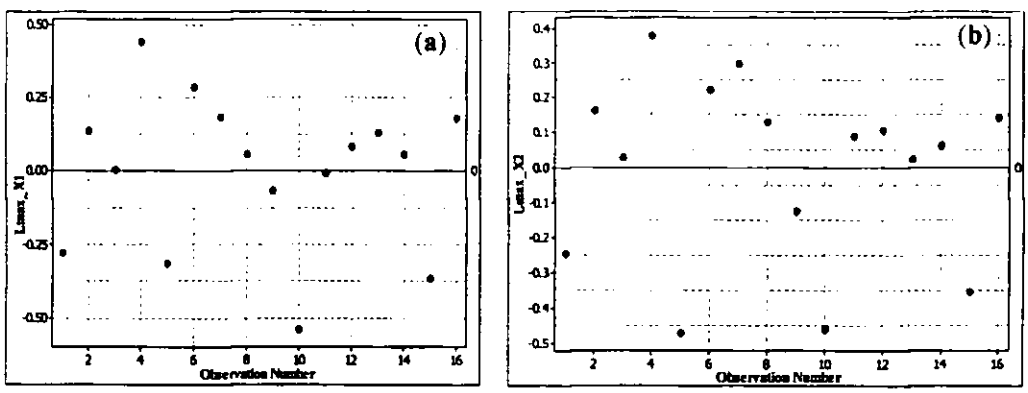
Walker and Birch's (1988) and Cook's method. However, it is a moderated and highly influential observation in Shi and Wang's (1999) and Jahufer and Chen (2008) papers. But, it is identified as one of the most influential case in this method.

Detection of Influential Cases When σ^2 is unknown

Local influential observations for Liu estimator is analyzed using equation (12), when we assumed that mean squared error σ^2 is unknown. The largest five C_{di} occurred for cases 10, 4, 5, 15 and 1 in this order. Therefore, in this method the most influential observations are cases 10, 4, 5, 15 and 1. These influential observations are approximately the same as in table 1 methods except that the order of magnitude is changed.

Detection of Influential Cases When Explanatory Variables Perturbed

Here, we consider the perturbation of individual explanatory variables. We use the scale factors of explanatory variables as suggested by Schall and Dunne (1992). The maximum value of C_{max} in (19) for separately perturbing explanatory variables x_i, i=1, ..., 6 are 581.701, 17.822, 2.955, 2.639, 358.909 and 3.031, respectively. Hence x_1, x_5 and x_2 are the largest (in this order) among the others on $\hat{\beta}_d$. The index plots of l_{max} based on perturbation of x_1, x_2, and x_5 are given in figure 2. It is obvious that most influential cases in figure 2, cases 10, 4, 15, 5, and 6 in (a) and cases 10, 4, 15, 6, and 5 in (c), respectively. But in (b) the most influential cases are 5, 10, 4, 15 and 7. These imply that the Liu estimator $\hat{\beta}_d$ is sensitive for values of x_1 and x_5 at cases 10, 4 and 15, and values of x_2 at cases 5 and 10.



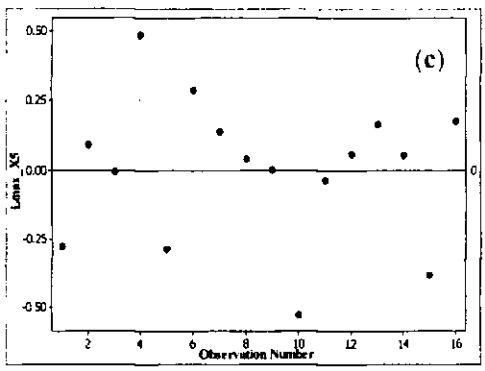


Figure 2: Index plot of l_{max}-X_1, l_{max}-X_2 and l_{max}-X_5, against Observation Number respectively

Discussion

Here, we propose a local influence diagnostic based on Cook's (1986) method to detect influential observations for ridge type Liu estimator when mean squared error is known or unknown, and perturbation of individual explanatory variable.

As many authors observed and signified in this paper, when the Liu estimator is used to replace OLSE for reducing the effect of multicollinearity, the influence of some cases can be modified. Hence, it is important to derive an effective diagnostics for detecting such anomalous points in Liu regression. Instead of using case deletion, this paper uses the local influence Cook's method to study the identification of anomalous observations. By perturbing different aspects of the model, the influence compact of the data on the Liu estimator can be analyzed.

Local influence diagnostics consider the joint influence of the data set, so it is useful to detect some influential patterns appearing in the data set. The using of signed elements in l_{max} will identify possible masking effects and canceling effects presented in the data as shown in the example.

The influential observations identified by this method is approximately same as the influence diagnostic methods of Cook's, Walker and Birch's, Shi and Wang's and Jahufer and Jianbao methods, but only the order of the magnitudes are changed.

ACKNOWLEDGEMENTS

The authors would like to thank the editor and the referee for their constructive and valuable comments.

This paper was written while the second author is a Ph.D. student under the supervision of the first author at the Xiamen university of China. This work was supported by the Chinese Scholarship Council (CSC) government of P.R. of China-2006.

References

Beckman, R.J., Nachtsheim, C.J., Cook, R.D., (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, 27 (4), 413-426.

Belsley, D.A., (1991). *Conditioning*

Diagnostics: Collinearity and Weak Data in Regression. Wiley New York.

Belsley, D.A., Kuh, E., Welsch, R.E., (1980). *Regression Diagnostics: Identifying Influence data and Source of Collinearity*, Wiley New York.

Cook, R.D., (1977). Detection of Influential Observations in Linear Regression, *Technometrics*, 19, 15-18.

Cook, R.D., (1986). Assessment of Local Influence. *Journal of Royal Statistical Association*, Series B, 48, 133-169.

Hoerl, A.E., (1964). Ridge Analysis. *Chemical Engineering Process Symposium* Series60, 67-77.

Jahufer, A., and Jianbao, C., (2008). Identifying Local Influence in Modified Ridge Regression Using Cook's Method. *Sri Lankan Journal of Applied Statistics* Vol. 9, 93-108.

Lawrance, A.J., (1988). Regression Transformation Diagnostics Using Local Influence. *Journal of American Statistical Association*, 83, 1067-1072.

Lawrance, A.J., (1991). Local and Deletion Influence. In: Stahel, W., Weisberg, S. (Eds.), Directions in Robust Statistics and Diagnostics, Part1. Springer, Berlin, pp. 141-157.

Liu, K., (1993). A New Class of Biased Estimate in Linear Regression. *Communications in Statistics - Theory and Methods*. 22, 393-402.

Longley, J.W., (1967). An Appraisal of Least Squares Programs for Electronic Computer from the point of View of the User. *Journal of American Statistical Association*, 62, 819-841.

Mallows, C.L., (1973). Some Comments on Cp. *Technometrics*, 15, 661-675.

Schall, R., Dunne, T.T., (1992). A note on the relationship between parameter collinearity and local influence. *Biometrika*. 79, 399-404.

Shi, L., (1997). Local Influence in Principal Component Analysis. *Biometrika*, 84 (1), 175-186.

Shi, L., Wang, X., (1999). Local Influence in Ridge Regression. *Computational Statistics & Data Analysis*, 31, 341-353.

Stein, C., (1956). Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. *Proceeding of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, 197-206.

Thomas, W., Cook, R.D., (1990). Assessing the influence on predictions from generalized linear model. *Technometrics*. 32, 59-65.

Walker, E., Birch, J.B., (1988). Influence Measures in Ridge Regression. *Technometrics*, 30, 221-227.